Discussion Paper

Data Linkage to Validate and Calibrate Traffic Estimations on a Nationwide Scale: A Framework for Official Statistics

Inan Bostanci (Utrecht University, Master Student)
Yvonne Gootzen (Statistics Netherlands)
Peter Lugtig (Utrecht University)

March 16, 2023

Abstract

Traffic estimation is an important area in official statistics and is used by policymakers in their decision-making process for regional planning. This paper describes a framework that relies on combining large amounts of data from traffic loop sensors with administrative data of the entire population of the Netherlands to estimate traffic intensities during rush-hour for all major roads in the Netherlands. Multiple calibration models are developed and compared, and although the models do suggest that traffic counts are sometimes over- or underestimated in some regions, the models overall perform well.

1 Introduction

Road traffic congestion has significant negative impacts on the environment, public health, and the economy Condurat et al. (2017); Hymel (2009); Levy et al. (2010); Barth and Boriboonsomsin (2008). Vehicles stuck in traffic jams dissipate nonrenewable fuel sources, increase air pollution, and hinder mobility. In 2018, traffic congestion cost the Dutch industry alone 1.4 billion euros NL Times (2019). In the same year, it cost the entire UK economy around 7.9 billion pounds, with drivers spending a total of 178 hours on average in congestion INRIX (2019). Reoccurring traffic congestion is often a result of commute travel during morning rush hour Falcocchio and Levinson (2015). Transport demand modeling has been used by researchers and policy makers to understand and predict traffic and improve regional planning strategies Möller (2014); Ortúzar S. and Willumsen (2011). For example, statistical and machine learning methods can inform stakeholders which roads might get particularly busy and improve traffic flow by building additional roads Eagle and Greene (2014); McFadden (1974).

A characteristic of transport demand research is its increasing usage of multiple data sources, such as the combination of travel survey data and data from traffic loop sensors or floating cars Lana et al. (2018); Willumsen (2021). The widespread distribution of traffic loop sensors across road networks makes them attractive for traffic research of large geographical areas. Traffic loop sensors count the total number of cars passing by per location and thereby produce data of enormous size. In the Netherlands, around 20,000 sensors are placed on many roads across the road network, giving insight into the total traffic intensity at specific locations Puts et al. (2019). However, while data from travel surveys and floating cars are linked to the individual driver, loop sensor data alone does not allow for conclusions about driving behavior on the individual level. Therefore, loop sensors can complement large-scale transport demand research, but are not sufficient alone.

Linking observed counts from traffic loop sensor data to individual data bears the potential of analyzing driving behavior on a large, nationwide scale Klingwort and Burger (2021). This could provide policy makers and regional planners with traffic modeling methods for scenarios, e.g., to estimate the local effects of planned housing developments or large factories on traffic demand of specific roads.

Gootzen et al. (2020) developed a framework at Statistics Netherlands (CBS) to estimate rush hour traffic on a nationwide scale, by linking administrative data, infrastructure

data, survey data, and observed passenger counts from public transport. Administrative data contains information on all residents in the Netherlands. The key information for the framework are the home and work or school locations of residents, and demographic variables. In a case study on the Rotterdam Metro, Gootzen et al. (2020) applied this framework to the case of public transport. First, they estimated the number of metro passengers. Then, they calibrated the estimates with observed counts from automatic card-readers of metro stations, resulting in more accurate expected counts. The calibration was then used in a scenario analysis of population growth.

The current paper is following up on the work of Gootzen et al. (2020) under the umbrella of the Data on Cities and Mobility in the Netherlands (DaCiMob) project of CBS, the Dutch central agency for statistics Roos and Gootzen (2021). In their case on public transport, all metro stations are equipped with card-readers. In the road traffic case studied in the current paper, only a share of roads is equipped with traffic loop sensors. The aim of the current paper is to validate and calibrate expected counts for the case of rush hour road traffic on a nationwide scale. To be able to calibrate counts on unobserved roads, a modeling approach is used in this paper, which utilizes road and geographical features. Therein, the paper answers two questions by extending the DaCiMob framework with traffic loop sensor data: 1) How valid are the nationwide traffic estimates that the DaCiMob framework produces? 2) How much can the validity be improved by the inclusion of sensor and infrastructure data?

The remainder of this paper is structured as follows: First, we give a background on the current state of affairs in traffic research and relate it to this study. Next, we describe the general framework and methodology of the DaCiMob project. Afterwards, we explain how the data sources are linked to enable traffic predictions and comparisons. This section also includes descriptive statistics and visualizations of the sensor data. Following, we compare the expected traffic counts to observed counts from traffic loop sensors and train a model for calibration. Then, we use the results to calibrate expected traffic counts on the entire road network during rush hour. Finally, we discuss results and limitations and draw a conclusion about the use and potential of the framework.

2 Background

Due to the complexity of traffic flow, studies often focus on traffic forecasting with simulation software Krajzewicz et al. (2002, 2012); Lana et al. (2018). These forecasts are mostly short-term and focus on compact geographical areas Lana et al. (2018), but recent advances in computational power and big data have enabled a growing body of literature on long-term forecasting He et al. (2019); Qu et al. (2019); Wang et al. (2021). This article contributes to the literature by looking at long-term traffic count predictions on a nationwide scope for the case of the Netherlands.

This country is a unique case for such a traffic study. In terms of population size, population density, and urbanization, the Netherlands is comparable to many metropolitan areas. For example, its population density of 522 per km^2 CBS (2022b) is similar to the metropolitan areas of Melbourne and Los Angeles. Australian Bureau of Statistics (2021); U.S. Census Bureau (2021). However, in terms of land area, the Netherlands is still approximately five times larger than Tokyo-Yokohoma Demographia (2016). Further, it contains countrysides and agricultural areas which are not typically

seen in metropolitan areas. Results from this study might therefore be interesting for traffic planning of metropolitan areas and smaller countries.

Observed traffic counts are often an important component for calibration in studies on traffic flow and demand. Observations may come from stationary traffic loop sensors or cameras Ma et al. (2021b); Behrisch et al. (2009); Tcheumadjeu et al. (2012); Shafiei et al. (2018). Some studies use floating car data from fleets of taxis or mobility surveys, which can track vehicles over their entire route (see Heyns et al., 2019). Currently, floating car data has a low coverage of the traffic network and is therefore of limited use for traffic demand modeling de Fabritiis et al. (2008). Some studies use a combination of both data types (see Ma et al., 2021a). Often, observed counts are used to validate and calibrate the simulated expected counts. Because observed traffic counts only provide data on a subset of the road network, there are many blank areas on the network where expected counts cannot be validated. It might therefore be that expected counts were specifically tailored to these observed sites on the network and are potentially biased.

One main culprit in traffic demand modeling for long-term urban planning is in estimating the origin-destination (OD) matrix. The OD matrix contains all possible origin- and destination points of a spatial area as the rows and columns, and the number of trips in the cells. To estimate the number of cars on each road, researchers need to know where drivers start and end their trip, and how many trips happen for an OD pair. OD matrices can be estimated such that their resulting traffic estimates are in accordance with observed counts, e.g., from sensors. However, there is often no unique solution for the OD matrix. In other words, there can be multiple different estimates of the OD matrix that fit the observed counts equally well. This is because observed counts are only available for specific sites on the road network, leaving many areas unobserved. To alleviate this problem, prior information from historical data or surveys is often used as a starting point. The prior information is used to generate an initial OD matrix, which then is optimized such that results fit to observed counts (Wang et al. (2012); Bauer et al. (2018); Shafiei et al. (2018); (for a review, see Bera and Rao, 2011)). It becomes challenging and computationally costly to estimate the OD matrix on a nationwide scale, because prior information needs to be representative of the population and spatial area.

Rather than using historical data or survey data, the DaCiMob project uses administrative data to build a spatially detailed OD matrix with nationwide coverage. Administrative data is updated yearly and includes information on the home and work location of the entire registered population in the Netherlands. This type of data is increasingly being used for official statistics, and research on data validity and correction is growing Bakker (2012); Zhang (2012); Scholtus et al. (2015); van Delden et al. (2016); Oberski et al. (2017); Pankowska et al. (2018). This puts the DaCiMob project into a unique position: The OD matrix for commuters now covers the entire population of interest without relying on estimation and optimization steps. To our knowledge, no study has investigated traffic estimates from such a rich and encompassing OD matrix yet. Considering that the main interest in this study is home-work commuting in the Netherlands, it is plausible to assume that this OD matrix covers the population rather precisely. Because the OD matrix does not require optimization, this study can use observed counts to validate traffic estimates. In doing so, it contributes to traffic demand research by proposing a method to predict the quality of traffic estimates on unobserved road segments.

Not all trips in the OD matrix are done by car. The mode of transport of each individual needs to be known to estimate road traffic. Insights on travel mode choice usually come

from travel surveys. Traditionally, respondents record their daily trips and the respective mode and duration in travel surveys, but smartphone applications utilizing GPS sensors have recently gained popularity McCool et al. (2021). Travel mode models classify the travel mode of individuals and are usually trained with data from travel surveys, which also include demographic variables McFadden (1974); Barff et al. (1982); Shamshiripour et al. (2019); Zhao et al. (2020); Wang and Ross (2018). Such models can be employed to classify the travel mode of each case in an OD matrix. In the case study on public transport in Rotterdam, such a travel mode model was used to estimate the number of public transport users for each OD pair Gootzen et al. (2020). By inspecting the relationship between geographical and road characteristics, the presented methods in this study could be used to draw conclusions about potential ways to improve travel mode models.

3 The DaCiMob project and Data

The main interest of the DaCiMob project lies in the average road traffic counts during rush hour. Two approaches are done in parallel to obtain traffic counts. The first approach builds on traffic loop sensors and leads to observed counts on a subset of the nationwide road network. The second builds on administrative data and leads to expected counts on the entire network. Then, the results are linked and optimized.

In the first approach, traffic loop sensor data is linked to infrastructure data to obtain observed traffic counts on a subset of the Dutch road network. Sensors are placed on road sections of many of the main roads and data is publicly accessible. A projection of sensor data onto infrastructure data leads to traffic counts of many roads for any desired time window. This can be used to get the traffic counts during rush hour and therefore gain a picture of the traffic demands in specific regions or on specific roads.

However, this data alone is insufficient to gain deeper insights for the scenarios proposed in the introduction for two reasons: 1) Sensors count the frequencies of cars without providing background information on who the drivers are. For meaningful projections into the future, this background information is crucial, because driving behavior is dependent on personal characteristics. If we do not know the composition of drivers at the time of measurement, we cannot project into a future with a different (or similar) composition of drivers. 2) Sensors only give insight into a subset of the road network. Inferring from that subset, which can be seen as a non probability sample, onto other roads is not possible without the inclusion of other data sources, such as regional or population data. If the subset were a random sample, inference could be possible without the inclusion of other data sources.

In the second approach, administrative data, survey data, and infrastructure data is linked to obtain expected traffic counts of the entire Dutch road network. Administrative data contains the home and work location of employees on a neighborhood level and characteristics of employees, such as demographic characteristics or car ownership. Travel surveys can be used to model the travel mode choice for commutes, i.e., whether an employee commutes by car, public transport or bike. By using modeling variables that are present in administrative data, and by introducing infrastructure data, the number of car users per neighborhood can be estimated and their probable commuting routes computed. Thus, traffic counts during rush hour can be estimated for the entire road

network.

The shortcoming of the second approach is the lack of validation through observed counts. This approach assumes that all traffic is due to commuting. Although commuting arguably accounts for the largest share of traffic during rush hour, a share of the traffic is also caused by other travel motives. There might be regional differences in the size of this share, which can lead to regional biases. Further, expected counts are based on yearly administrative sources and describe the typical mobility of an average day, while real-life traffic is known to have daily and hourly fluctuations due to circumstances such as weather and school holidays. These patterns cannot be distinguished from administrative sources, but they can be found in observed counts. Aggregating the observed counts over a period that corresponds to the morning rush hour allows for a fairer comparison on expected counts. From a policy-makers perspective, getting an idea of the scope of error and being able to correct for it is important when costly decisions are dependent on those expected counts.

In the third step, the shortcomings of both steps are overcome by linking the observed on edges and expected traffic counts and calibrating expected counts. This is achieved by projecting them on the same level. Now, expected counts can be validated, assuming that the observed counts are ground truth. Next, the accuracy of expected counts on roads with no sensors can be modeled with road (segment-based) and regional (area-based) characteristics. The final result is calibrated estimates of traffic counts.

Figure 3.1 gives a schematic overview of the project. A detailed description can be found in the project documentation Roos and Gootzen (2021). In the remainder of this section, the data sources used to acquire observed and expected traffic counts will be described. After this, preprocessing and linkage will be explained and the resulting datasets will be shown. Then, the calibration method will be described.

3.1 Traffic Loop Sensor Data

Traffic loop sensor data for the Dutch road network is provided by the National Road Traffic Data Portal (NDW) and is openly accessible NDW (2020), and data quality has been studied (see Melnikov et al., 2015; Puts et al., 2019; Tennekes and Puts, 2018). Meta data contains the sensor ID, location (measured in longitude and latitude), the bearing of the road at the point of measurement (i.e., the trajectory), and the road type. The location does not refer to the point of measurement, but to a device placed next to the road containing most of the hardware. Sensor data contains the sensor ID and the measurements with a timestamp. Due to practical and regulatory reasons, this study used a reformatted version of traffic loop sensor data that is stored on an internal server at CBS. Sensors measure the traffic intensity, which is the number of vehicles passing by per minute.

Data from 2019, the most recent year before the pandemic, was chosen to exclude the effects of lock downs and quarantine regulations. Since the focus of this study are morning commutes, only observations between 5 and 9 am were included. In 2019, 16,734 sensors were placed on Dutch roads. Out of these, around half also measured the vehicle length. This information is crucial, because this study focuses on commuting in personal vehicles, and traffic from buses or trucks are disturbing factors. Therefore, sensors that do not provide the vehicle length were removed, as well as measurements of vehicles that are longer than 5.6m. The remaining data set contains 7,775 sensors that made 41,942,720 measurements with a median of 6 cars per minute (IQR = 12) in 2019. A visual inspection suggests an adequate road network coverage, because the

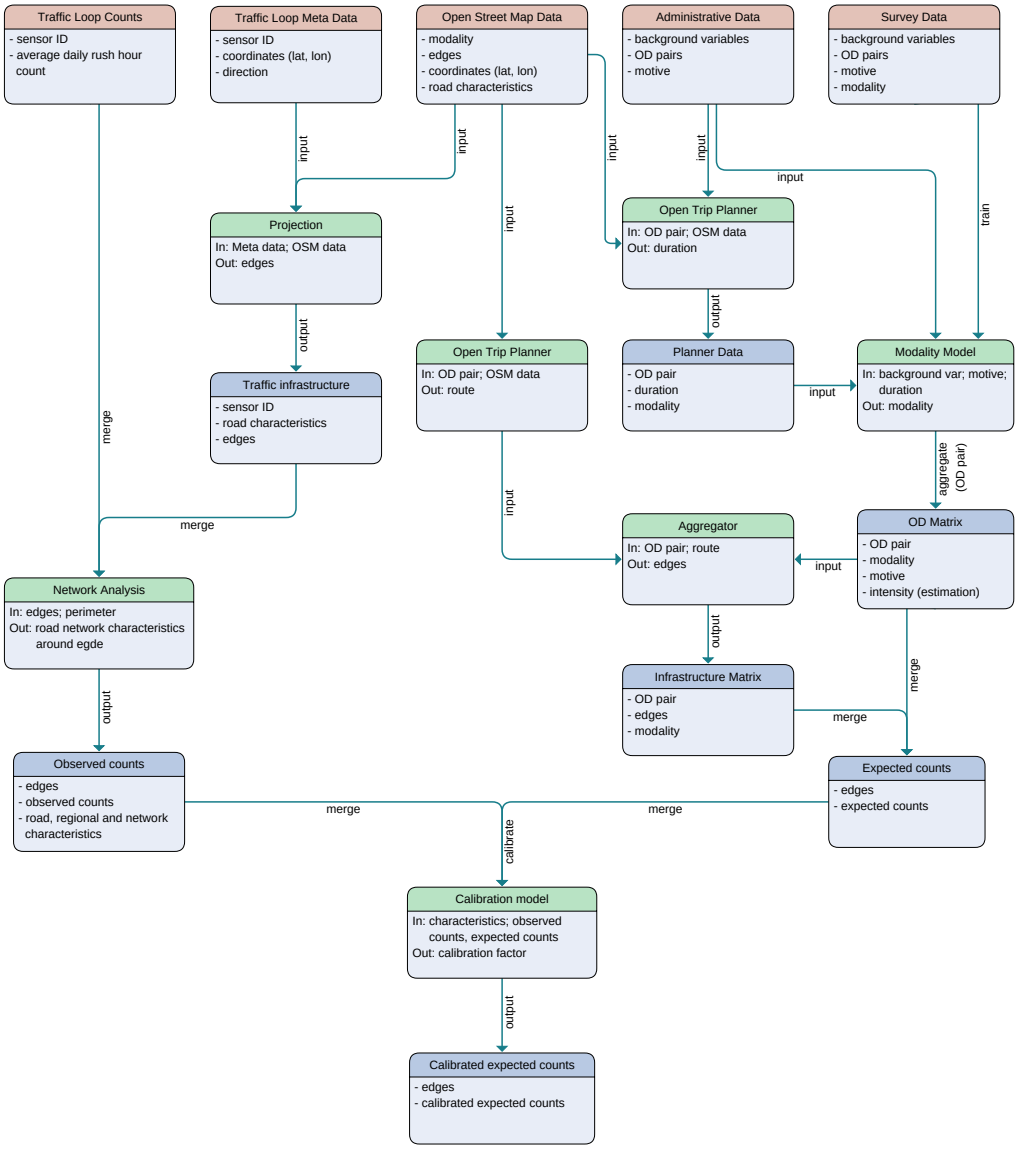


Figure 3.1. Schematic overview of DaCiMob project (modified version of Figure 1 in Roos and Gootzen (2021)).

sensor density of a region seems to correspond to road density and population density.

Figure 3.2 shows an example of the cars per minute as measured by one sensor during morning rush hour on December 11th, 2019.

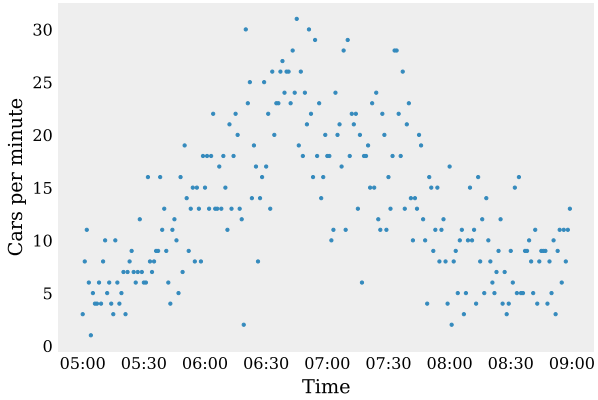


Figure 3.2. Measurements of one sensor on 12-11-2019.

3.2 Infrastructure Data

The infrastructure data on the Dutch road network originates from OpenStreetMap OpenStreetMap (2021). The Python package OSMnx v1.1.2 Boeing (2017) was used to load a geospatial graph of the Dutch road network. This can be seen as a graph G consisting of a set of nodes U and a set of edges E, where $E = \{(u, v) \in U \times U\}$. In other words, an edge is defined by its starting node u and its ending node v. There are 681,994 nodes and 1,669,220 edges in G. The nodes are conjunctions and turning points, and the edges are road segments connecting the nodes. Many main roads (e.g., highways) are one-directional and have a parallel road that leads into the opposite direction. Let us think of a highway road segment A that leads south, with a parallel segment B that leads north. u of segment A will be equal to v of segment B. The geospatial graph also contains information about each node (i.e., the location) and each edge (i.e., the road type, speed limit, etc.).

3.3 Administrative Data

Administrative data was used by CBS to build a data set of all registered employees in the Netherlands. This is possible at CBS because of the wide usage of registers in Dutch administrative bodies and their accessibility for CBS. Administrative data covers the entire population of registered employees in the Netherlands. Relevant variables for this study are the home and work locations on the district level, and personal characteristics, such as demographic variables and car ownership (for a detailed description, see Gootzen et al., 2020; Roos and Gootzen, 2021).

3.4 Survey Data

Survey data originates from the Dutch National Survey Onderweg in Nederland (ODiN; Boonstra et al. (2021); CBS (2021)). This is a travel diary survey, where respondents provide demographic variables and track each of their trips, including trip characteristics, over a course of a few weeks.

4 Methodology

4.1 Preprocessing of Traffic Loop Sensor Data

Preliminary analyses showed significant data quality issues for the first quarter of 2019 and seasonal effects on traffic intensity. For example, the intensity decreased in all regions in December, and increased in touristic regions during summer. To reduce the time frame to a time that is relatively unaffected by seasonality, the month of May was chosen for this study. However, the calibration procedure can be applied to other months as well. A visual inspection of the data points out that workdays are not distributed equally over employees, but it is expected that the largest portion of employees travels to work between Tuesday and Thursday. While it could be possible to include the aforementioned in the model and compensate for some of the effects, the nature of the expected counts does not advocate for this approach. The expected counts are based on a transportation modality model that was trained on a selection of survey data where it was given that someone was travelling for work. To match this assumption, we chose to base the observed counts on days where the largest portion of employees travel to work. Therefore, data from other days was excluded. Because sensors count traffic on each road lane separately, traffic counts were aggregated per road. Per sensor, the average daily sum was computed. 859 sensors with an average of 0 were removed, since this indicates malfunctioning Puts et al. (2019).

4.2 Preprocessing of Infrastructure Data

Sensors are mostly placed on main roads (A-roads and N-roads) and their connected ramps. To decrease computational costs, edges from other types were removed, such as pedestrian ways. This left 68,476 edges in E. A full list of all removed road types can be found in Appendix A.

4.3 Linking Sensor Data to Infrastructure Data

To obtain the road segment (u, v) that a sensor is placed on, sensor data was linked to infrastructure data by finding the nearest edge for each sensor. For this, we projected both data sets and coded them in the same coordinate reference system of EPSG:28992. The nearest edge of each sensor was relatively close, with a mean distance of 37cm. Some sensors were linked to the road leading towards the wrong direction. Such errors occur because rather than providing the location of measurement, the meta data refers to the location of a device placed next to a street, which gathers measurements from lanes leading towards both directions.

We computed the bearing of the edge and compared it to the bearing according to the sensor's meta data to get a scope of this linkage error. Figure 4.1 shows the correlation between the sensor bearing and the edge bearing. A perfect linkage would lead to a straight line from the bottom left to the top right. Most points scatter around this line. Some deviation is expected, because road segments that are not perfectly straight might have a slightly different bearing overall than at the exact point of measurement. However, some clusters that are far off from this line can be seen. We corrected this by switching u and v for edges where the difference between the sensor and edge bearing was between 90 and 270. Due to missing data in the sensor's bearing, this reduced the total amount of sensors to 4,771 (28.5% of the original 16,734). The final result of this

linkage step is the average daily count of cars during morning rush hour for a subset of road segments (u, v) in E.

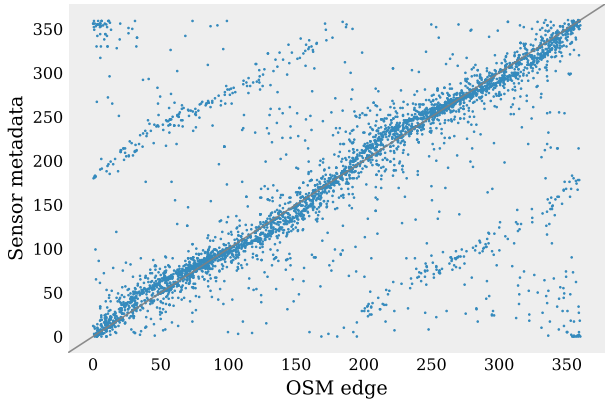


Figure 4.1. Bearing of road: OSM edge vs. sensor metadata.

4.4 Linking Administrative Data to Infrastructure Data

Gootzen et al. (2020) developed a framework to estimate Rotterdam metro traffic counts with administrative data. For the DaCiMob project, this framework was then extended to estimate road traffic counts on a nationwide level. For the sake of completeness, we briefly explain the steps that were taken at CBS to provide the expected traffic counts for this study. For a detailed documentation, see Roos and Gootzen (2021).

An origin-destination (OD) matrix was created using every possible home (origin) and work (destination) district from the administrative data, where the entries resemble the number of employees per OD combination. The OD matrix also includes demographic variables and trip characteristics of each employee. After removing unusable combinations, this matrix contains 69,830 unique OD pairs.

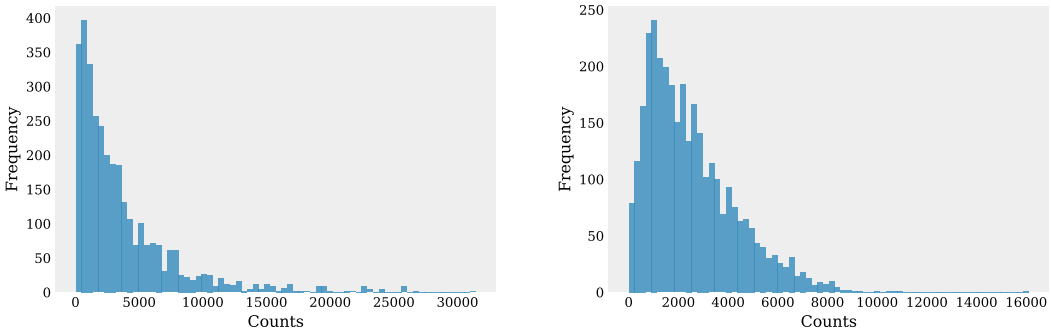
Next, a travel mode model was trained with the survey data. This is a Naïve Bayes model, which estimates the probability for each travel mode per person, given demographic variables and trip characteristics. Model details can be found in Roos and Gootzen (2021).

The model was applied onto the OD matrix. Because one parameter of the Naïve Bayes model is the travel duration, the duration for each mode of each OD pair had to be estimated first. This was done using the open source multi-modal trip planner OpenTripPlanner (2022) in combination with OpenStreetMap. Since the origin and destination of each entry are districts, the centroids of districts were used for OpenTripPlanner. In this step, the trip route for car trips was also obtained as a polyline containing the start and end of the route and the locations passed on the route. The OD matrix and the travel times were then fed into the Naïve Bayes model to estimate the choice probability for each travel mode per person. The probabilities were then aggregated for each OD pair. This led to an estimated share of the total commute traffic for each mode per OD pair, which was multiplied by the number of employees for that pair. For example, if there are 100 employees living in district A who all have work locations in district B, and the model estimates aggregated probabilities of 50% for using a car, 30% for using a bike and 20% for using the public transport, we estimate 50 car commutes, 30 bike commutes and 20 public transport commutes from district A to district B. The entries in the OD matrix were then reduced to the estimated number of

car commutes per OD pair (i.e., 50 entries would remain in the OD pair of the previous example).

The number of expected car trips from A to B were linked to the respective polyline connecting A and B. Next, the locations in each polyline were projected onto the road segments E of the road network. The result of this linkage step is the expected daily number of cars during morning rush hour for each (u, v) in E. A total of 60,621 edges (out of 68,476) have an expected count above 15. Counts under were not included in the study because of privacy preservation.

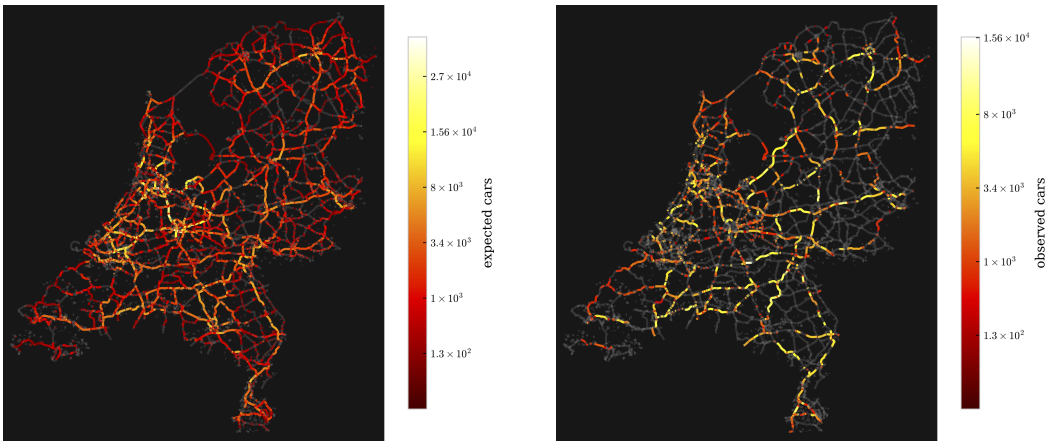
4.5 Resulting Data Sets



(a) Expected traffic counts. (b) Observed traffic counts.

Figure 4.2. Histograms of expected and observed traffic intensities per road segment.

Figures 4.2a and 4.2b show histograms of the expected and observed traffic intensities. Both resemble a right-skewed distribution. The expected counts in figure 4.2a seem to have a slightly lower mode of around 1,000, but the tail has a long spread up to 30,000 traffic counts, which is around twice as large as the outliers in the observed traffic counts.



(a) Expected traffic counts. (b) Observed traffic counts.

Figure 4.3. Expected vs. observed traffic intensities on the Dutch road network. Note that the legends differ due to differences in value range.

Figures 4.3a and 4.3b show the expected and observed traffic intensities on the Dutch road network. Note that fewer road segments have values in Figure 4.3b and 4.2b, because sensors are only present on some road segments. There are some agreements between figure 4.3a and 4.3b. For example, segments that have observed traffic counts around 8,000 (yellow in 4.3b) tend to also have expected traffic counts around 8,000

(orange in 4.3a). However, although both figures show higher counts around Rotterdam, Amsterdam, and Utrecht, the expected counts are relatively far north of the observed counts. This area seems to be where most of the outliers in expected counts are clustering. One possible explanation is that the transportation modality model does overestimates the probability of people in these highly urban areas to travel by car compared to the lesser urban areas of the country.

To be able to compare observed to expected counts, both data sets have to be linked. This is now possible, because they were brought onto the same level of granularity of road segments. In other words, because u and v refer to the same underlying road network G, the data sets can simply be merged.

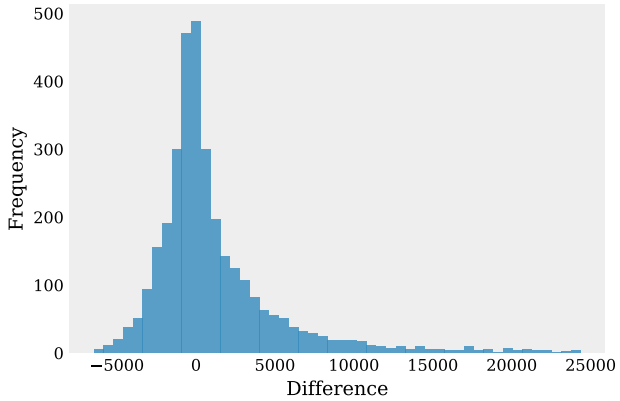


Figure 4.4. Difference between expected and observed counts.

Figure 4.4 shows the difference between the expected and observed traffic counts of road segments. The mode of the distribution is around 0, which indicates that most expected counts are close to the observed counts (median = 25.5). However, the distribution is right-skewed with a long tail. Underestimation of traffic counts seems to stop at around 5,000 counts, but overestimation can range up to around 25,000 counts. This reflects the difference in outlier spread between the two data sets.

4.6 Introduction to the Calibration Factor

First, we will compute a calibration factor following the approach of the Rotterdam metro study Gootzen et al. (2020). There, observed counts were used to calibrate the model for expected counts. Unlike in the metro network, observed counts of the road network are only present for a subset of road segments. Thus, we will model the calibration factor using road and regional characteristics. The purpose of this is two-fold: First, this will help us investigate how input features, such as road characteristics, are related to under- and overestimation (i.e., the quality of estimates). Second, we can use this to estimate the calibration factor on road segments that are not equipped with sensors. The estimated calibration factor can then be used to correct the expected counts of all road segments, including those without sensors.

The calibration factor C is a random variable that takes the value

$$c_{(u,v)} = \frac{obs_{(u,v)}}{exp_{(u,v)}} \quad \text{for all } (u,v) \text{ in } E,$$

which is the quotient of the observed count obs divided by the expected count exp of a road segment (u,v).

Since both OBS and EXP strictly take positive values, C is guaranteed to only take positive values. For roads that are equipped with sensors, C is known, because observed counts are available. For other roads, C is hidden, but the denominators (the expected counts) are known. By multiplying the expected counts of these roads with $\hat{c}_{(u,v)}$, the estimated calibration factor, we can correct expected counts.

While the primary use of C is calibration, it can also be seen as a quality metric for expected counts: If $c_{(u,v)}$ is larger than 1, we have more observed than expected cars on road segment (u, v) and hence underestimate the number of cars. If it is equal to 1, an expected count perfectly estimates the observed count. To see whether the calibration improved the quality, we can compute $c^*_{(u,v)}$ as $obs_{(u,v)}/\hat{c}_{(u,v)} \cdot exp_{(u,v)}$.

4.7 Modeling the Calibration Factor

Due to the exploratory nature of this approach, we will model C both with a linear regression and with a random forest regression. Linear regression has the advantage that parameters are easy to interpret, but is restricted due to its assumptions about the data. Random forest regression cannot be interpreted as straightforwardly, but it does not impose any assumptions. This can be beneficial when there is a lack of prior knowledge about the data. We will use the package statsmodels v0.13.1 Seabold and Perktold (2010) for the linear regression and Scikit-learn v0.23.2 Pedregosa et al. (2011) for the random forest regression. The hyperparameters in the random forest model will be tuned to maximize the model performance in a test set.

The variables used as input for the model will be characteristics of the road segment, the surrounding region, the surrounding road network, and the province. These variables will be denoted by $input$. Because C is a function of OBS and EXP, and EXP is available for all road segments, we will also include the expected counts in both models. We will additionally include EXP^{-1} in the linear regression, because a visual inspection suggested that EXP has an inverse squared relationship with C. The variable EXP and its derivatives will be available when the model is applied in practice.

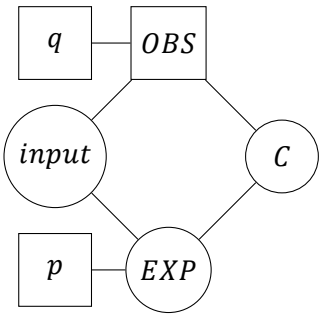


Figure 4.5. Graphical representation of assumed underlying model. Note: q = unobserved variables, p = other parameters for expected counts, boxed variables are not included in modeling C.

The assumed underlying model in figure 4.5 shows how C is directly caused by observed and expected counts. These are, in turn, both results of model input variables, but also of components that are not included in modeling C. By modeling C with $input$ and EXP, the estimate of $input$ represents the indirect effect that $input$ has on C through OBS, while the estimate of EXP represents the indirect effect that other parameters p of the expectation model have on C. Unobserved variables q remain a not-estimated source of error.

Table 4.1. Input variables to model the calibration factor C

Segment characteristics	Regional Characteristics	Road network characteristics	Province	Interaction variables (linear regression only)
Road type (dummy variable):	Population density	*12km perimeter:*	*Dummy variable:*	Population density · Motorway
Motorway	*Dummy variable:*	Nr of edges	Drenthe	Population density · Zuid-Holland
Motorway_link	Border in 10km perimeter	Max speed limit	Flevoland	Population density · Noord-Holland
Primary		Min speed limit	Friesland	Population density · Utrecht
Primary_link		Max lanes	Gelderland	Lanes · Max Lanes (1.5km)
Trunk		Min lanes	Groningen	Max lanes (12km) · Zuid-Holland
Trunk_link		*1.5km perimeter:*	Limburg	Max lanes (12km) · Noord-Holland
Number of lanes		Max lanes	Noord-Brabant	Max lanes (12km) · Utrecht
			Noord-Holland	Max lanes (12km) · Noord-Brabant
			Overijssel	
			Utrecht	
			Zeeland	
			Zuid-Holland	

Five sets of input variables can be found in table 4.1. We will start with a model that only contains EXP and then subsequently add each set of input variables. This will result in five models for each modeling approach, and has the purpose to gain insight into the added explanation of each set. Interaction variables are only included in the linear regression and will be added simultaneously to the corresponding sets of input variables. Using road segment characteristics as predictors is based on the expectation that the error originates in part from unexpected route choices of employees. One can imagine that an employee might prefer another route than suggested by OpenTripPlanner, e.g., due to habit or because it is less straining. Additionally, expected counts might be more accurate in areas with few alternative roads, because driver's choices are limited. To account for this, network characteristics of the local road network surrounding a segment are included. We obtained these characteristics by creating a 12km perimeter (based on exploratory findings) around a segment and analyzing intersecting segments with the package GeoPandas v0.10.1 Jordahl et al. (2021). It is expected that incoming traffic from neighboring countries is not covered by the expected counts. One solution would be to extend the network Klingwort and Burger (2021). This was however outside the scope of the current paper. Instead, we computed a dummy variable as a regional characteristic, that marks whether a segment is within 10km of the national border. This distance was chosen to plausibly cover traffic to and from bordering regions without covering too much of unrelated traffic. Additionally, we linked road segments to official statistics on the population density from CBS (2022a). Using regional characteristics as predictors is based on the observation that there are regional differences in traffic counts in figure 4.3b. Finally, the province is included based on the expectation that C and the expected relationships differ between provinces due to differences in infrastructure. In the linear regression model, we also include interaction variables that seem plausible and improve adjusted R^2.

4.8 Assessing Model Performance

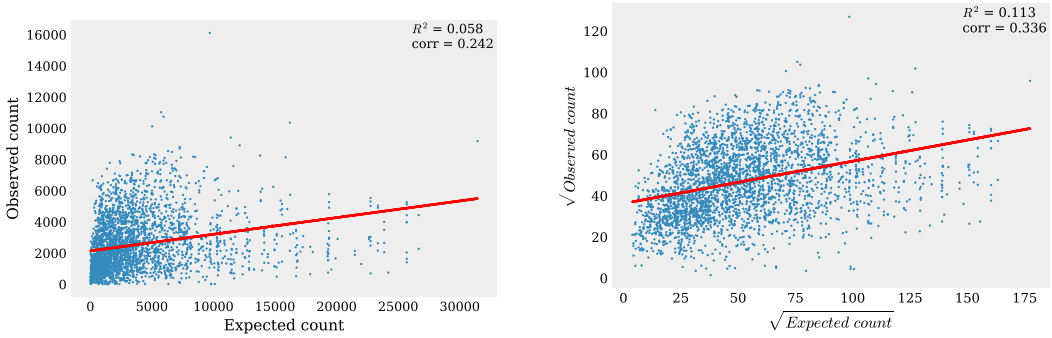
We will assess the performance of the calibration models as follows: First, we will perform a validation analysis V_1 with a linear regression that predicts the observed counts with the expected counts. The R^2 will inform us how accurate the expected counts are before calibration. In other words, this tells us how well the expected counts

fit to the observed counts and will therefore serve as a baseline. After training each model, we will compute the prediction $\hat{c}$ for each road segment of a test set and calibrate expected counts accordingly. We will randomly sample 60% of the data for training (1,341 edges) and leave 40% for testing (894 edges). To inspect whether the models are overfitting to the train data, we will compare the Root Mean Squared Error ($RMSE$) for the train and test data. We will apply each model to the test data to obtain the prediction $\hat{C}$. Next, we will perform another validation analysis V_2 in the test set, where we predict the observed counts with the calibrated expected counts ($\hat{c}_{(u,v)} \cdot exp_{(u,v)}$). The changes in R^2 will inform us how much the expected counts improved due to the calibration of each model. Finally, we will inspect how much the expected counts improved in quality after calibration by computing and visually inspecting C^*.

5 Results

5.1 Quality of Expected Traffic Counts

Figure 5.1a shows a scatter plot with the expected traffic counts on the x-axis and the observed traffic counts on the y-axis. Each point resembles a road segment. The difference in outlier spread mentioned in 4.5 becomes more evident, since an increase in expected counts does not lead to a proportional increase in observed counts. The red line is the fitted regression slope, which shows a slightly positive relationship and yields an R^2 of .058. Due to the right-skewed distributions, both variables were square root transformed (see Figure 5.1b). This yields a slightly steeper regression slope that has a better fit with an R^2 of .113, meaning that the expected counts explain 11.3% of the variance in the observed counts. This R^2 serves as the baseline from V_1. The data points do not scatter closely to the regression line, suggesting that the quality of expected counts is rather low.

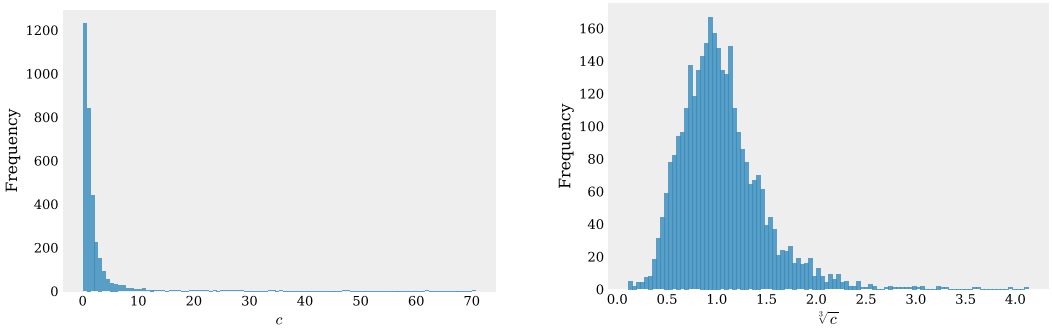


(a) Raw values. (b) √-transformed values.

Figure 5.1. Validation analysis V_1: Relationship between observed and expected traffic counts. Red line resembles estimated regression slope.

Figure 5.2a shows the distribution of the calibration factor C. Most values are close to 1 (i.e., perfect expectation), but the data is right skewed with a long tail ranging up to around 70. This makes it difficult to inspect the range of values below 1, which indicate overestimation. Figure 5.2b shows the distribution of C after cube root transformation. Since C strictly takes positive values, this transformation pushes all values towards 1. This allows us to keep 1 as the point of reference, while inspecting the distribution around 1. In this histogram, we can see that the mode of the distribution is indeed

around 1. Although there are more outliers above 1, it appears as if the area below 1 has a slightly higher density, suggesting that overestimating is more common than underestimating traffic.



(a) Raw values. (b) $\sqrt[3]{}$-values.

Figure 5.2. Distribution of calibration factor C (ratio of observed to expected count per road segment).

Figure 5.3 shows how C is distributed across the Dutch road network. C was mapped to a diverging color scale, with white indicating a value close to 1, red indicating a value larger than 1 (underestimation) and blue indicating a value smaller than 1 (overestimation). The clustering of blue colors in the triangle of Utrecht, Rotterdam, and Amsterdam suggests that this area particularly suffers from overestimation of expected counts, hinting at a regional bias. This corresponds to the clustering of outliers in this area as observed in figure 4.3a. The overestimation could be due to the dense public transportation network in this area, which might incentivize more commuters to use public transport than in other areas. It is also possible that commuters choose other modes than cars in dense areas to avoid traffic congestion. In general, blue colors appear to be more likely in areas where the road network has a higher density. On the other hand, for the majority of road segments close to the national border, the traffic counts were underestimated. The expected traffic counts are based on the assumption that all traffic is caused by commuters that were covered in the OD matrix. Because the matrix only covers employees working and living in the Netherlands, commuters from and to neighboring countries can cause unexpected traffic.

5.2 Results of Modeling the Calibration Factor C

Table 5.1 shows the results of the linear regression and the random forest regression.

In the first linear regression model LM 1, only the expected counts were included as input. Using the predicted $\hat{C}$ from this model in validation analysis V_2 shows that this explains 10.5% of the variance in the observed counts of the test set. This is slightly less than the baseline of 11.3%. By including segment characteristics in LM 2, the explained variance in V_2 becomes almost twice as large. Adding regional and network characteristics and the province raises the explained variance to 26.1% in LM 5. The $RMSE$ in the train and test set are close, indicating that the model does not suffer from overfitting.

The first random forest model RF 1 can calibrate the expected counts such that they explain 15.8% of the variance in the observed counts. Once segment characteristics are added in RF 2, the percentage becomes more than twice as large with 38.5%. By adding regional and network characteristics in RF 3 and RF 4, the R^2 in V_2 is raised to 41.2% and

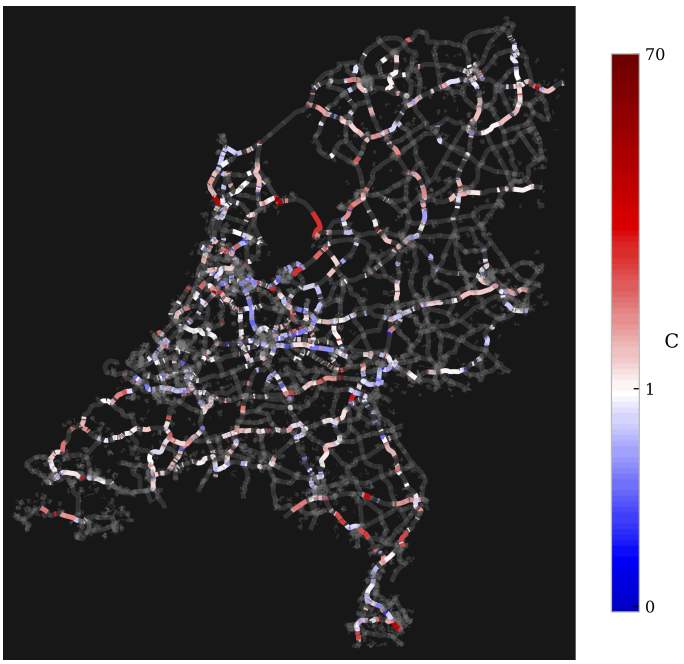


Figure 5.3. Quality of expected traffic counts on Dutch road network (C < 1 indicates overestimation, C > 1 indicates underestimation).

42.6%, respectively. However, adding the province does not lead to an improvement. In all random forest models, the $RMSE$ of the train and test set are comparable and close, indicating that the models are not overfitting to the training data.

All random forest models outperformed their corresponding linear regression models in terms of the explained variance in observed counts after calibration. Both models suggest that segment characteristics have the highest importance to model C, once expected counts are accounted for. This is due to the fact that the increase in explained variance was highest between model 1 and 2 in both cases. The increase is considerably larger in the random forest regression, but the increase from models 2 - 4 is similar for both models. Unlike the linear regression model, the random forest model does not seem to profit from the province. This suggests that the random forest model captures all the necessary information from the input sets of RF 4. RF 4 raised the explained variance in the observed counts to 42.6%, which is almost four times the baseline level of 11.3% from V_1. Because this model had the best performance with the fewest variables, it will be used as the final model to calibrate the expected counts.

The difference in performance between linear regression and random forest regression might be due to two reasons: 1) There is a non-linear component in the relationship between C and the input variables. 2) Assumptions of linear regression are not met by the data. For example, although C was cube root transformed, figure 5.2b shows that outliers are still present. We decided to keep these because they are of particular interest when correcting expected counts. However, model diagnostics of LM 5 showed signs of heteroskedasticity in the errors, which could both be a result of the outliers and of a non-linear relationship.

Another source of error might be multicollinearity. Because it is plausible to consider that some of the input variables might be correlated (e.g., population density and the number of segments), we inspected the variance inflation factor (VIF). Most variables are unproblematic, but the VIF of the variable for the number of road segments in the

Table 5.1. Comparison of model performance. Model outcome = $\sqrt[3]{c}$

Linear regression

	LM 1	LM 2	LM 3	LM 4	LM 5
Input	Exp	Exp Segment	Exp Segment Regional	Exp Segment Regional Network	Exp Segment Regional Network Province
$rmse_{train}$	0.304	0.284	0.281	0.279	0.274
$rmse_{test}$	0.317	0.302	0.298	0.298	0.296
R^2 in V_2 (explained var in OBS_{test})	0.105	0.206	0.235	0.241	0.261

Random Forest Regression

n_trees = 500, max_features = $\sqrt{n_{features}}$, min_split = 15, min_leaf = 8

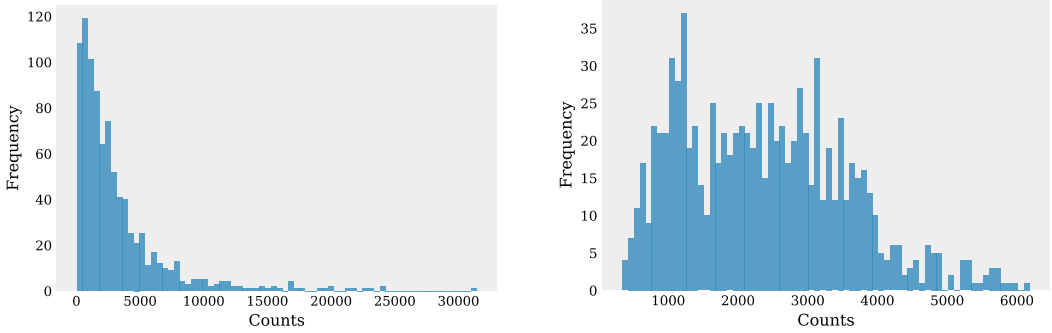
	RF 1	RF 2	RF 3	RF 4	RF 5
Input	Exp	Exp Segment	Exp Segment Regional	Exp Segment Regional Network	Exp Segment Regional Network Province
$rmse_{train}$	0.254	0.215	0.198	0.192	0.191
$rmse_{test}$	0.284	0.239	0.235	0.233	0.233
R^2 in V_2 (explained var in OBS_{test})	0.158	0.385	0.412	0.426	0.426

surrounding road network indicates moderate multicollinearity.

Appendix C.1 shows model results for LM 5 and the permutation importance of variables for RF 5. Importance measures for the random forest show that the five most important variables are the expected count, the motorway road type, the number of lanes, the population density and the number of edges in the local network. The linear regression coefficients can indicate how input variables are related to C. For example, an increase in the number of lanes or in the population density is associated with an increase in $\sqrt[3]{C}$. However, coefficients should be interpreted with caution due to the violations mentioned in the previous paragraphs. Direct comparisons to the random forest regression should also be made with caution, because input variables could have a different form of relationship in the random forest regression.

5.3 Expected Traffic Counts after Calibration

Figure 5.4a and 5.4b show the distributions of observed and calibrated expected counts in the test set using the final model RF 5. We can see that the expected counts do not suffer from the big spread to extreme outliers as they did before calibration (see figure 4.2a). Instead, they resemble the observed counts more closely now.



(a) Observed counts in test set. (b) Calibrated expected counts in test set.

Figure 5.4. Comparison of observed and expected counts after calibration step in test set.

Figure 5.5 shows the observed counts on the y-axis and the expected counts on the x-axis after they were calibrated with $\hat{C}$ from RF 4, which is the result of V_2 . Data was square root transformed. The red line resembles the estimated regression slope which has a moderate fit, with an R^2 of .43.

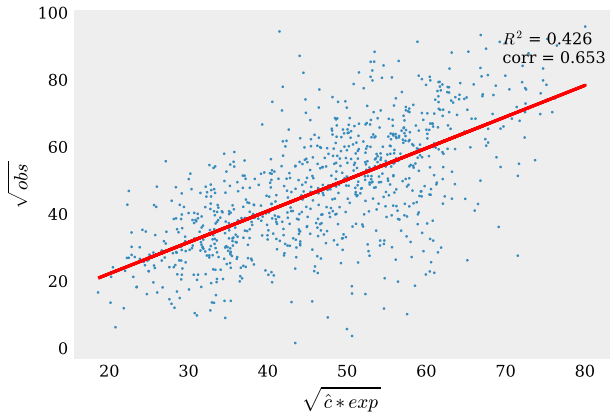


Figure 5.5. Validation analysis V_2: Relationship between observed and calibrated expected counts ($\sqrt{}$-transformed) using RF 4 on the test set. Red line resembles estimated regression slope.

Figure 5.6 shows the quality of the expected counts on the road network after calibration (C^*) for the test set. An improvement is clearly visible, as most road segments now are white or very pale colored. The problem of overestimation has dropped strongly compared to 5.3: The highest value of C^* is 5, compared to 70 in the case of C. Additionally, red colors are less frequently spotted in this map than in figure 5.3. Apparently, the problem of overestimation still remains as strongly for some road segments, because C^* has a minimum value close to 0. However, deep blue colored sensors, which represent a c^* that approximates 0, cannot be spotted on the map. This indicates that the problem of extreme overestimation only seems to affect very few segments. A regional bias cannot be identified as directly from a visual inspection alone.

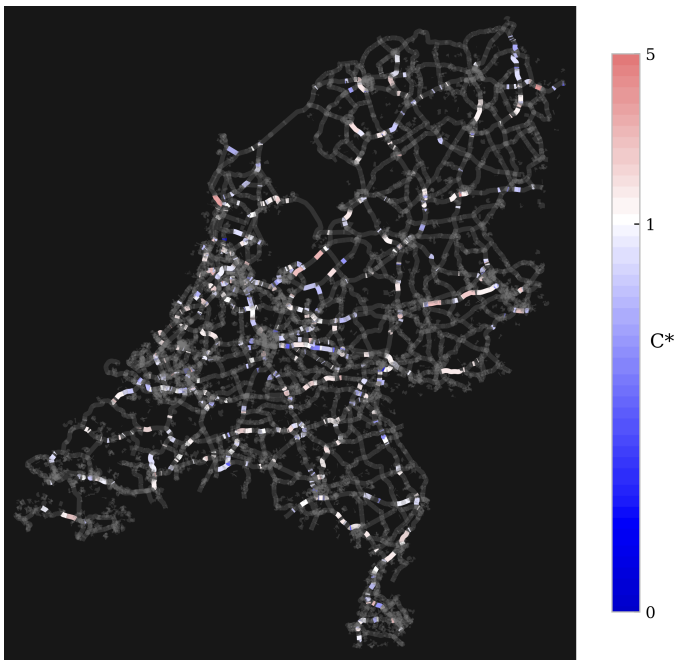


Figure 5.6. Quality of expected traffic counts on Dutch road network after calibration (C* < 1 indicates overestimation, C* > 1 indicates underestimation).

6 Discussion

The DaCiMob framework uses administrative data for traffic modeling on a nationwide scale. Two main findings come out of this study. First, this study shows that traffic loop sensor data can be utilized to validate expected traffic counts. Second, the results suggest that a calibration could improve the quality of expected traffic counts. This shows that the inclusion of traffic loop sensor data and the calibration model into the DaCiMob framework is profitable. Further, the results show that features of the road segment, network, and region play a large role in explaining and improving the quality of traffic predictions that are based on administrative data.

The random forest model was significantly better at this task. Due to the difficulty in interpreting random forests, this leaves room for further investigation of associations between input variables and the calibration factor. Some limitations arise from the complexity of the DaCiMob framework, which includes many steps. First, many assumptions had to be made to realize the framework. For one, it was assumed that commuters use an intelligent navigation system to find the fastest way to work, and do not take detours. Before calibration, expected counts are based on the assumption that all traffic during rush hour is commuting traffic. After calibration, this assumption can be lifted, but it consequently becomes difficult to disentangle commuting from other travel motives. Another broad assumption is that the correction factor of a road segment does not change over time. Our study suggests that it is in fact time-varying. To verify this, the model should be trained on different time periods and auto-correlation of $\hat{C}$ from different models should be inspected. In the case of a stable C, auto-correlation would be high for all lags of $\hat{C}_t$. Seasonal differences in traffic counts suggest that C might also have a seasonal pattern, which could be taken into account by the calibration model. It was also assumed that the traffic loop sensors are representative for all roads on the road network. The reasoning behind placements of sensors in the road network is

unknown and might affect the results of this study. If factors such as expected variability in traffic and risks of queues were considered during sensor placement, a bias might be introduced due to a not missing at random effect. Weighting and stratification techniques, adding additional features and adding a time component in a multi-level model could be explored as potential solutions.

Second, linking multiple data sources also means that the error of each data source is included, and in some cases, might be amplified. For example, although observed counts from sensors are viewed as ground truth in this study, sensor data can also carry errors due to malfunctioning or issues in data storage (see Section 4.3). Further errors can occur during the projection of sensors onto the infrastructure, leading to an accumulation of errors over the steps of the framework. This could be investigated under an approach similar to the Total Survey Error Framework Biemer et al. (2017).

Third, changes earlier in the process of the framework might lead to very different results later in the process. For example, it is possible that the calibration model loses its value once a different travel mode model is employed to produce expected counts. Generally, this would be a favorable improvement, but it is difficult to pinpoint how much each piece in the framework contributes to the final result at this point.

Given that there are many unexplored possibilities in the building blocks of this framework, this can also be seen as a strength. Comparing different travel mode models might just improve the quality altogether, compared to the current baseline. Also, it is possible to create a custom segmentation of roads that optimally suits the location of sensors Tennekes and Puts (2018), which could be used to investigate whether there are errors in the linkage of sensors to OpenStreetMap data. The framework could also be extended by incorporating public transport data. For one, information about the local public transportation network as model input could be explored to improve calibration for road traffic. Second, observed data from public transport could be used to calibrate nationwide public transport expectations, as was done in the Rotterdam case study Gootzen et al. (2020). Here too, unobserved areas could be predicted with a calibration model. Further, road traffic and public transport expectations could be calibrated simultaneously, as it is plausible to assume that they are dependent on each other. An even more exhaustive approach could additionally include observed counts from bike usage.

The calibration factors resulting from the model could potentially be traced back through the DaCiMob framework and provide individual calibrations for each of the steps. If this is achieved, calibrated OD matrices can be obtained. It would be relevant to see how the calibration affects more subtle and derived statistics.

7 Conclusion

The proposed inclusion of traffic loop sensor data and a calibration model was found to significantly improve nationwide traffic estimates from administrative data. Visual inspections showed that the quality of expected traffic counts was initially good for many road segments, but can be volatile both regionally and locally, with huge bias in some cases. This leads to high uncertainty in estimates. A direct application for policy makers, e.g., in urban planning, would lead to wrong decisions, which are costly.

We compared two predictive models for calibration and showed that a random forest

regression can improve estimates, raising the explained variance in observed counts from 11.3% to 42.6%. This shows that even crude traffic prediction models with a high uncertainty can produce good predictions when multiple data sources are combined with observed data, yet room for improvement remains.

By validating and calibrating expected counts, this paper underlines that the DaCiMob framework presented in this paper is a promising tool for regional planners. For example, before a nationwide housing project is approved, the framework can be used to estimate the effects on traffic and consequently make adjustments or initiate additional projects to expand the road network.

This paper demonstrates the potential of data linkage in traffic estimation for official statistics. It shows how data from multiple sources can be combined to estimate traffic on a nationwide scale and validate estimations. Expanding the number of data sources in future research might improve estimations further.

References

Australian Bureau of Statistics (2021). Greater Melbourne. 2021 Census All persons QuickStats.

Bakker, B. F. M. (2012, February). Estimating the validity of administrative variables: *Validity of administrative variables*. *Statistica Neerlandica 66*(1), 8–17.

Barff, R., D. Mackay, and R. W. Olshavsky (1982, March). A Selective Review of Travel-Mode Choice Models. *Journal of Consumer Research 8*(4), 370.

Barth, M. and K. Boriboonsomsin (2008, January). Real-World Carbon Dioxide Impacts of Traffic Congestion. *Transportation Research Record: Journal of the Transportation Research Board 2058*(1), 163–171.

Bauer, D., G. Richter, J. Asamer, B. Heilmann, G. Lenz, and R. Kolbl (2018, June). Quasi-Dynamic Estimation of OD Flows From Traffic Counts Without Prior OD Matrix. *IEEE Transactions on Intelligent Transportation Systems 19*(6), 2025–2034.

Behrisch, M., M. Bonert, D. Hinkeldein, D. Krajzewicz, G. Kuhns, and Y.-P. Wang (2009). DELPHI – a joint web decision support application for real time traffic situation analysis and prognosis, information exchange and cooperation. In *Proceedings of ITS 2009*, pp. 6.

Bera, S. and K. V. K. Rao (2011). Estimation of origin-destination matrix from traffic counts: the state of the art. (49), 21.

Biemer, P. P., E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West (Eds.) (2017, January). *Total Survey Error in Practice*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Boeing, G. (2017, September). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems 65*, 126–139.

Boonstra, H. J., J. van den Brakel, and S. Das (2021). Modeling mobility trends - update based on new ODiN data and level breaks. Technical report.

CBS (2021). Dutch National Travel survey.

CBS (2022a, March). CBS StatLine. Regionale kerncijfers Nederland.

CBS (2022b). Population; key figures.

Condurat, M., A. M. Nicuţă, and R. Andrei (2017). Environmental Impact of Road Transport Traffic. A Case Study for County of Iaşi Road Network. *Procedia Engineering 181*, 123–130.

de Fabritiis, C., R. Ragona, and G. Valenti (2008, October). Traffic Estimation And Prediction Based On Real Time Floating Car Data. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, Beijing, China, pp. 197–203. IEEE.

Demographia (2016). Demographia World Urban Areas (Built Up Urban Areas or World Agglomerations). 12th Annual Edition.

Eagle, N. and K. Greene (2014). *Reality mining: using big data to engineer a better world*. Cambridge, Massachusetts: The MIT Press.

Falcocchio, J. C. and H. S. Levinson (2015). Concentration of Travel Demand in Space and Time. In J. C. Falcocchio and H. S. Levinson (Eds.), *Road Traffic Congestion: A Concise Guide*, pp. 39–51. Cham: Springer International Publishing.

Gootzen, Y. A. P. M., M. R. Roos, and B. O. Mussman (2020). Combining data sources to gain new insights in mobility. A case study. *CBS Working paper*(03-20), 18.

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant (2020, September). Array programming with NumPy. *Nature 585*(7825), 357–362.

He, Z., C.-Y. Chow, and J.-D. Zhang (2019, June). STCNN: A Spatio-Temporal Convolutional Neural Network for Long-Term Traffic Prediction. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pp. 226–233. Journal Abbreviation: 2019 20th IEEE International Conference on Mobile Data Management (MDM).

Heyns, E., S. Uniyal, E. Dugundji, F. Tillema, and C. Huijboom (2019). Predicting Traffic Phases from Car Sensor Data using Machine Learning. *Procedia Computer Science 151*, 92–99.

Hymel, K. (2009, March). Does traffic congestion reduce employment growth? *Journal of Urban Economics 65*(2), 127–135.

INRIX (2019, December). Congestion Costs U.K. Nearly £8 Billion in 2018. Technical report, INRIX, Inc.

Jordahl, K., J. Van Den Bossche, M. Fleischmann, J. McBride, J. Wasserman, A. G. Badaracco, J. Gerard, A. D. Snow, J. Tratner, M. Perry, C. Farmer, G. A. Hjelle, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, B. Ward, G. Caria, M. Taves, N. Eubank, Sangarshanan, J. Flavin, M. Richards, S. Rey, Maxalbert, A. Bilogur, C. Ren, D. Arribas-Bel, D. Mesejo-León, and L. Wasser (2021, October). geopandas/geopandas: v0.10.1.

Klingwort, J. and J. Burger (2021). Inferring network traffic from sensors without a sampling design. *CBDS Working paper*(02-21), 29.

Krajzewicz, D., J. Erdmann, M. Behrisch, and L. Bieker (2012). Recent Development and Applications of SUMO – Simulation of Urban MObility. pp. 11.

Krajzewicz, D., G. Hertkorn, C. Rössel, and P. Wagner (2002). SUMO (Simulation of Urban MObility) - an open-source traffic simulation. In A. Al-Akaidi (Ed.), *4th Middle East Symposium on Simulation and Modelling*, pp. 183–187.

Lana, I., J. Del Ser, M. Velez, and E. I. Vlahogianni (2018). Road Traffic Forecasting: Recent Advances and New Challenges. *IEEE Intelligent Transportation Systems Magazine 10*(2), 93–109.

Levy, J. I., J. J. Buonocore, and K. von Stackelberg (2010, December). Evaluation of the public health impacts of traffic congestion: a health risk assessment. *Environmental Health 9*(1), 65.

Ma, X., X. Hu, T. Weber, and D. Schramm (2021a, March). Evaluation of Accuracy of Traffic Flow Generation in SUMO. *Applied Sciences 11*(6), 2584.

Ma, X., X. Hu, T. Weber, and D. Schramm (2021b, January). Experiences with Establishing a Simulation Scenario of the City of Duisburg with Real Traffic Volume. *Applied Sciences 11*(3), 1193.

McCool, D., P. Lugtig, O. Mussmann, and B. Schouten (2021, March). An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. *Journal of Official Statistics 37*(1), 149–170.

McFadden, D. (1974, November). The measurement of urban travel demand. *Journal of Public Economics 3*(4), 303–328.

Melnikov, V. R., V. V. Krzhizhanovskaya, A. V. Boukhanovsky, and P. M. Sloot (2015). Data-driven Modeling of Transportation Systems and Traffic Data Analysis During a Major Power Outage in the Netherlands. *Procedia Computer Science 66*, 336–345.

Möller, D. P. (2014). *Introduction to Transportation Analysis, Modeling and Simulation*. Simulation Foundations, Methods and Applications. London: Springer London.

NDW (2020). Open Data Portaal.

NL Times (2019, November). Traffic jams cost businesses €1.4 billion last year.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter (2017, October). Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models. *Journal of the American Statistical Association 112*(520), 1477–1489.

OpenStreetMap (2021).

OpenTripPlanner (2022). Documentation.

Ortúzar S., J. d. D. and L. G. Willumsen (2011). *Modelling transport* (4th ed. ed.). Oxford: Wiley-Blackwell.

Pankowska, P., B. Bakker, D. L. Oberski, and D. Pavlopoulos (2018). Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. *Statistical Journal of the IAOS 34*(3), 317–329. Publisher: IOS Press.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Puts, M. J. H., P. J. H. Daas, M. Tennekes, and C. de Blois (2019). Using huge amounts of road sensor data for official statistics. *AIMS Mathematics 4*(1), 12–25.

Python Software Foundation (2020). Python.

Qu, L., W. Li, W. Li, D. Ma, and Y. Wang (2019, May). Daily long-term traffic flow forecasting based on a deep neural network. *Expert Systems with Applications 121*, 304–312.

Roos, M. and Y. Gootzen (2021). Projection of register- and big data sources on the mobility network [to be published in 2022].

Scholtus, S., B. Bakker, and A. Van Delden (2015). Modelling measurement error to estimate bias in administrative and survey variables. *CBS Discussion Paper* (17). Publisher: Statistics Netherlands.

Seabold, S. and J. Perktold (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Shafiei, S., Z. Gu, and M. Saberi (2018, August). Calibration and validation of a simulation-based dynamic traffic assignment model for a large-scale congested network. *Simulation Modelling Practice and Theory 86*, 169–186.

Shamshiripour, A., N. Golshani, R. Shabanpour, and A. K. Mohammadian (2019, October). Week-Long Mode Choice Behavior: Dynamic Random Effects Logit Model. *Transportation Research Record: Journal of the Transportation Research Board 2673*(10), 736–744.

Tcheumadjeu, L. C. T., S. Ruppe, M. Ali, and L. Bieker (2012). EmerT – Supporting traffic parameter estimation from low cost and low resolution uncalibrated web cameras. In *Proceedings CD ROM*, pp. 11.

Tennekes, M. and M. Puts (2018). Projection of road sensors to the Dutch road network. Technical report.

The pandas development team (2020, February). pandas-dev/pandas: Pandas.

U.S. Census Bureau (2021). American Community Survey one year estimates.

van Delden, A., S. Scholtus, and J. Burger (2016, September). Accuracy of Mixed-Source Statistics as Affected by Classification Errors. *Journal of Official Statistics 32*(3), 619–642.

Wang, F. and C. L. Ross (2018, December). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation Research Record: Journal of the Transportation Research Board 2672*(47), 35–45.

Wang, N., M. Gentili, and P. Mirchandani (2012, January). Model to Locate Sensors for Estimation of Static Origin–Destination Volumes Given Prior Flow Information. *Transportation Research Record: Journal of the Transportation Research Board 2283*(1), 67–73.

Wang, Z., X. Su, and Z. Ding (2021). Long-Term Traffic Prediction Based on LSTM Encoder-Decoder Architecture. *IEEE Transactions on Intelligent Transportation Systems 22*(10), 6561–6571.

Willumsen, L. (2021, January). Use of Big Data in Transport Modelling. International Transport Forum Discussion Papers 2021/05. Series: International Transport Forum Discussion Papers Volume: 2021/05.

Zhang, L.-C. (2012, February). Topics of statistical theory for register-based statistics and data integration: *Developing theory for data integration*. *Statistica Neerlandica 66*(1), 41–63.

Zhao, X., X. Yan, A. Yu, and P. Van Hentenryck (2020, July). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society 20*, 22–35.

Appendix

A Access to Data and Scripts

Traffic loop sensor data and infrastructure data are openly accessible at their aforementioned sources. Administrative and survey data is stored on secure servers at CBS and were not directly accessible for this study due to their sensitivity. Instead, CBS provided traffic count estimates that resulted from the methods described in section 4.4. Interested researchers can contact the infoservice for further details (cbs.nl/en-gb/about-us/contact/infoservice). All work of the authors of this paper was done in python v3.8.5 Python Software Foundation (2020). Data cleaning and management were done with the packages pandas v1.3.4 The pandas development team (2020) and NumPy v.1.21.4 Harris et al. (2020). Jupyter notebook scripts to reproduce the model can be found on https://github.com/iebos/dacimob. Specifically, Section can be resproduced with the script inspect_model_c.ipynb.

This study was approved by the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University under file number 21-2133.

B Removed edge types in E

secondary, tertiary, unclassified, residential, secondary_link, tertiary_link, living_street, service, pedestrian, track, bus_guideway, raceway, road, busway, footway, bridleway, steps, corridor, path, cycleway, elevator, emergency_bay, platform, User Defined

C Results from LM 5 and RF 5

Acknowledgments and Funding: The authors would like to acknowledge Marko Roos for help in working with the administrative data. Part of the work for this project was funded via the DACIMOB programme within Statistics Netherlands.

Data availability statement: All replication code, including instructions for how to access the not-publicly accessible administrative data used in this paper, can be found at https://github.com/iebos/dacimob

Table C.1. Coefficients from linear regression and permutation importance from random forest. N = 1,341

Linear regression

Input	Coefficient (SE)	P>|t|
Const	0.318 (0.098)	0.001
Exp	-0.000 (0)	0
Exp^{-1}	47.447 (1.944)	0
Motorway	0.204 (0.034)	0
Motorway_link	-0.098 (0.03)	0.001
Primary_link	-0.167 (0.128)	0.191
Trunk	0.055 (0.032)	0.09
Trunk_link	-0.125 (0.057)	0.03
Mixed Highwaytypes	0.046 (0.063)	0.46
Lanes	0.121 (0.034)	0
Population density	0 (0)	0.045
Border in 10km	-0.067 (0.03)	0.027
Nr of edges (12km)	-0.000 (0)	0.83
Max speed limit (12km)	0.003 (0.001)	0.021
Min speed limit (12km)	0.001 (0.001)	0.518
Max lanes (12km)	-0.024 (0.019)	0.209
Min Lanes (12km)	0.318 (0.098)	0.001
Max Lanes (1.5km)	0.052 (0.018)	0.003
Flevoland	-0.028 (0.055)	0.609
Friesland	-0.033 (0.053)	0.53
Gelderland	-0.06 (0.047)	0.197
Groningen	-0.121 (0.062)	0.052
Limburg	0.055 (0.053)	0.298
Noord-Brabant	0.121 (0.165)	0.462
Noord-Holland	-0.339 (0.142)	0.017
Overijssel	-0.054 (0.05)	0.277
Utrecht	-0.643 (0.25)	0.01
Zeeland	-0.005 (0.064)	0.935
Zuid-Holland	0.001 (0.174)	0.994
Population density · Motorway	-0.000 (0)	0
Population density · Noord-Holland	-0.000 (0)	0.358
Population density · Utrecht	0 (0)	0.935
Population density · Zuid-Holland	-0.000 (0)	0.009
Lanes · Max Lanes (1.5km)	-0.021 (0.007)	0.002
Max lanes 12km · Noord-Brabant	-0.017 (0.032)	0.594
Max lanes (12km) · Noord-Holland	0.069 (0.027)	0.01
Max lanes (12km) · Utrecht	0.106 (0.04)	0.008
Max lanes (12km) · Zuid-Holland	0.026 (0.03)	0.389

Random forest

Input	Permutation importance (normalized)
Exp	1.522 (1)
Motorway	0.153 (0.102)
Maxlanes	0.035 (0.025)
Population density	0.024 (0.017)
Nr of edges (12km)	0.012 (0.009)
Motorway_link	0.009 (0.007)
Max lanes (1.5km)	0.005 (0.005)
Noord-Holland	0.003 (0.004)
Trunk_link	0.003 (0.003)
Trunk	0.002 (0.003)
Border in 10km	0.001 (0.002)
Noord-Brabant	0.001 (0.002)
Overijssel	0 (0.002)
Min speed limit (12km)	0 (0.002)
Limburg	0 (0.002)
Utrecht	0 (0.002)
Drenthe	0 (0.002)
Max speed limit (12km)	0 (0.002)
Mixed Highwaytypes	0 (0.002)
Groningen	0 (0.002)
Friesland	0 (0.002)
Flevoland	0 (0.002)
Primary	0 (0.002)
Primary_link	0 (0.002)
Zeeland	0 (0.001)
Gelderland	0 (0.001)
Zuid-Holland	-0.001 (0.001)
Max lanes (12km)	-0.002 (0)

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information