Discussion Paper

Predictive inference for non-probability samples: a simulation study

2015 | 13

Bart Buelens
Joep Burger
Jan van den Brakel

Non-probability samples provide a challenging source of information for official statistics, because the data generating mechanism is unknown. Making inference from such samples therefore requires a novel approach compared with the classic approach of survey sampling. Design-based inference is a powerful technique for random samples obtained via a known survey design, but cannot legitimately be applied to non-probability samples such as big data and voluntary opt-in panels. We propose a framework for such non-probability samples based on predictive inference. Three classes of methods are discussed. Pseudo-design-based methods are the simplest and apply traditional design-based estimation despite the absence of a survey design; model-based methods specify an explicit model and use that for prediction; algorithmic methods from the field of machine learning produce predictions in a non-linear fashion through computational techniques. We conduct a simulation study with a real-world data set containing annual mileages driven by cars for which a number of auxiliary characteristics are known. A number of data generating mechanisms are simulated, and—in absence of a survey design—a range of methods for inference are applied and compared to the known population values. The first main conclusion from the simulation study is that unbiased inference from a selective non-probability sample is possible, but access to the variables explaining the selection mechanism underlying the data generating process is crucial. Second, exclusively relying on familiar pseudo-design-based methods is often too limited. Model-based and algorithmic methods of inference are more powerful in situations where data are highly selective. Thus, when considering the use of big data or other non-probability samples for official statistics, the statistician must attempt to obtain auxiliary variables or features that could explain the data generating mechanism, and in addition must consider the use of a wider variety of methods for predictive inference than those in typical use at statistical agencies today.

Keywords: big data; pseudo-design-based estimation; predictive modelling; algorithmic inference

Contents

1. Introduction

With the emergence of big data as a potential source of official statistics, there is increased attention for estimation procedures for non-probability samples. Big data are sometimes referred to as found data, reflecting that they happen to be available but were not originally intended for statistical purposes. Big data sources typically describe large subgroups in great detail, making them potentially interesting for small area estimation with high precision. Unlike administrative registers, however, big data sources typically do not cover the entire population, making unbiased estimation of population parameters precarious. The data generating mechanism is often unknown and is generally very different from random sampling.

Examples of big data sources that have been discussed in the literature include internet search behavior (Ginsberg et al. 2009), social media and mobile phone metadata (Daas et al. 2015). While the data sets in these examples are large, they are generated by a subset only of the entire population. The examples mentioned concern respectively users of the Google search engine, active Twitter users and mobile phone users. None of these groups coincide with a population of interest to official statistics. Such data sets are said to be selective and can lead to biased estimates when basing inference on them (Buelens et al. 2014). An additional problem may be conceptual differences between quantities measured in big data sets and variables of interest to the statistician. We do not address such measurement bias here, and concentrate on selection bias specifically.

In the present article predictive inference methods are investigated as a technique of removing non-random selection bias. Three classes of predictive methods are considered: pseudo-design-based methods, which proceed as if the data set was generated through random sampling; model-based methods, which formulate explicitly some statistical model; and algorithmic methods, which are popular in data-mining and machine-learning communities.

A simulation study is conducted based on real data from the Online Kilometer Registration (OKR) system in the Netherlands. Various non-probability data selection scenarios are implemented, and the performance of the predictive inference methods is assessed by comparing the estimation results to the—in this case—known population totals.

The main risk of basing inference on selective big data sources is biased estimates. The extent to which predictive methods can correct such bias is crucially dependent on the availability of auxiliary variables that can be used as auxiliary data in the models and algorithms. The algorithmic approaches are more flexible than the pseudo-design-based and traditional model-based methods and should be considered in real world settings, in particular when the relations between auxiliary and target variables are complex and non-linear. In certain situations however equally good results can be obtained through pseudo-design-based methods.

Section 2 covers a brief history of inference in official statistics and the emergence of models in the last decades. Methods of predictive inference used in the presented study are presented in section 3. Section 4 discusses the simulation setup and the OKR data source. The main results are shown in section 5, with additional results in the supplement. Section 6 draws final conclusions.

2. History

National statistical institutes are mandated by law to publish statistical information about economic and social developments of a society. This information is generally referred to as official statistics and it is often defined as totals, means or proportions at the national level as well as breakdowns in various subpopulations. The finite population values for these variables are unknown. Until the beginning of the twentieth century this kind of information was obtained by a complete census of the target population. The concept of random probability sampling has been developed, mainly on the basis of the work of Bowley (1926), Neyman (1934) and Hansen and Hurwitz (1943) as a method of obtaining valid estimators for finite population parameters based on a modest but representative sample, rather than on a complete census.

National statistical institutes traditionally use probability sampling in combination with design-based or model-assisted inference for the production of official statistics. This refers to estimation procedures that are predominantly based on the randomization distribution of the sampling design. This means that statistical properties of the estimator, like expectation and variance, are derived under the concept of repeatedly drawing samples from a finite population according to the sample design while keeping all other parameters fixed. Statistical modelling of the observations obtained in the survey does not play any role so far. Under this approach, an estimator of unknown population totals is obtained as the sum over the observations in the sample, expanded with the so called design weights. These weights are constructed such that the sum over the weighted observations is a design-unbiased estimate of the unknown population total and are obtained as the inverse of the probability that a sampling unit is included in the sample. In sampling theory this is a well-known estimator and is called the Horvitz-Thompson estimator, Narain (1951), and Horvitz and Thompson (1952).

National statistical institutes often have auxiliary information about the target population from external sources. The precision of the Horvitz-Thompson estimator can be improved by taking advantage of this auxiliary information. One way is to improve the efficiency of the sampling design, for example through stratified sampling with optimal allocation or sampling with unequal inclusion probabilities proportional to the size of the target variable. Another way is to use this auxiliary information in the estimation procedure via the so called general regression estimator proposed by Särndal et al. (1992) or calibration estimators (Deville and Särndal 1992). These estimators adjust the design-weight of the HT estimator such that the sum over the weighted auxiliary variables in the sample equates the known population totals. In the model-assisted approach, developed by Särndal et al. (1992), this estimator is derived from a linear regression model that specifies the relationship between the values of a certain target variable and a set of auxiliary variables for which the totals in the finite target population are known. Most estimators known from sampling theory can be derived as a special case from the general regression estimator.

The general regression estimator has two very attractive properties. Although this estimator is derived from a linear model, it is still approximately design unbiased. If the underlying linear model explains the variation of the target parameter in the population reasonably well, then the use of this auxiliary information will result in a reduction of the design variance compared to the Horvitz-Thompson estimator and it might also decrease the bias due to selective non-response, Särndal and Swenson (1987), Bethlehem (1988), and Särndal and Lundström (2005). Model misspecification might result in an increase of the design variance but the property that

this estimator is approximately design unbiased remains. From this point of view, the general regression estimator is robust against model misspecification. The linear model is only used to derive an estimator that uses auxiliary information but the resulting estimator is still judged by its design-based properties, such as design expectation and design variance. This is the reason that this approach is called model assisted.

Design-based and model-assisted inference are very powerful concepts that are still used in modern statistical science because:

1) It allows drawing valid inference of unknown variables of a large population based on a relatively small but representative sample.
2) Uncertainty of using an estimator of the unknown population total can be measured by calculating the design variance of this estimator.
3) The precision of the estimator can be improved by taking advantage of auxiliary information in the design of the sample or in the estimation procedure.

Model-based inference procedures for finite population parameters have also been developed. They rely on the probability structure of an explicitly assumed statistical model, whereas the probability structure of the sampling design plays a less pronounced role. Royal (1970) proposed the prediction approach, where a model is assumed for the realized values of the finite population. The observations in the sample are used to fit the model and predict the values of the unobserved units that are not included in the sample. To avoid selection bias, sampling features can be incorporated in the model. An extensive treatment of the predictive modelling approach is given by Vaillant, Dorfman and Royal (2000).

The major drawback of the model-based approach is that model misspecification can result in poor inference. Hansen, Madow and Tepping (1983) show that even small model misspecification in a large sample can result in spurious inference. This is one of the major reasons that design-based and model-assisted modes of inference are traditionally used by national statistical institutes. For decades, there has been the prevailing opinion that official statistics must be free from model assumptions, since model misspecification easily translates into wrong statements about the variable of interest.

Design-based and model-assisted inference procedures, however, also have some limitations. A major drawback is that they have large design variances in the case of small sample sizes and do not handle measurement errors effectively. In such situations model-based estimation procedures can be used to produce more reliable estimates, see Rao (2011) for an appraisal. In situations where the sample size within the sub populations is too small to produce reliable estimates with design-based or model-assisted procedures, explicit model-based procedures can be used to increase the effective sample size within the separate domains using cross-sectional sample information from other domains or temporal sample information from preceding periods. This is often referred to as small area estimation, Rao (2003), Pfeffermann (2002, 2013). Model-based procedures are also required to account for non-sampling errors, like selective nonresponse and measurement errors. See Rao (2011) for a more complete discussion.

There is a persistent pressure on national statistical institutes to reduce administration costs and response burden. This must be accomplished by using register data like tax registers, or other large data sets that are generated as a by-product of processes unrelated to statistical

production purposes. Examples are data available from mobile phone companies and social media like Twitter and Facebook, as well as Google trends, to describe all kinds of developments that can be derived from activities on the Web. A recent related development is the increasing use of non-probability samples, like opt-in panels observed through the Web. In areas like marketing research this type of survey sampling is popular since voluntary opt-in samples are cheap and have a fast turnaround time.

A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. A challenging problem in this context is to use this kind of data for the production of official statistics that are representative of the target population. There is no randomized sampling design that facilitates the generalization of conclusions and results obtained with the available data to an intended larger target population. As a result, the traditional design-based inference framework is not appropriate to these situations. Baker et al. (2013) evaluate available inference procedures for non-probability samples. One approach is to apply weighting and calibration techniques, originally developed for design-based inference in probability sampling, to construct weights that correct for possible selection bias. This is referred to as the pseudo-design-based approach. An alternative approach is the aforementioned model-based prediction approach developed by Royal (1970). The extent to which both approaches successfully remove selection bias depends on the extent to which the available auxiliary information, used to construct pseudo-design weights or model-based predictions, explains the data generating process. Baker et al. (2013) mention that probability sampling is not the only way of making valid statistical inference about intended target populations. In many other scientific areas advanced statistical methods are used to draw inference from data that are not obtained through probability samples. A class of techniques that can be used are algorithmic inference procedures known from the area of machine learning (Hastie et al. 2009). In this paper, the use of algorithmic inference procedures, like neural networks and support vector machines as well as the pseudo-design-based and predictive inference approach are investigated in a simulation study.

3. Methods

3.1 Predictive inference

Methods of inference are aimed at providing an estimate of some population quantity based on a partial observation of that population. The units for which observations are available constitute the sample, which may be a non-probability sample. Many texts discuss inference in general. This section is mainly based on Valliant et al. (2000) and Hastie et al. (2009).

For a variable Y, the quantity of interest G_Y is generally a function $G_Y = G(y_i)$ for $i = 1, \dots, N$, with y_i the values of Y for each of the N units in the population. Common functions G are the sum and the mean. The y_i are known for the units in the sample S, and unknown for the units in the remainder of the population, R. Through predictive inference, an estimate of G_Y is obtained as

$$\hat{G}_Y = G(\{y_i\}_{i \in S} \cup \{\hat{y}_i\}_{i \in R}),$$

with $\hat{y}_i$ predictions—or estimates—of the values of Y for units not in the sample. The variance of the estimate $\hat{G}_Y$ depends on the uncertainty associated with the unit predictions $\hat{y}_i$ and can be quantified through bootstrapping. Any method capable of producing the unit predictions can be used to arrive at the estimate $\hat{G}_Y$.

Such methods typically use the observed Y values of the sample to establish appropriate values for the unobserved part of the remainder of the population. In addition, auxiliary characteristics X known for the observed and the unobserved parts of the population are often used. In data-mining and machine-learning contexts these are referred to as attributes or features. Prediction methods ordinarily contain some parameters β that need to be estimated. Henceforth, a prediction method F is written in general terms as

$$\hat{y}_i = F(x_i, \hat{\beta}) \text{ for } i \in R,$$

with the estimates for β obtained from the sample through some procedure P,

$$\hat{\beta} = P_F(x_i, y_i) \text{ for } i \in S,$$

which is commonly referred to as model fitting, training or learning. In addition, many methods require some prior optimization in that choices must be made about some specific properties of the method. This will be clarified below for the methods under consideration.

Predictive inference is rarely perfect. First, the prediction method *F* may be inappropriate for the data at hand or make wrong assumptions. Second, even if *F* is perfect in principle, the estimation of its parameters is subject to error. Both types of error are carried forward into the predictions $\hat{y}_i$ which in turn give rise to an error in the estimated quantity of interest, $\hat{G}_Y$. For some specific methods—typically those stemming from sample survey theory—it may be possible to calculate MSE estimates by means of analytical formulae. The accuracy of the population parameter estimated using algorithmic methods typically does not have an analytical expression. Fortunately, the total error can also be quantified through a bootstrap approach.

A bootstrap sample is drawn from the original sample through simple random sampling with replacement, giving rise to estimates $\hat{\beta}^{(b)}$, $\hat{y}_i^{(b)}$ and $\hat{G}_Y^{(b)}$. Repeating this *B* times, the bootstrap Mean Square Error is defined as

$$MSE(\hat{G}_Y) = \frac{1}{B}\sum_{b=1}^{B}(\hat{G}_Y^{(b)} - G_Y)^2 \,,$$

which can be decomposed as

$$MSE(\hat{G}_Y) = \; Bias(\hat{G}_Y)^2 + Var(\hat{G}_Y),$$

with

$$Bias(\hat{G}_Y) = \hat{G}_Y^B - G_Y,$$
$$Var(\hat{G}_Y) = \frac{1}{B}\sum_{b=1}^{B}(\hat{G}_Y^{(b)} - \hat{G}_Y^B)^2,$$

and

$$\hat{G}_Y^B = \frac{1}{B}\sum_{b=1}^{B}\hat{G}_Y^{(b)}.$$

Assessment of the results presented in this study is conducted using relative root mean square error (RRMSE),

$$RRMSE\left(\hat{G}_Y\right) = \frac{\sqrt{MSE(\hat{G}_Y)}}{\hat{G}_Y}.$$

In practice, G_Y is unknown and the bias cannot be obtained. In simulation studies like the one presented in this article, G_Y is known and the (RR)MSE, bias and variance are easily obtained and can be used as indicators of the performance of different methods of predictive inference.

The methods considered in this article are described below. For each method an abbreviation is given in parentheses in the section titles, and will be used in subsequent chapters.

3.2 Sample mean (SAM)

While a sample at hand may not have been obtained through some sampling design—or the design is unknown to the analyst—it can be regarded as the result of a simple random sampling design, where units have equal inclusion probabilities. This is a simple but naïve approach which can always be conducted. In a predictive inference context, predictions for unobserved units are equal to the mean of the observed units,

$$\hat{y}_j = \bar{y} = \frac{1}{n}\sum_{i\in S} y_i,$$

for all units $j \in R$ and n the number of elements in S. The SAM approach is the only method considered here that does not make use of auxiliary auxiliary variables X. When one or more X variables are available, more sophisticated methods can be applied. These are discussed in the sections below.

3.3 Pseudo-design-based estimation (PDB)

Methods proceeding as if the sample is obtained through some complex sampling design while that is not the case are known as pseudo-design-based methods (Elliott 2009, Baker et al. 2013). The SAM approach discussed above is a special case of PDB where the design is simple random sampling and no auxiliary variables X are utilised. In this article the term PDB is used only for situations in which at least one auxiliary variable is available. The auxiliary characteristics X are used to form strata as is commonly done for post-stratification in survey sampling settings.

For a given data set, the initial optimization consists of defining the post-strata based on the available auxiliary characteristics X. A balance is sought in which the strata differentiate between subgroups and at the same time contain sufficiently large numbers of sample units.

In the predictive inference framework given above, the model fitting or learning phase consists of calculating strata means for the sample:

$$\bar{y}_h = \frac{1}{n_h}\sum_{i\in h\cap S} y_i,$$

with n_h the number of observed units in stratum h. With the strata sample means available, the Y values for the unobserved units are predicted by their strata means,

$$\hat{y}_j = \bar{y}_h \; \forall j \in h \cap R.$$

For the specific case of G_Y defined as the total of Y, the pseudo-design-based prediction estimator is equal to the sum of the observed and the predicted values,

$$\hat{G}_Y = \sum_h \left(\sum_{i\in h\cap S} y_i + \sum_{j\in h\cap R} \hat{y}_j \right) = \sum_h (n_h\bar{y}_h + (N_h - n_h)\bar{y}_h) = \sum_h N_h\bar{y}_h$$

with N_h the number of population units in stratum h. This expression can be written as

$$\hat{G}_Y = \sum_{i\in S} w_i y_i,$$

with weights w_i given by

$$w_i = \frac{N_h}{n_h} \; \forall i \in h \cap S.$$

Estimating the total of Y by the weighted sum of the observed values with stratum weights as defined above is exactly the expression of a design-based post stratification estimator for simple random sampling designs. Hence the name, pseudo-design-based estimator.

3.4 Generalised linear models (GLM)

A generalised linear model expresses a function of the dependent variable Y as a linear combination of predictors X. In the present setting, the auxiliary characteristics or features provide the predictors,

$$g(E(Y_i)) = \beta x_i,$$

where the function g is known as the link function. If g is the identity function, the model is said to be linear. Common choices for g include the inverse, logarithmic and logistic functions. Texts on generalised linear models are plentiful; Gelman and Hill (2006) is an example.

Optimising this approach for a given data set boils down to establishing the exact model terms to be included in the model specification, and a choice for g. For example, interaction terms or higher order terms may or may not be included.

In the model fitting phase, the coefficients β are estimated from the sample through a standard method such as Ordinary Least Squares (OLS) or Maximum Likelihood (ML). The estimated coefficients allow for prediction of Y values for unobserved units,

$$\hat{y}_j = g^{-1}(\hat{\beta}x_j).$$

3.5 K-nearest neighbours (KNN)

The k-nearest neighbour method (Hastie et al. 2009) predicts the Y value of unseen units by averaging the k nearest observed Y values. This requires some distance function or metric for population units. Typically the X space is used for this purpose with a common distance function such as the (weighted) Euclidean distance.

Optimisation in this case encompasses the choice of X, an appropriate distance metric, and the choice of k. In the learning phase distances are calculated between each unseen unit and all observed sample units. The algorithm proceeds by finding the set A_j of size k for each unobserved unit j, such that A_j contains the k nearest neighbours of j obtained from the sample S. The prediction for the Y value for the unobserved units equals the mean of the observed k nearest neighbours,

$$\hat{y}_j = \frac{1}{k}\sum_{a\in A_j} y_a.$$

3.6 Artificial neural networks (ANN)

An artificial neural network, or neural network in short, is an algorithmic method supposedly mimicking the working of the human brain in a simplified manner. It arrives at a prediction of some variable Y by propagating inputs—in this case the predictors X—through a network of artificial neurons (nodes) laid out in a layered network structure. Each node in the network applies a multiplicative weight to all its input and a so-called activation function to the weighted sum. The output of a node in one layer serves as input to the nodes in the next layer. This is known as a feed-forward neural network. Hastie et al. (2009) provide an overview of neural network methods.

In the present article neural networks with a single hidden layer are considered as these are among the simplest types of networks. The number of nodes in the input layer is determined by the dimension of the feature space X, and that of the output layer by the dimension of the variable Y, which is one in the case of regression. In an initial optimisation routine, the appropriate number of hidden nodes is established and an activation function is chosen.

Learning or training involves establishing the weights and any parameters that may occur in the activation functions in the nodes. Standard back propagation (Hastie et al. 2009) is used as the learning method. The weights and any other parameters are repeatedly adjusted in small steps so as to minimize the networks prediction error.

Finally, the neural network prediction is obtained algorithmically by propagating the X values of some unit *j* with unknown *Y* value through the trained network, arriving at a predicted value

$$\hat{y}_j = ANN(x_j, \hat{\beta}),$$

where the vector $\hat{\beta}$ includes all parameters that need to be established during the learning phase.

3.7 Regression trees (RTR)

Originally developed for classification, tree-based algorithms can be used for regression too (Breiman et al. 1984; Hastie et al. 2009). A binary tree splits the data in two subsets at each node according to some criterion typically maximizing the variance between the two groups. Starting from a single root node, a data set is split into two parts which are taken down branches of the tree to two other nodes, which in turn split the data again. Some stopping criterion is applied to decide whether data at a particular node get split—if not, the node is said to be a leaf node.

An initial optimisation routine is conducted to establish an appropriate value for the stopping criterion, expressed as a percentage of improvement that is required for the total prediction error. If a split does not improve the total error by this percentage, the node is not split further. In addition, one may choose to specify a minimum number of data points required at leaf nodes.

The sample is used in the learning phase to construct the regression tree. The result is a particular layout of nodes and branches, and at each node the splitting rule expressing how to split a data set arriving at that node.

A regression tree prediction is obtained for a unit with unknown Y value by taking it down the tree to a leaf node. The predicted value for that unit is the mean of the sample units in that node with known Y values,

$$\hat{y}_j = RTR\left(x_j, \hat{\beta}\right) = \frac{1}{n_\lambda} \sum_{k \in \lambda \cap S} y_k,$$

with $\hat{\beta}$ the parameters characterising the regression tree, λ the leaf node to which unit j is assigned, n_λ the number of sample units from S that are assigned to node λ.

While history and background are very different, the regression tree approach and the pseudo-design-based methods are similar to some extent. This is easily seen by considering the collection of leaf nodes of a regression tree as a stratification. While strata are defined manually by the analyst using the pseudo-design-based method, they are constructed algorithmically when using the regression tree approach. In both cases the predicted values are the known sample means of the strata or leaf nodes respectively.

3.8 Support vector machines (SVM)

A support vector machine is an algorithm projecting input data into a higher dimensional space, and applying linear—hence simple—algorithms there (Vapnik 1996; Hastie et al. 2009). In the original space, the SVM can predict non-linear behaviour. Originally, support vector machines were developed for classification tasks. Using support vector machines for regression is referred to as support vector regression. When an input vector x is projected into another space—typically of higher dimension—a vector machine seeks to approximate the target variable Y through a linear function in the projected space,

$$y = b + w.\rho(x),$$

with $\rho(x)$ the projection of x, and b and w parameters to be estimated from the data. The second term in this expression is the dot-product of the vectors w and $\rho(x)$.

For a particular choice of ρ, optimal values for b and w are determined through a procedure known as ε-insensitive loss optimization, essentially developed to reduce computation time for very large data sets: data points that are less than a distance ε away from the regression line are ignored. The other data points do contribute—loosely speaking in an OLS-type of approach—to the estimation of b and w. When the optimization problem is infeasible, deviations larger than ε are tolerated. Smola and Scholkopf (2004) provide details of the estimation method.

Centring the data prior to conducting SVM removes the need for the parameter b. It has been shown (Smola and Scholkopf, 2004) that the solution of SVM regression can then be written as

$$\hat{y} = \hat{w}.\rho(x) = \sum_i \hat{\alpha}_i(\rho(x_i).\rho(x)),$$

where summation is over all data points contributing to the estimate of $\hat{w}$, also known as the support vectors, and $\hat{\alpha}_i$ scalars. From this formula it is seen that the projection as such does not need to be performed to arrive at predicted values $\hat{y}$, as the only quantity needed is the product of the vector $\rho(x)$ with each of the $\rho(x_i)$. Furthermore, the $\hat{\alpha}_i$ too are obtained from pairwise products of projected input data points. Hence, the explicit projections are not needed, only their product with other projected points is ever required. To this end, the product of the projected vectors is defined as a dedicated function, called the kernel function K,

$$K(x_i, x_j) = (\rho(x_i).\rho(x_j)),$$

with K a function that must fulfil certain regularity conditions (Smola and Scholkopf, 2004).

The choice of the kernel is an optimisation that needs to happen prior to applying the support vector machine. Typical kernel functions include Gaussian, polynomial and sigmoid functions. If the kernel is the identity function, the support vector machine is linear, in other cases it is non-linear.

Once learned, the support vector machine produces predictions for given x_j values,

$$\hat{y}_j = \sum_i \hat{\alpha}_i K(x_i, x_j),$$

which requires evaluation of the kernel function at each combination of the input x_j and all support vectors x_i.

4. Simulation

4.1 General setup

A major concern of non-probability samples is that they cover only a selective portion of the target population and that the selection mechanism is unknown. This can lead to biased estimates if methods of inference fail to correct for the selectivity in an appropriate manner.

To test the performance of the inference methods described in the previous section, several non-probability samples are constructed from a data set from the Online Kilometer Registration (see Section 4.2). Different selection mechanisms are implemented, each depending on one or more variables and resulting in data sets that mimic non-probability samples with selective coverage (Section 4.3). Different scenarios are simulated in which certain auxiliary variables are or are not available (Section 4.4). After optimizing the methods of predictive inference (Section 4.5), they are applied to all samples under the different scenarios (Section 4.5).

This setup mimics real-world situations that are encountered where data sets are available but their data generating mechanisms are unknown. The results of the simulation and the performance of the various methods are discussed in Section 5.

4.2 The Online Kilometer Registration

For our simulation we consider vehicles in the Online Kilometer Registration (OKR) as our target population. The OKR is a facility provided by the government Agency for Road Transport (ART) and allows entry of vehicle odometer readings into a central database. When visiting a service station or motor vehicle test center, the car's registration number and odometer reading are entered into OKR together with the date. While primarily aimed at detecting odometer fraud through reversing the meter, the OKR database is also used for statistical purposes. From the odometer readings annual mileages are derived, which are used to estimate the total mileage for all vehicles registered in the Netherlands. The register of all vehicles is maintained by the ART and also contains auxiliary characteristics, such as registration year, vehicle weight, fuel type, and name and address of the owner, which can be a company or a private person. For privately owned vehicles the date of birth of the owner is available as well.

From the reported odometer readings, daily mileages are computed in a straightforward manner by distributing the total mileage between two readings evenly over the days in between. Annual mileages are obtained by aggregation of the daily mileages of all the days in a year. For the purpose of this study, annual mileages of the year 2012 are used. The mileage is taken to be the target variable Y.

Annual mileages cannot be computed for vehicles for which no suitable set of odometer readings is available. Circumstances in which this can occur include: the last reported reading is before the end of the year of interest; missing or erroneous readings due to technical problems of the OKR system; missing or erroneous readings due to human factors such as failing to enter a reading into the system or making a typing mistake when doing so.

In reality, annual mileages for vehicles for which they are not available must be estimated. For the purpose of the simulation study only those vehicles are used that have a valid 2012 mileage. This subset is called the population in the present article.

4.3 Generating non-probability samples

Both qualitative and quantitative features correlating with the target variable (mileage) are used to create selective non-probability samples (see Supplementary Fig. S1 for an example of the correlation between an auxiliary variable and the target variable).

The qualitative or categorical features include:

- registration year: year in which vehicle was first registered; eight 1-year bins for the most recent vehicles, three 2-year bins for older vehicles, and one bin for vehicles registered in or before 1998
- legal form: privately owned or company car
- vehicle weight: four weight categories
- fuel: petrol, diesel or other (including electricity and gas)

The quantitative or numeric features include:

- registration year: year in which vehicle was first registered; not grouped; considered a numeric variable
- owner age: the age of the vehicle owner in 2012

Selectivity with respect to qualitative variables is studied in a population data set A, consisting of 7.6 million vehicles for which all four qualitative variables are available and which have valid 2012 annual mileages obtained from the OKR (Table 1: population A).

Typically a non-probability sample starts recording at a certain point in time but does not contain any information about past events. We hypothesized that numerical features are more powerful to estimate beyond the observed range of values than categorical features. To test this, we created a subpopulation of population A containing two numerical features: registration year and age of owner (Table 1: population B). Since the age of the owner of the vehicle only applies to privately owned cars, this population consists of 6.7 million privately owned vehicles for which the age of the vehicle owner is known. Population B is used to generate cut-off samples that are selective with respect to the continuous variables, which are only observed over a certain range.

Selective samples are generated according to the following scheme:

- Population A: All vehicles (N=7,582,065)
    - 1: young vehicles overrepresented
    - 2: vehicles of juridical persons overrepresented
    - 3: heavy vehicles overrepresented
    - 4: young, heavy vehicles of juridical persons overrepresented
    - 5: high-mileage vehicles overrepresented
    - 6: low-mileage vehicles overrepresented
- Population B: Vehicles of natural persons with known age (N=6,670,509)
    - 7: representative sample of young (2003–2012) vehicles of owners aged 17–64, i.e. excluding elderly people and older vehicles

- 8: representative sample of high-mileage (≥ 10000 km) vehicles only
- 9: representative sample of young (2003–2012) vehicles only
- 10: representative sample of owners aged 17–64, i.e. excluding the elderly

In sample 1, younger vehicles are overrepresented by increasing the inclusion probability with registration year (compare the probability mass functions in Supplementary Fig. S2). In the next three samples, we overrepresent vehicles of juridical persons (sample 2), heavier vehicles (sample 3) or younger and heavier vehicles of juridical persons (sample 4). In sample 5, vehicles with high annual mileage were overrepresented by increasing the inclusion probability with annual mileage. Here the data are missing not at random, because the propensity for a data point to be missing depends directly on its value. In sample 6, vehicles with low annual mileage were overrepresented. This sample was introduced to distinguish bias correction from underestimation.

Samples 7–10 are generated from population B. In sample 7, a representative sample was drawn but exclusively of vehicles up to ten years old, owned by persons under 65 (Supplementary Figs. S3 and S4). In the remaining samples a representative sample was drawn but exclusively of vehicles with an annual mileage of 10,000 km or more (sample 8), of vehicles up to ten years old (sample 9) or of vehicles owned by persons under 65 (sample 10).

4.4 Scenarios

The scenarios simulate which auxiliary information is available to make inference about the target population, given a non-probability sample (Table 1). In the simplest case, the information causing the missingness is also available as auxiliary variables (*complete*: scenarios 1a, 2a, 3a, 4a, 7a, 9a, 10a). For example, in scenario 1a the missingness is caused by registration year, and registration year is also available to compare its distribution between the sample and the population. The data are thus missing at random. More realistically, we only have *indirect* information that correlates with the target variable (scenarios 1b, 2b, 3b, 4e, 5a, 6a, 8a, 9b, 10b). For example, in scenario 1b the missingness is caused by registration year, but only legal form, vehicle weight and fuel type are available to compare their distribution between the sample and the population. The data are thus missing not at random.

In Supplement 8.3, we additionally discuss cases where we have some but not all information causing the missingness (*partial*: scenarios 4b–d). For example, in scenario 4b young, heavy company cars are overrepresented, but only registration year is available to model the missingness while no information is available on legal form or vehicle weight. In the same Supplement we also discuss cases where we have *extra* information in addition to the information causing the missingness (scenarios 9c, 10c). For example, in scenario 9c the missingness is only caused by registration year, but both registration year and the age of the vehicle owner are available as auxiliary information.

Some methods, such as KNN, ANN and SVM, are developed for numerical variables. In case categorical variables were available as covariates, these were transformed to a numerical scale by numbering the categories. To make variables with different number of categories comparable, each variable was scaled between 0 and 1 by subtracting the smallest number and dividing by the difference between the largest and smallest number. For example, vehicle

weight classes {0–850 kg, 851–1150 kg, 1151–1500 kg, 1501+ kg} would be transformed to $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$.
Conversely, continuous numerical variables are made categorical by binning: the range of values is split into a limited number of intervals, and each interval is considered a category. This is required for the PDB method.

4.5 Optimization

Each inference method requires choices about its properties (see Section 3). These properties were chosen through the following optimization procedure. A sample was drawn with replacement from the original non-probability sample. This bootstrap sample was randomly split into a training set (70%) and a test set. For a range of likely input parameter values, a model was trained using the training set, model predictions were made using the auxiliary information in the test set, and the mean square error (MSE) was calculated by comparing the predicted values with the observed values in the test set. This was repeated for ten bootstrap replicates. For each input parameter value, the MSE was averaged over the ten bootstrap replicates. The input parameter value that gave the lowest average MSE was chosen to train the model using the entire non-probability sample and to predict the annual mileage of the vehicles in the remainder of the population. An example is given in Supplementary Fig. S5.

The input parameters that need optimization take different forms depending on the inference method. For the SAM and PDB methods there is nothing to optimize. For the GLM, we optimized the model by comparing the average MSE between candidate models including all possible interactions and—in case of numerical variables—quadratic terms. For the KNN, we optimized the number of neighbours k, where k was varied by powers of two. For ANN, we optimized the number of nodes in the (single) hidden layer. For the RTR, we optimized the complexity parameter: the data in a node are split until the model fit is not improved by a factor of this parameter. For the SVM, we optimized the kernel function, where we tried linear, polynomial, radial and sigmoid functions. For each scenario, the resulting optimal input values for the five models are given in Table S1 in Supplement 8.1.

4.6 Predictive inference

The methods presented in Section 3 are applied in all scenarios. While the samples are split to determine the optimal parameters and settings of the methods, they are used entirely to train the models for inferential purposes. Inference is made for the non-sample part of the population as if it is unknown. In this simulation study the non-sample part is known and is used to assess the bias and variance of the predictions, as discussed in Section 3. Variance estimates are obtained through bootstrapping by sampling with replacement from the original sample. In the present study we conducted 100 bootstraps, which was sufficient to discern substantial differences between the methods In real-world applications convergence of the bootstrap distribution should be monitored and stopping criteria defined.

4.7 Implementation

This simulation study is implemented in the statistical computing environment R (R Core Team, 2014). The packages *ff* (Adler et al. 2014) and *ffbase* (de Jonge et al. 2014) are used to handle big data files. The figures are produced using the *ggplot2* package (Wickham 2009). A number of packages are used for the estimation methods: *survey* (Lumley 2004) for PDB, *FNN*

(Beygelzimer et al. 2013) for KNN, *nnet* (Venables and Ripley 2002) for ANN, *rpart* (Therneau et al. 2014) for RTR and *e1071* (Meyer et al. 2014) for SVM.

Table 1 Population A: all vehicles (N=7,582,065), categorical features

Sample	Scenario	Sample size		Registration year *eerste_toel_dat_1* Ordinal	Legal form *srt_ref* Nominal	Vehicle weight *afl_gewkl* Ordinal	Fuel type *afl_brankl* Nominal	Annual mileage *jaarkm_A* Ratio
1		137,351	Selectivity	✓				
	a		Availability (complete)	✓				
	b		Availability (indirect)		✓	✓	✓	
2		15,477	Selectivity		✓			
	a		Availability (complete)		✓			
	b		Availability (indirect)	✓		✓	✓	
3		40,368	Selectivity			✓		
	a		Availability (complete)			✓		
	b		Availability (indirect)	✓	✓		✓	
4		48,288	Selectivity	✓	✓	✓		
	a		Availability (complete)	✓	✓	✓		
	b		Availability (partial)	✓				
	c		Availability (partial)		✓			
	d		Availability (partial)			✓		
	e		Availability (indirect)				✓	
5		25,467	Selectivity					✓ (high)
	a		Availability (indirect)	✓	✓	✓		
6		151,248	Selectivity					✓ (low)
	a		Availability (indirect)	✓	✓	✓		

Table 1 (*continued*) Population B: vehicles of natural persons with known age (N=6,670,509); numerical features

Sample	Scenario	Sample size		Registration year *eerste_toel_dat_1* Ratio	Age of vehicle owner *leeftijd* Ratio	Annual mileage *jaarkm_A* Ratio
7		26,394	Selectivity	✓ (cut-off)	✓ (cut-off)	
	a		Availability (complete)	✓	✓	
8		34,982	Selectivity			✓ (cut-off)
	a		Availability (indirect)	✓	✓	
9		34,066	Selectivity	✓ (cut-off)		
	a		Availability (complete)	✓		
	b		Availability (indirect)		✓	
	c		Availability (extra)	✓	✓	
10		53,809	Selectivity		✓ (cut-off)	
	a		Availability (complete)		✓	
	b		Availability (indirect)	✓		
	c		Availability (extra)	✓	✓	

5. Results

In this section the results of the study are presented. A discussion and interpretation follow in Section 6. The simulation results are separated into two parts depending on the type of auxiliary variables available. For samples 1 through 6 the auxiliary variables are categorical, for samples 7 through 10 they are numerical. The main results are included here, Supplement 8.2 contains additional outcomes.

5.1 Categorical auxiliary variables

The results of the simulation are presented graphically in Figures 1 and 2 for sample 1. Similar plots for samples 2 through 6 are included in Supplement 8.2 (Figs. S6–10).

In Fig. 1 the bootstrap results for sample 1 are given using boxplots. The known population level is indicated by the solid horizontal line: 13,109 km. Thus, the difference between the boxplot median and the horizontal line is an indication of the bias, and the boxplot height an indication of the variance (the exact definitions of bias and variance are given in Section 3.1). The left panel shows the results for the scenario in which the variable explaining the selectivity—in this case registration year—is also available, and is used as a covariate in the models. The right panel shows the results for the scenario in which the variable explaining the selectivity—registration year—is not available, but other variables—legal form, vehicle weight and fuel type—are available and used as covariates (see Table 1 for the scenario definitions).

It is clearly seen in Fig. 1 that the sample mean (SAM) is biased. All other methods, aimed at correcting the bias, are successful and perform roughly equally well, apart from SVM, which is less successful at removing all bias in this case. In the scenario where the covariate explaining selectivity is used (left panel), all bias can be removed. In the other scenario bias can be removed to some degree, but not completely. The extent to which they are successful depends on the correlation between the indirect variables and that explaining the bias.

Fig. 2 shows the root relative mean squared error (RRMSE) associated with the results in Fig. 1. Since the bootstrap variance is comparable for all methods it is the bias that dominates the RRMSE. Of the methods capable of removing bias—PDB, GLM, KNN, ANN and RTR—PDB and RTR have the smallest RRMSE in the scenario with covariates explaining the selectivity available (left panel). When only indirect covariates are available (right panel), RTR compares slightly favorably to the other methods. Nevertheless, the scenario (complete versus indirect information) determines the accuracy more than the inference method.

The results for the other samples with categorical variables (2 through 6) are very similar as for sample 1 and lead to the same conclusions (Figs. S6–S10 in Supplement 8.2). Note that all methods also remove bias when vehicles with low rather than high annual mileage are overrepresented (Fig. S10), suggesting that they do not simply lower the population parameter estimate.

Only the PDB, GLM and RTR methods are naturally appropriate for categorical data. The KNN, ANN and SVM methods are in fact developed for continuous variables. The latter methods are based on an appropriate metric in the covariate space allowing for measuring distances between data points. Applying these methods to categorical data is awkward and counter intuitive. Nevertheless, they still perform rather well, apart from the SVM.

As mentioned in Section 2, the RTR method can be seen as a flexible form of the PDB method. In the simulations here they both perform well for categorical variables.

The recommendation based on these findings is to use the RTR method for the data at hand. It has the advantage of resulting in a stratification—meaning a subdivision of the data set in cells defined by categorical auxiliary variables—that is optimal in the sense that the between-strata variance is maximized. This will generally lead to estimates with lower variance than using an ad-hoc and potentially suboptimal stratification in the PDB method. Furthermore, stratifications that can be achieved naturally using RTR are sometimes awkward to implement in a PDB context, for example complex cross-classifications and interactions of multiple categorical variables.

An approach sometimes followed with continuous variables is to create categorical variables based upon them, by binning the values into a limited number of bins or cells, using for example quantiles of the distribution to define bin boundaries. Regression trees can also be used as a means to define optimal boundaries between such bins. Bins defined in this way provide an optimal categorization of a continuous variable.

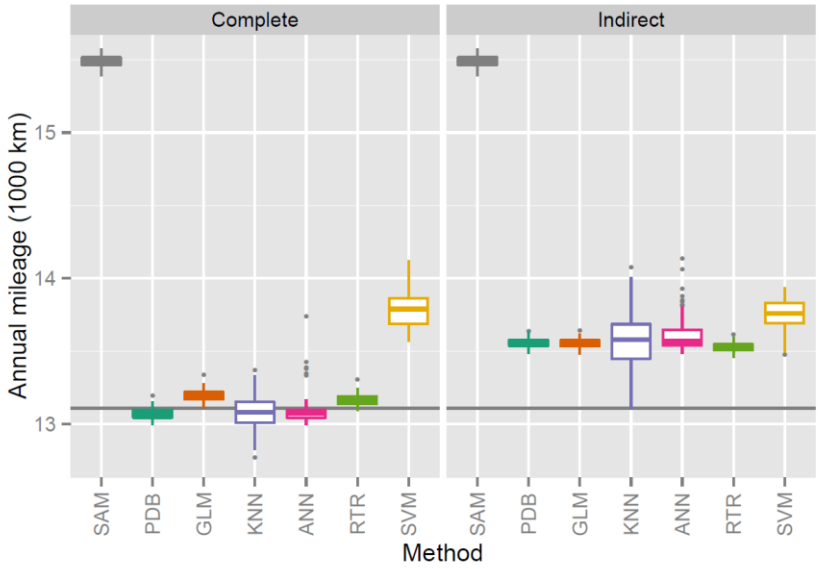

Figure 1 Effect of inference method on predicted mean annual mileage (boxplot of 100 bootstrapped predictions), given sample 1 (young vehicles overrepresented) and (Left) complete information (registration year) or (Right) indirect information (legal form, vehicle weight and fuel type) available for prediction. Horizontal line is true population level. Note that the y-axis does not start at 0.

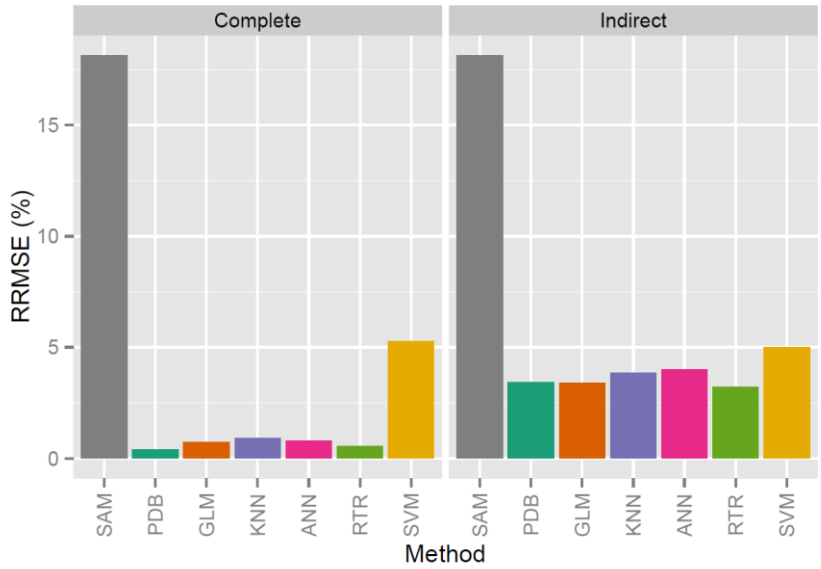

Figure 2 Effect of inference method on relative root mean square error, given sample 1 (young vehicles overrepresented) and (Left) complete information (registration year) or (Right) indirect information (legal form, vehicle weight and fuel type) available for prediction.

5.2 Continuous auxiliary variables

For the cut-off samples 7 through 10 we use continuous variables as covariates. Here sample 9 is discussed in detail, since it is selective in one (registration year) but not the other (age of owner) variable and hence provides a relevant example. Figures 3 and 4 show the results for sample 9. Supplement 8.2 (Figs. S11–S13) contains similar plots for the other samples. Note that the true population level (11,735 km) is lower than in the previous section because population B only contains privately owned vehicles, for which the age of the owner is known.

The left panel of Fig. 3 shows the results for sample 9 when the variable explaining selectivity is available, in this case registration year. Three methods perform better than the others in terms of removing bias: GLM, ANN and SVM. The PDB approach results in estimates comparable to the sample mean (SAM) and is essentially unable to remove bias. The KNN and RTR methods are not very successful either. From the height of the boxes it is clear that in terms of variance the ANN method stands out as having the largest variance. Despite this, the RRMSE of the ANN method is still lower than that of the KNN and RTR methods, which in turn are lower than the SAM and PDB approaches; see left panel of Fig. 4. The GLM and SVM have the lowest, and almost equal, RRMSEs.

The right panels of Figs. 3 and 4 show the results of the scenario in which variables are available that only indirectly explain the selectivity. None of the methods perform well. Even the SAM approach is not worse than the other methods, in this case. This is due to the weak correlation between the available variable (age of owner) and the variable explaining selectivity (registration year). It could be expected that with stronger correlating variables the results would resemble those presented in the previous section.

Fig. 5 shows the predicted values for sample 9 in the scenario where the variable explaining selectivity is available (registration year). Fig. 5 corresponds to the results in Fig. 3 but now shows the predictions at the level of registration year rather than for the population as a whole. The black line indicates the average mileages in the population; the solid line marks the sample,

the dotted line the missing part, which is the target of inference. Predictions using the various methods are shown by solid colored lines.

Two methods that perform well for categorical variables—PDB and RTR (Figs. 1 and 2)—do not perform well for continuous variables, even when variables explaining the selectivity are available (left panels of Figs. 3 and 4; Fig. 5). The advantage of the GLM, ANN and SVM methods for continuous variables is their capacity to produce predictions outside the range of the values observed in the training data set, as seen in Fig. 5. The SAM, PDB, KNN and RTR methods use (weighted) averages of available observations and are incapable of producing predictions that are very different from available observations. On the other hand, the GLM, ANN and SVM methods can extrapolate outside the range of available observations and are therefore more powerful in predicting missing values for strongly selective data sets. In the samples at hand, entire parts of the population range of values are missing, and prediction outside the range of the sample is necessary to obtain unbiased results. The success of such approaches depends on the models or algorithms fitted to the data at hand to be applicable to unseen data too. In real-world situations this assumption must be made.

While GLM, ANN and SVM all produce lower estimates for the unknown part of the population than the other methods, none of the three are very precise. Some details in particular are impossible to predict, such as the bump around registration year 1987 (Fig. 5). This is caused by vehicles aged 25+, considered as old-timers under Dutch law, to which beneficial taxation rules apply. As a side effect of this, some people deliberately use such old cars even though they drive considerable distances. Hence, while there is an explanation for the bump, it is not possible to predict it without variables that would explain its presence.

The results for samples 7, 8 and 10 are given in Figs. S11–S13 of Supplement 8.2. When the sample contains only vehicles up to ten years old owned by persons under 65, but both registration year and age of owner are available as covariates (Fig. S11), GLM and SVM still outperform the other methods. Although ANN also effectively removes the bias, its variance is exceptionally large. When the sample contains only vehicles owned by persons under 65, but age of owner is available as a covariate (Fig. S13, left panel), GLM and ANN outperform the other methods. In this scenario, SVM is less effective in removing the bias. With extra information available for prediction, the performance of ANN declines, whereas the performance of SVM improves and may even outperform GLM (compare Figs. 4 and S15, and Figs. S13 and S16). ANN is possibly more sensitive to overfitting than SVM.

GLMs perform well and are recommended in situations in which model evaluations can be conducted to establish validity of the assumed models. ANN and SVM techniques are more flexible and less restrictive in their assumptions, which may be beneficial in certain circumstances when the more rigid models underpinning the GLM approach do not hold. In the present simulation, it was found that quadratic terms were necessary for the GLM to produce unbiased results. In general, model and algorithm selection needs to be conducted to establish the most suitable method of inference.

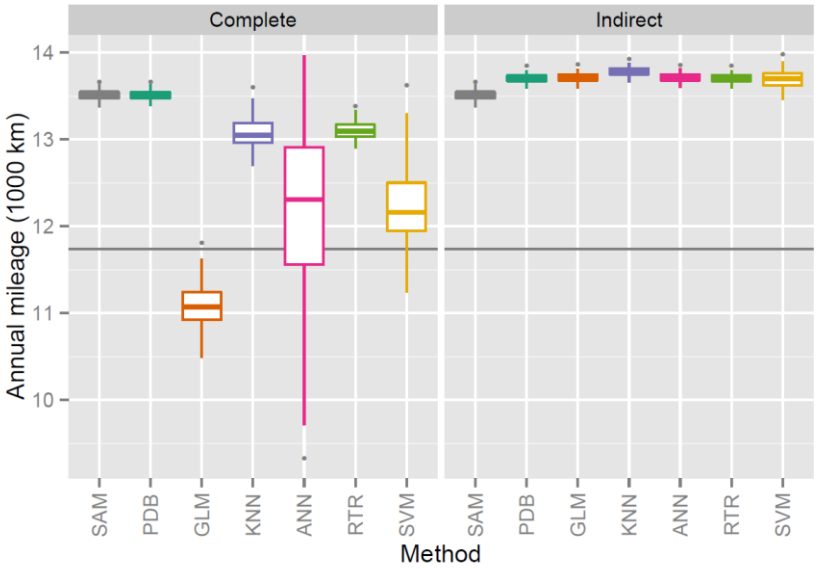

Figure 3 Effect of inference method on predicted mean annual mileage (boxplot of 100 bootstrapped predictions), given sample 9 (only vehicles up to ten years old) and (Left) complete information (registration year) or (Right) indirect information (age of owner) available for prediction. Horizontal line is true population level. Note that the y-axis does not start at 0.

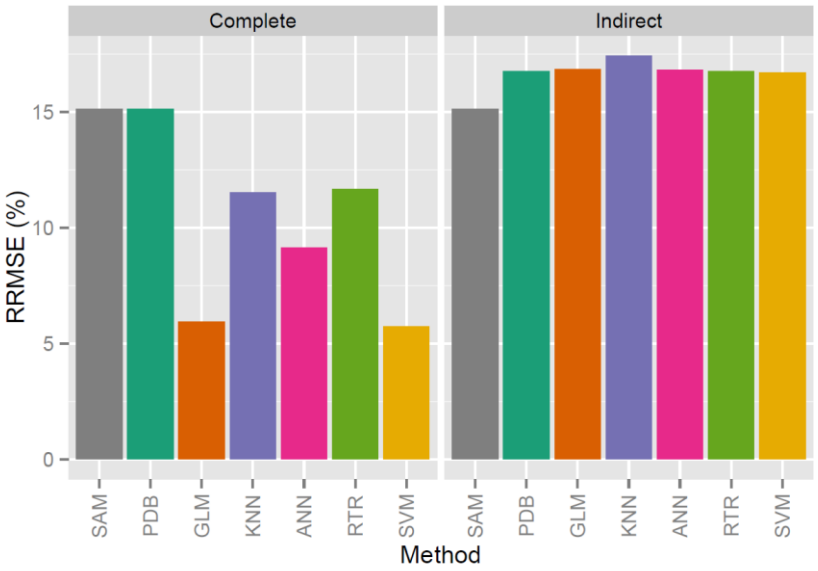

Figure 4 Effect of inference method on relative root mean square error, given sample 9 (only vehicles up to ten years old) and (Left) complete information (registration year) or (Right) indirect information (age of owner) available for prediction.

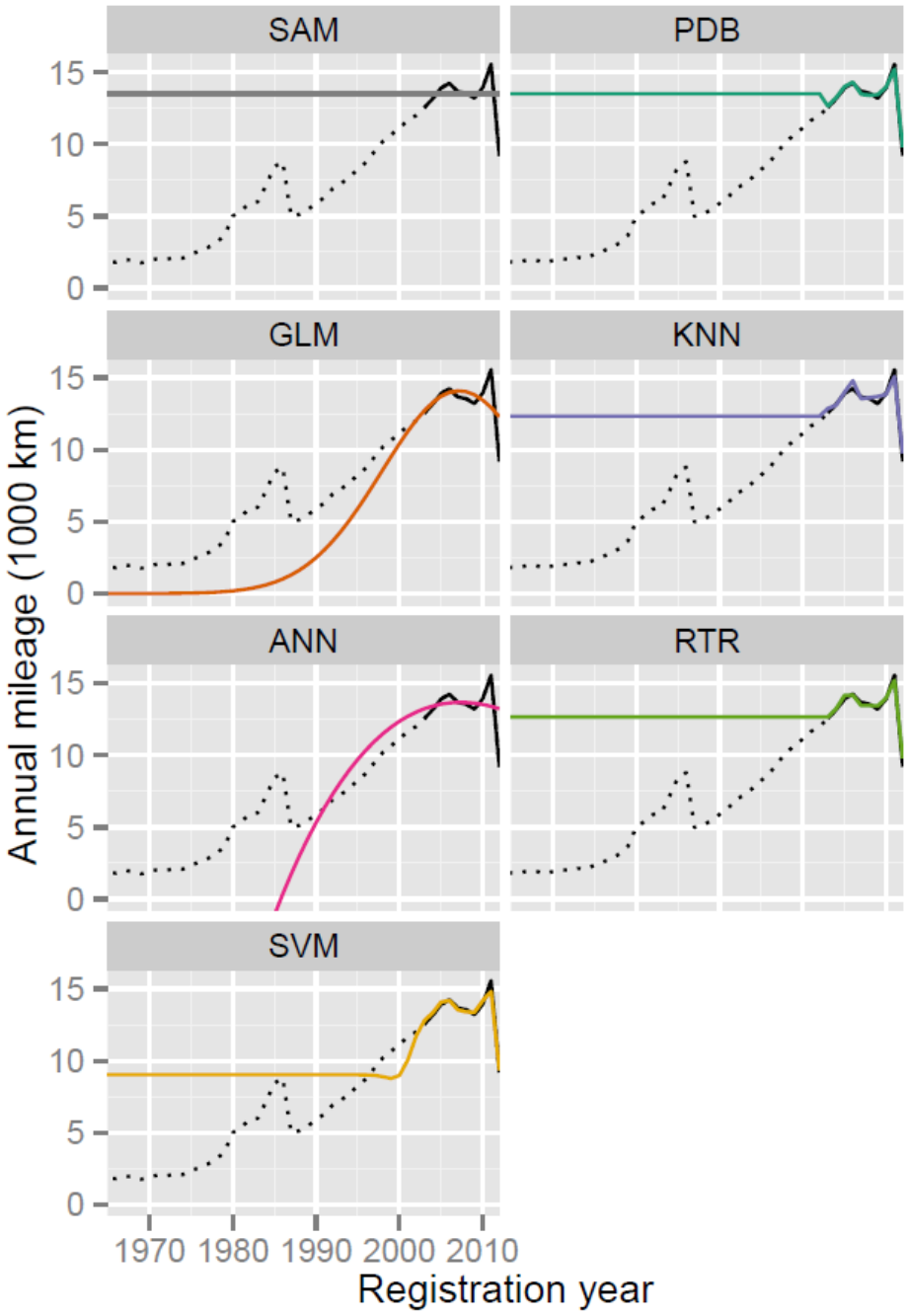

Figure 5 Mean annual mileage in population B (black) and predicted (color) by inference method based on sample 9 (only vehicles up to ten years old; solid black line) and complete information (registration year) available for prediction. The dotted line is the target for inference.

5.3 Consequences for official publications

While the OKR data are used here as the basis for simulation and comparative study of selection mechanisms and inference procedures, it is worthwhile to briefly discuss consequences of the results of this study for OKR-based statistics routinely produced at Statistics Netherlands. The official statistics based on OKR are totals and means of annual mileages of vehicles, by vehicle type, registration year, legal form of ownership, weight class and fuel type.

In the simulation study the population is defined as the passenger cars for which 2012 mileages are available. In reality, mileages are missing for approximately 10%-15% of the true population of cars. Mileages are missing when the available odometer readings are not sufficient to calculate an annual mileage. This is the case for new cars which do not have an odometer reading yet (other than zero when they are first registered); for old-timers (cars aged 25+) which are not required to have annual motor vehicle tests; for errors made when entering the odometer readings in the OKR system—either accidentally or fraudulently: implausible sequences of odometer readings are not used for statistics production.

At present, inference proceeds using the PDB method using a stratification defined by the publication variables, which are the four covariates in Table 1. The results of section 5.1 suggest that this as an acceptable approach. However, the data generating mechanism and the variables determining the missingness for the real data are unknown and may be different from the simulated regimes. From the possible reasons for missing mileages mentioned above, it is seen that mileages are missing in particular for very new cars, for old-timers and for cars that have been subject of fraud. It could be the case that the mileages that are available for these groups are selective and that this selectivity is not fully explained by the four covariates under consideration.

Since the share of cars with unknown mileages is relatively small it can be expected that the use of methods of inference other than PDB, or additional covariates, will not substantially affect the current estimates of means and totals for the whole population. Nevertheless, it would be valuable to conduct further analyses for the real situation, investigating the effect additional covariates might have on the predictions, in particular for the specific subpopulations mentioned. Many additional covariates can be obtained by linking the OKR data base to other administrative registers maintained at Statistics Netherlands, from which extra variables can be sourced with respect to the owner of the car including employment position, income, household size, value of house, etc. Such analyses may prove beneficial especially if statistics about subpopulations with large proportions of missing mileages are required.

Another approach to improving mileage statistics is a reconsideration of the derivation of annual mileages from the odometer readings. In the present study the annual mileages are assumed fixed, known, and without error. However, this derivation must be subject to error or uncertainty as well, as various assumptions are made. This perspective is not addressed in the present study but may warrant further research.

6. Conclusions

When considering non-probability samples for use in official statistics, the methods used for inference should be adapted to this setting. In this article a range of prediction methods are proposed that may fit this purpose. A simulation study is conducted using real-world data for which non-random selection regimes are simulated, mimicking data generating mechanisms often encountered in the context of big data sources. Pseudo-design-based, model-based, and machine-learning methods for predictive inference are studied and evaluated.

A key element determining the success of these methods in removing bias incurred by the selection mechanism of the data generating process is the availability of auxiliary characteristics explaining the selectivity of the data at hand. If these variables are categorical, the regression tree is found to be a suitable method with advantages compared to the simpler pseudo-design-based approach, in that it offers more flexibility while at the same time maximizing the between-variance of the data. When only certain ranges of auxiliary variables are observed but numerical auxiliary variables are available, (generalized) linear models, neural networks and support vector machines are more appropriate for the data considered in the simulation. These methods can generalize outside the domain of the available data and are capable of better predictions for new, unseen data. When a suitable explicit statistical model can be found for the data, that model can be used as the basis for predictive inference. In other situations neural networks and support vector machines may be better choices as they offer more flexibility and do not require a statistical model to be specified.

Formulating inference for non-probability sampling as a prediction problem creates a framework in which a wide range of methods can be applied. The tendency sometimes seen at NSIs to stay on familiar ground and use pseudo-design-based methods is too restrictive and will often not be sufficient to remove selection bias from data sets with non-random data generating mechanisms. While a range of methods are considered in the present study, there are many more that one could consider; for example random forests (Breiman 2001), and neural networks with different architectures (Bishop 1996).

An issue not exhaustively addressed in the present study is that of model selection. Given a data set including auxiliary variables, which of the various inference methods should one use, and which auxiliary variables? The classic approach taken in machine learning is to use predictive accuracy as a basis for choosing a method (Hastie et al. 2009). One would split the available data into two subsets—one for training and one for testing—and predict values for the units in the test set. Minimization of the prediction error for the test set can be used as a criterion for model selection.

Since the availability of auxiliary variables explaining the data generating mechanism is crucially important, research on big data specifically and on non-probability samples in general should include efforts on the collection or derivation of auxiliary variables. Typical sources of common auxiliary variables are administrative registers. Recently, process variables have been shown to be a promising type of auxiliary variables; in survey sampling contexts these are sometimes referred to as para-data (Kreuter 2013). Finally, attempts to recover auxiliary variables from the data—a procedure also known as profiling—could be valuable in big data settings; see Nguyen et al. (2014) for an example.

A set of variables explaining the selection mechanism of the data generating procedure and an appropriate predictive inference method are essential ingredients for successfully employing non-probability samples and big data in official statistics.

References

Adler D., C. Gläser, O. Nenadic, J. Oehlschlägel and W. Zucchini (2014). ff: memory-efficient storage of large data on disk and fast access functions. R package version 2.2-13. http://CRAN.R-project.org/package=ff

Baker R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology* 1: 90–143, first published online September 26, 2013 doi:10.1093/jssam/smt008.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 4: 251–260.

Beygelzimer, A., S. Kakadet, J. Langford, S. Arya, D. Mount and S. Li (2013). FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1. http://CRAN.R-project.org/package=FNN.

Buelens, B., Daas, P., Burger, J., Puts, M. and van den Brakel, J. (2014). Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Bishop C. (1996). *Neural Networks for Pattern Recognition*. Clarendon Press.

Bishop C. (2007). *Pattern Recognition and Machine Learning*. Springer.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l' Institute International de Statistique* 22**,** Supplement to Book 1: 6–62.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5–32.

Daas, P.J.H., Puts, M.J., Buelens, B. and van den Hurk, P.A.M. (2015). Big data as a source for official statistics. *Journal of Official Statistics* 31(2): 249–262.

de Jonge E., J. Wijffels and J. van der Laan (2014). ffbase: Basic statistical functions for package ff. R package version 0.11.3. http://CRAN.R-project.org/package=ffbase

Deville, J., and C.-E. Särnal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87: 376–382.

Elliott R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2(6).

Gelman A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Ginsberg J, Mohebbi, M.H., Patel, R.S., et al. (2009). Detecting influenza epidemics using search engine query data. *Nature* 457:1012–1014.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14: 333–362.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Society* 78: 776–793.

Hastie T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second edition. Springer.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663–685.

Kreuter, F. (Ed.) (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information* (Vol. 581). Wiley.

Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software 9(1): 1–19

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch (2014). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4. http://CRAN.R-project.org/package=e1071

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* 3: 169–174.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97: 558–625.

Nguyen, D., Trieschnigg, D. and Meder, T. (2014). Tweetgenie: Development, evaluation, and lessons learned. *Proceedings of the 25th International Conference on Computational Linguistics*: 62–66

Pfeffermann, D. (2002). Small Area Estimation – New developments and directions. *International Statistical Review* 70: 125–143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* 28: 40–68.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley.

Rao, J.N.K. (2011). Impact of frequentist and Bayesian methods on survey sampling practise: a selective appraisal. *Statistical Science* 26: 240–256.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

Särndal, C.-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review* 55: 279–294.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Smola, A.J. and Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* 14: 199–222.

Therneau, T., B. Atkinson and B. Ripley (2014). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8. http://CRAN.R-project.org/package=rpart

Valliant, R., Dorfman, A. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* Wiley.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory.* Springer.

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Wickham, H. (2009) ggplot2: elegant graphics for data analysis. Springer New York.

7. Supplement

7.1 Population and sample documentation

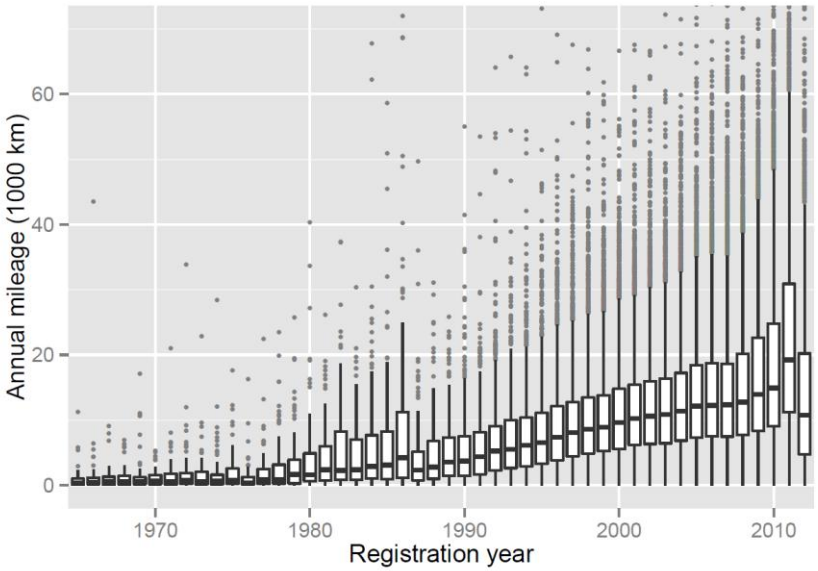

Figure S1 Correlation between auxiliary variable registration year and target variable annual mileage in (a random sample of) the Online Kilometer Registration. Vehicles older than 25 years have a relatively high annual mileage due to tax benefits for old-timers. The most recent vehicles have a relatively low mileage because they are introduced during the year.

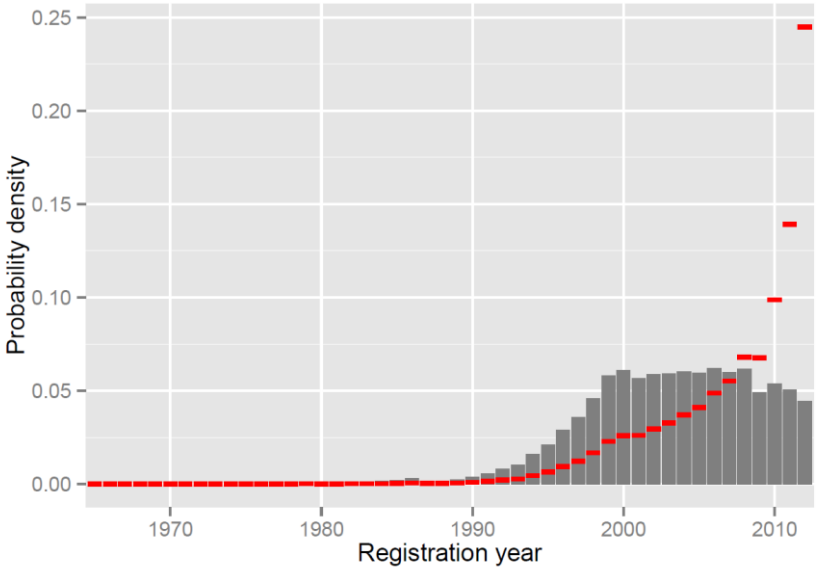

Figure S2 Probability mass function of registration year in population A (grey bars) and sample 1 in which younger vehicles are overrepresented (red lines).

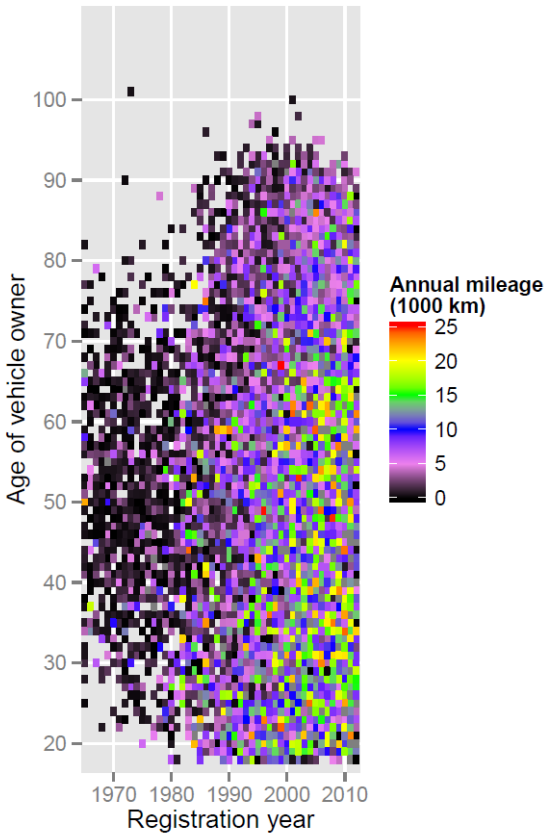

Figure S3 Correlation between registration year, age of vehicle owner, and annual mileage in (a random sample of) population B.

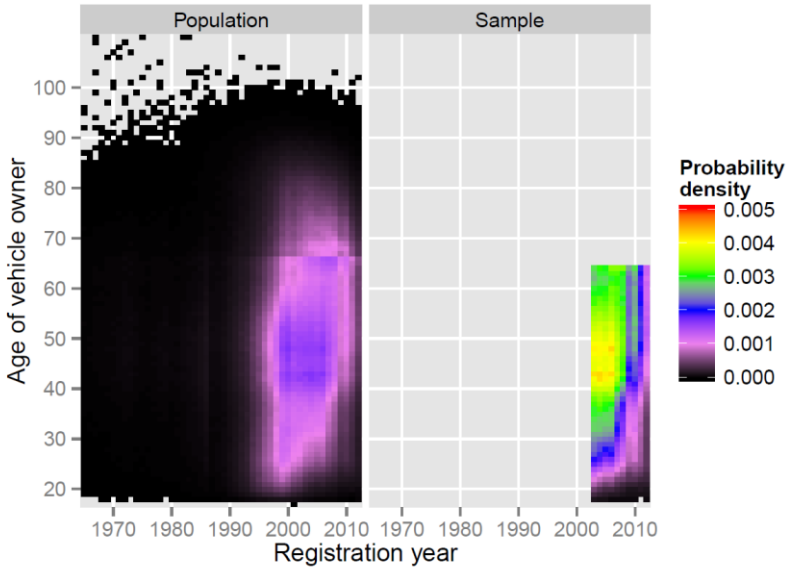

Figure S4 Probability density function of registration year by age of vehicle owner in (left) population B and (right) sample 7.

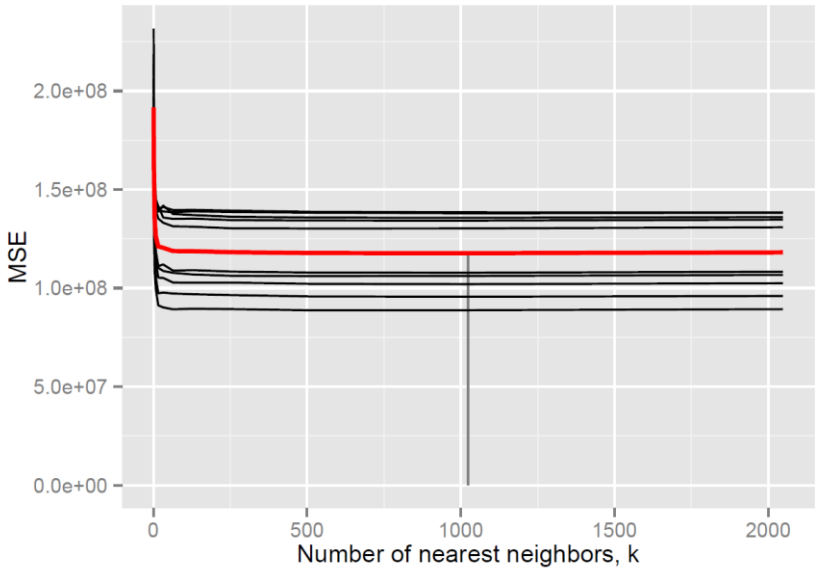

Figure S5 Example of model optimization: choosing the optimum number of nearest neighbors for the KNN model, given sample 9 (only vehicles up to ten years old) and complete information (registration year) available for prediction. The model was trained and tested on ten bootstrap resamples (black lines). The optimum k is indicated by the vertical line where the average MSE (red line) is smallest.

Table S1 Optimal input values for each model by scenario.

Sample	Scenario	GLM model*	KNN k	ANN node size	RTR stopping criterion	SVM kernel
1	a	$y{\sim}p$	512	10	10^{-4}	linear
	b	$y{\sim}l + w + f + lw + lf + wf + lwf$	512	6	10^{-7}	polynomial
2	a	$y{\sim}l$	4096	6	10^{-3}	sigmoid
	b	$y{\sim}p + w + f + pw + pf + wf$	128	5	10^{-8}	polynomial
3	a	$y{\sim}w$	2048	7	10^{-4}	radial
	b	$y{\sim}p + l + f + pl + pf + lf + plf$	128	4	10^{-8}	polynomial
4	a	$y{\sim}p + l + w + pl + lw$	512	3	10^{-7}	polynomial
	b	$y{\sim}p$	64	8	10^{-5}	polynomial
	c	$y{\sim}l$	8	4	10^{-3}	sigmoid
	d	$y{\sim}w$	256	5	10^{-3}	polynomial
	e	$y{\sim}f$	256	8	10^{-3}	polynomial
5	a	$y{\sim}p + l + w + pl + pw + lw$	128	4	10^{-4}	polynomial
6	a	$y{\sim}p + l + w + pl + pw + lw + plw$	256	5	10^{-6}	radial
7	a	$y{\sim}r + r^2 + a + a^2 + ra$	256	7	10^{-3}	radial
8	a	$y{\sim}r + r^2 + a + a^2 + ra + ra^2$	1024	6	10^{-3}	radial
9	a	$y{\sim}r + r^2$	1024	8	10^{-4}	radial
	b	$y{\sim}a + a^2$	2048	7	10^{-4}	radial
	c	$y{\sim}r + r^2 + a + a^2 + ra + r^2a$	256	8	10^{-3}	radial
10	a	$y{\sim}a + a^2$	1024	6	10^{-4}	radial
	b	$y{\sim}r + r^2$	512	6	10^{-5}	radial
	c	$y{\sim}r + r^2 + a + a^2 + ra + ra^2 + r^2a$	512	10	10^{-4}	radial

*p = registration period (categorical), r = registration year (numerical); l = legal form; w = vehicle weight; f = fuel; a = age of vehicle owner

7.2 Additional results

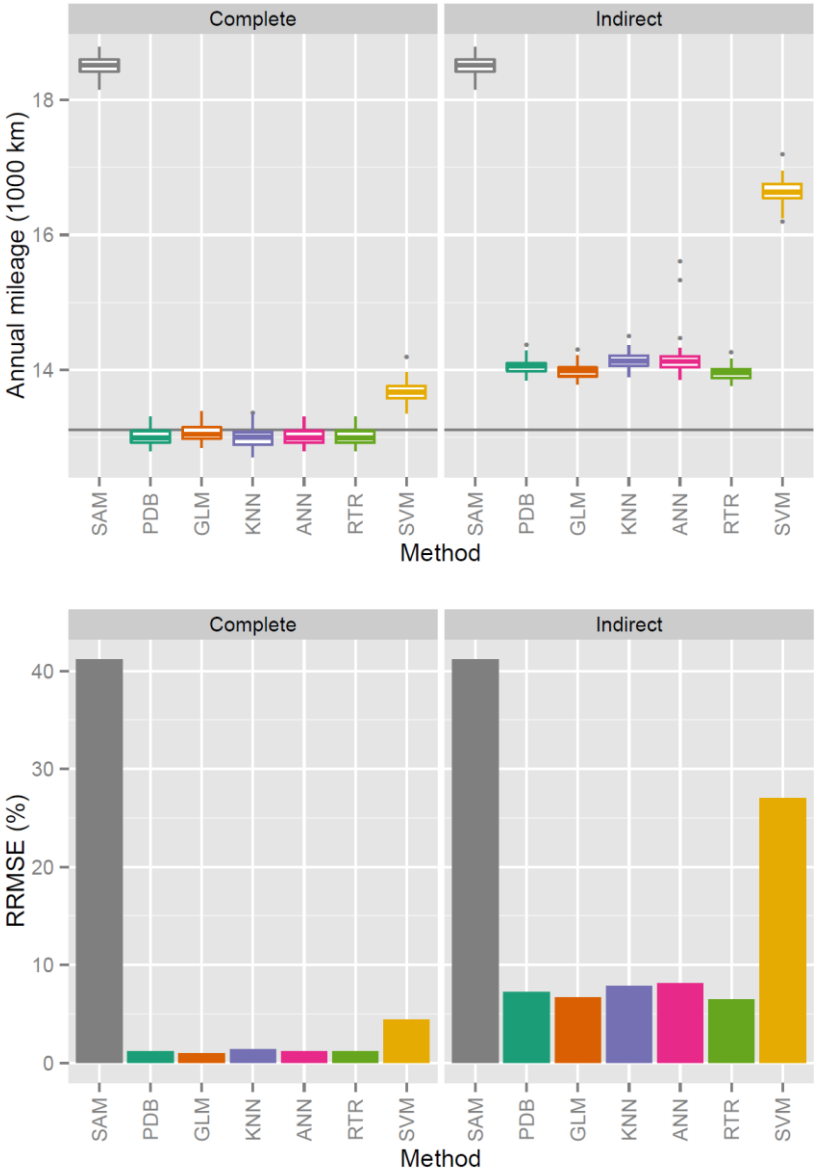

Figure S6 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 2** (company vehicles overrepresented) and (Left) complete information (legal form) or (Right) indirect information (registration year, vehicle weight and fuel type) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

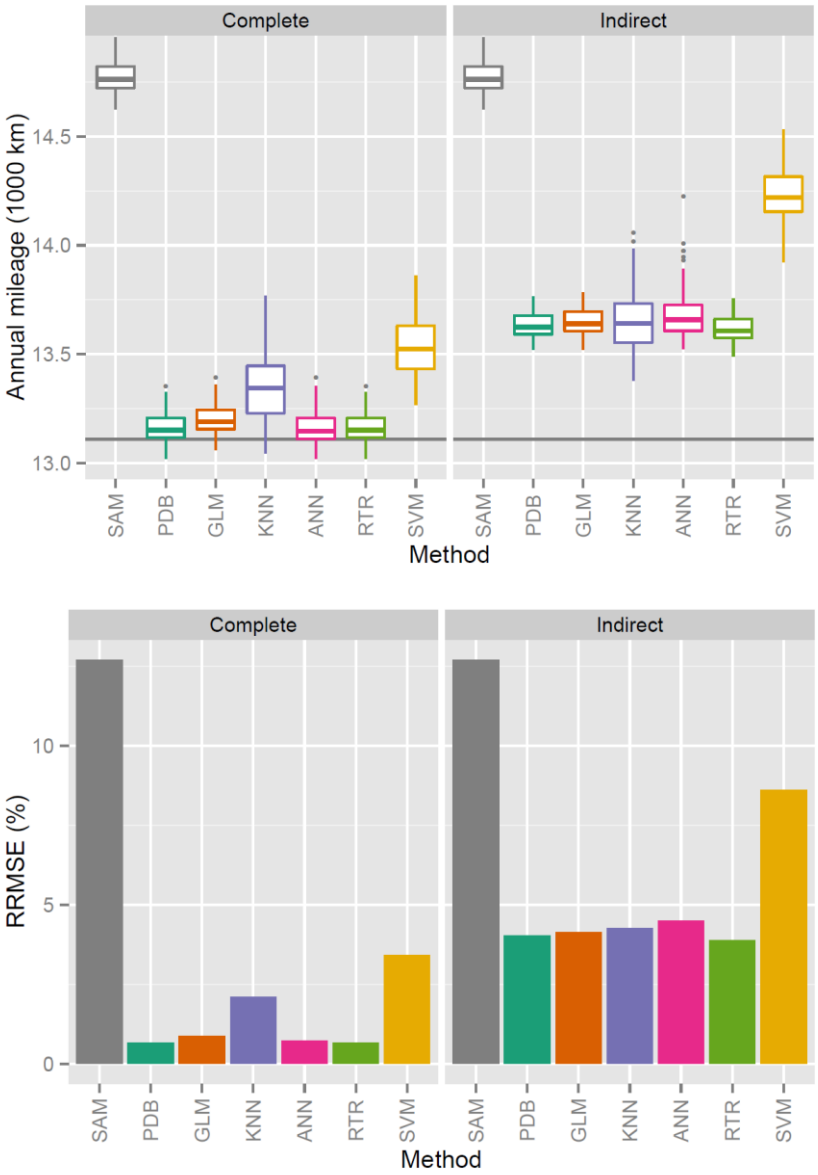

Figure S7 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 3** (heavy vehicles overrepresented) and (Left) complete information (vehicle weight) or (Right) indirect information (registration year, legal form and fuel type) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

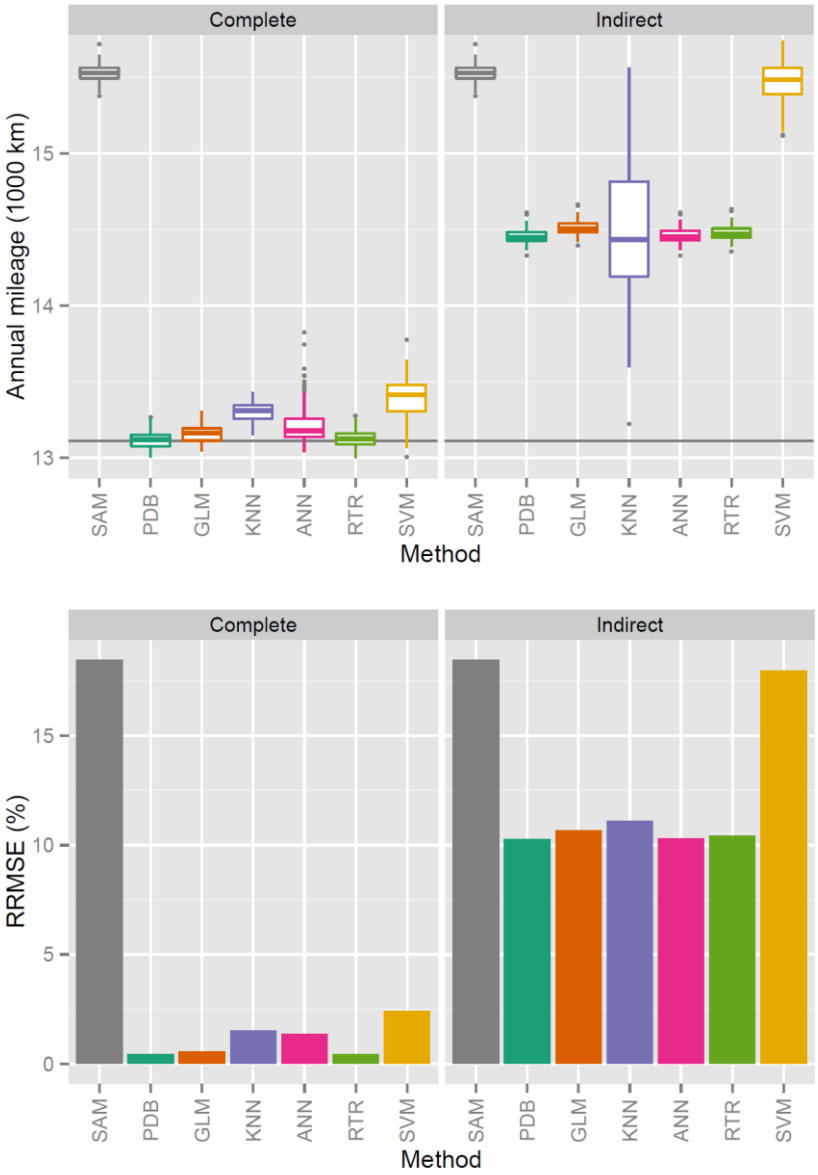

Figure S8 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 4** (young, heavy company vehicles overrepresented) and (Left) complete information (registration year, legal form and vehicle weight) or (Right) indirect information (fuel type) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

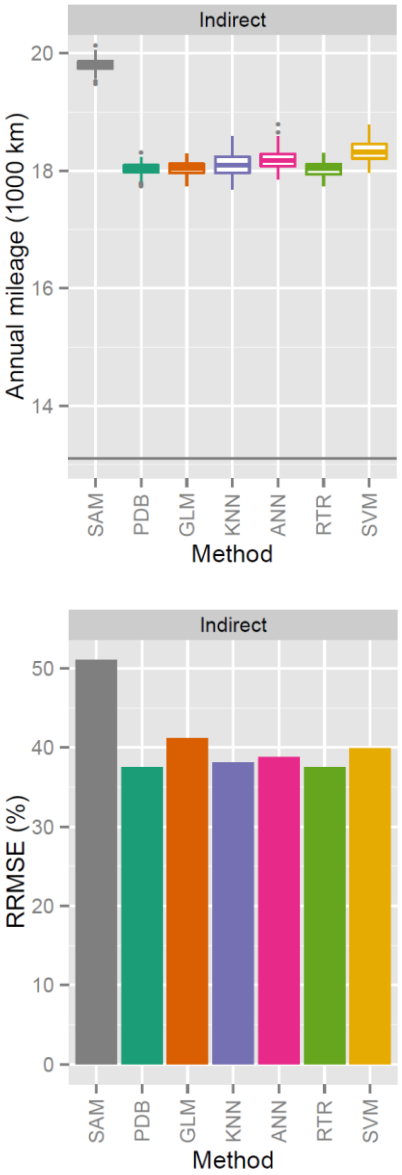

Figure S9 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 5** (high-mileage vehicles overrepresented) and indirect information (registration year, legal form and vehicle weight) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

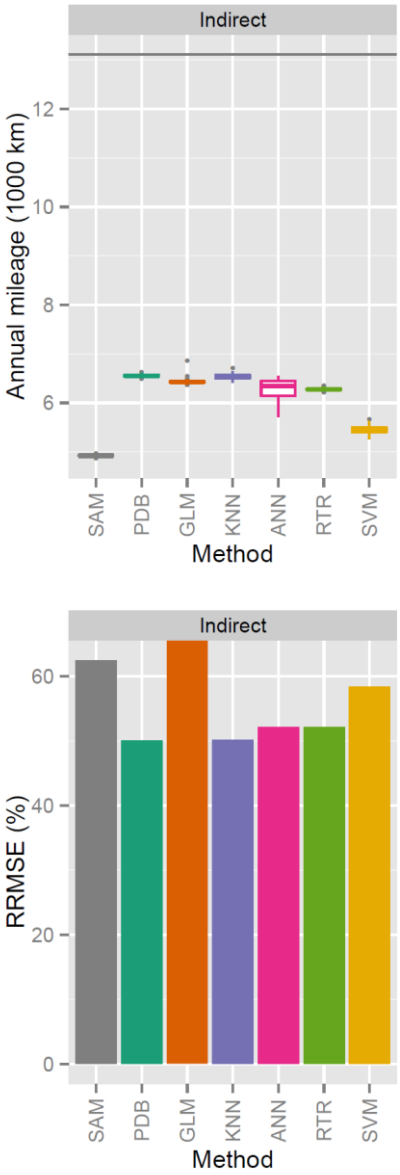

Figure S10 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 6** (low-mileage vehicles overrepresented) and indirect information (registration year, legal form and vehicle weight) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0. The RRMSE of the GLM is outside the plotted range due to one extreme bootstrap prediction.

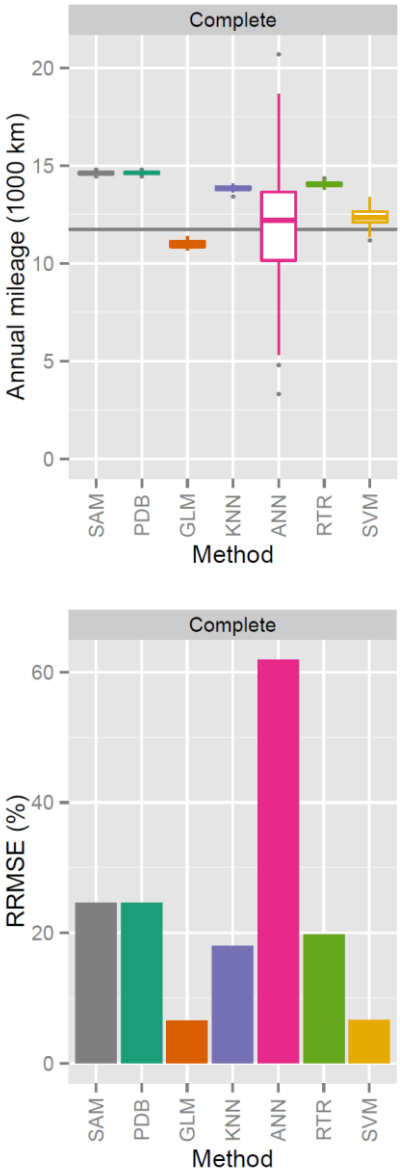

Figure S11 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 7** (only vehicles up to ten years old, owned by persons under 65) and complete information (registration year and age of owner) available for prediction. In upper panels, horizontal line is true population level.

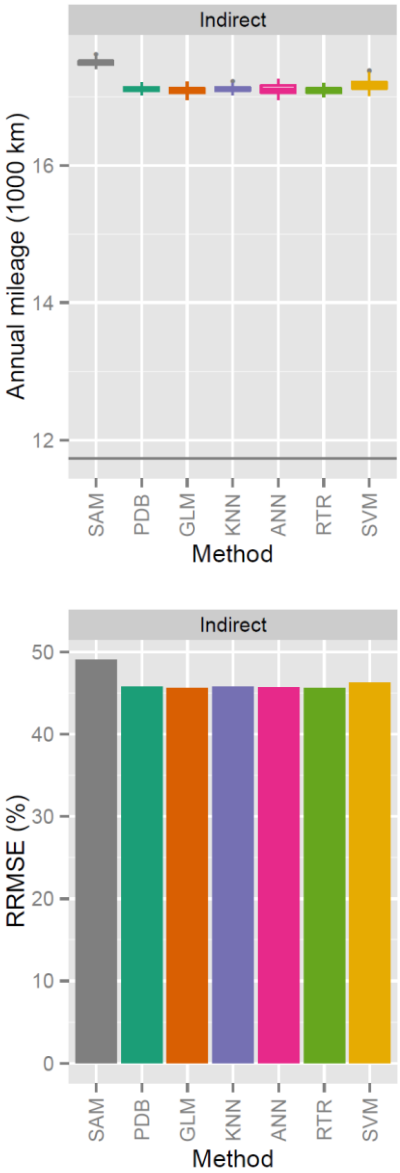

Figure S12 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 8** (only vehicles with an annual mileage of 10,000 km or more) and indirect information (registration year and age of owner) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

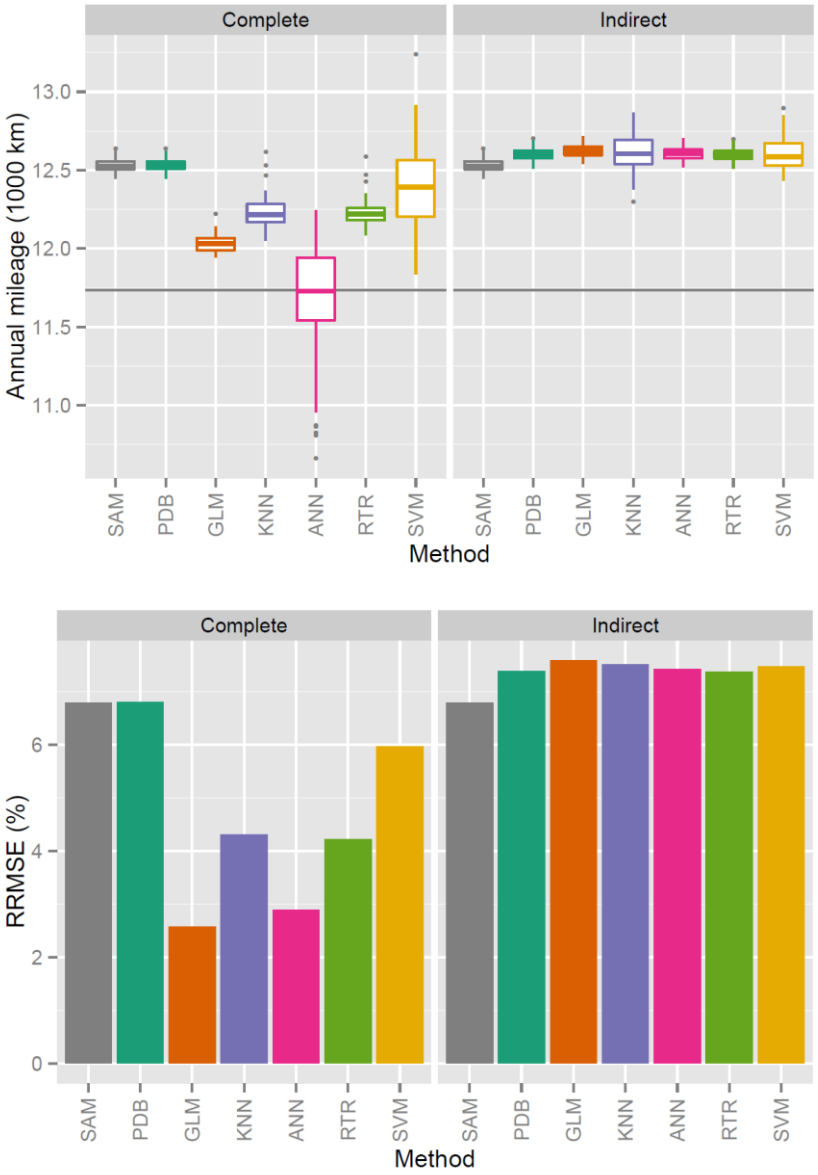

Figure S13 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 10** (only vehicles owned by persons under 65) and complete information (age of owner) or indirect information (registration year) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

7.3 Extra scenarios

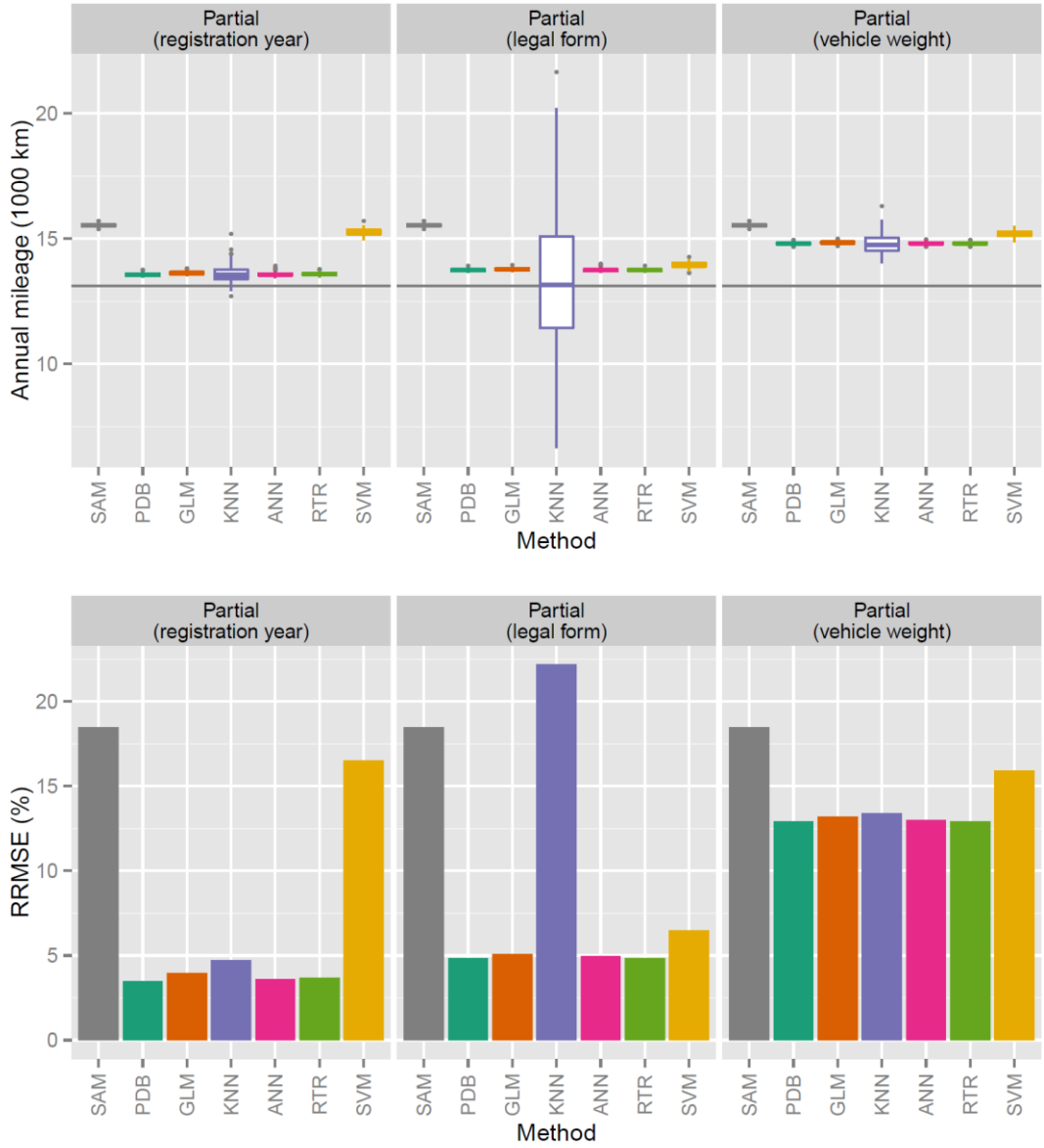

Figure S14 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 4** (young, heavy company vehicles overrepresented) and partial information (registration year, legal form or vehicle weight) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

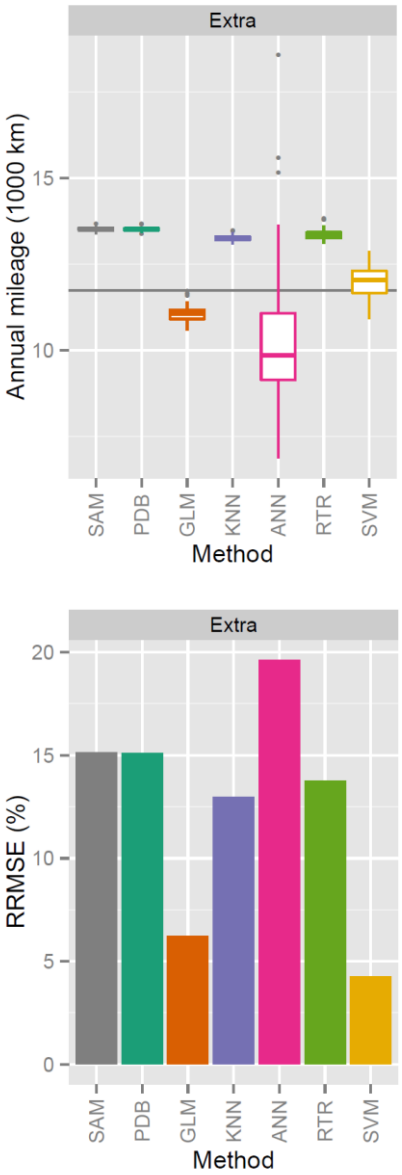

Figure S15 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 9** (only vehicles up to ten years old) and extra information (registration year and age of owner) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

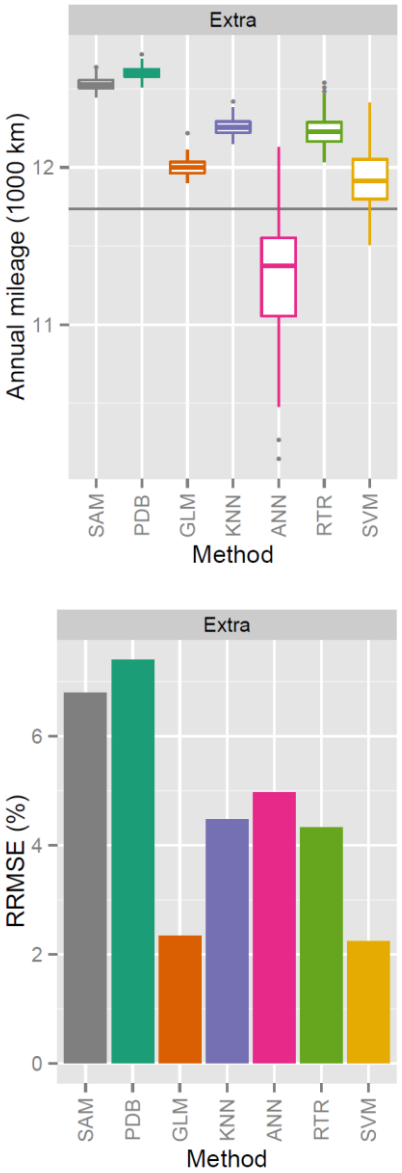

Figure S16 Effect of inference method on (Upper panels) predicted mean annual mileage (boxplot of 100 bootstrapped predictions) and (Lower panels) relative root mean square error, given **sample 10** (only vehicles owned by persons under 65) and extra information (registration year and age of owner) available for prediction. In upper panels, horizontal line is true population level; note that y-axis does not start at 0.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

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

Prepress
Studio BCO, Den Haag

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information