Center for
Big Data Statistics

Working paper

Using Road Sensor Data to Correct Survey Measurement Error Through Multiple Systems Estimation

Working paper no.: 01-20

Jonas Klingwort
Joep Burger
Bart Buelens
Rainer Schnell

January 2020

Contents

1 Introduction 4

2 Research Background 4
2.1 Underreporting in Transport, Mobility, and Travel Surveys 5

3 Data 6
3.1 Survey Data 6
3.2 Weigh-in Motion Road Sensor Data 6
3.3 Vehicle and Business Register Data 8

4 Methods 8
4.1 Definitions and Notation 8
4.2 Survey Estimates 9
4.3 Capture–recapture Techniques 10
4.4 Capture–recapture Assumptions 10
4.5 Lincoln–Petersen Estimator 10
4.6 Logit Model 11
4.7 Log-Linear Model 11
4.8 Model Selection 12
4.9 Variance Estimation 12
4.10 Linking Survey and Sensor Data 13

5 Results 13
5.1 Bootstrap Diagnostics 15
5.2 Stratification 15
5.3 Simulation 15
5.4 Summary 16

6 Discussion 16

7 Conclusion 16

Abstract

Time-based diary surveys collect data over specified time intervals and impose a heavy response burden. To reduce the effort of reporting, respondents may omit spells or may not respond at all. Correspondingly, those surveys may suffer from underreporting and item-non-response. Accordingly, survey estimates might be biased downwards. Capture-recapture methods (CRC) are used to estimate underreporting in the Dutch Road Freight Transport Survey (RFTS). The heterogeneity of the vehicles concerning capture and recapture probabilities is modeled through logistic regression and log-linear models. Six different estimators are discussed and compared. The obtained CRC estimates suggest considerable amounts of underreporting in the RFTS, although the levels are not surprising compared with findings of other validation studies on underreporting in transport, mobility, and travel surveys. All estimators applied yield larger point estimates than the RFTS, although the estimated amount of underreporting varies between the estimators. Linking sensor data to surveys and applying capture-recapture techniques is a promising method to estimate underreporting in surveys. However, the used sensor data are more suitable to complement the survey rather than replacing it.

1 Introduction

Nonprobability-based big data sources have become increasingly popular in social science research and official statistics.[1)] Due to their unknown data generating processes, these data sets are rarely used in the production of official statistics. However, in practice, it is often required to use this kind of data to provide statistics cheaper and faster and to reduce response burden (Buelens, 2012; Daas et al., 2015). In most cases, big data sources are partial observations of one or a few variables of a subset of a population (Buelens et al., 2014). One important type of big data is often produced and collected by sensors, which can be any device storing information about physical elements and human behavior (Ganguly et al., 2009). Sensor data are often not collected for research purposes (Connelly et al., 2016). Rather, the resulting data sets are large, complex, and unsystematic. Finally, the data are often held by commercial agencies (Schnell, 2019a). Nevertheless, the information the data hold should be utilized in the production of official statistics (Citro, 2014; Lohr & Raghunathan, 2017). A current promising concept seems to be the production of 'multisource statistics' (De Waal et al., 2017). For these concepts, record linkage on a microlevel is an essential tool (Schnell, 2016), since big data sources often contain only a few or no covariates, resulting in low information content.

When a sensor and a survey that independently measure an identical target variable can be linked by a unique identifier and can be enhanced with administrative data, a maximum information gain is achieved (Japec et al., 2015). In this case, the term 'big data' can be expanded to 'identifiable big data' (Shlomo & Goldstein, 2015). Hence, to evaluate the enhancement of survey data with administrative and big data, empirical research on linkable data sets is needed. In this article, the Dutch Road Freight Transport Survey (RFTS), the weigh-in-motion (WIM) road sensor data, the Dutch Vehicle Register (BVR), and the Dutch business register (BR) are linked on a microlevel for analysis.

An important aim of the RFTS is to provide estimates of transported shipment weight at quarterly and annual intervals. Due to nonresponse and underreporting, a downward bias in the RFTS point estimates is expected. We use WIM data to assess, quantify, and correct this bias associated with estimates of the number of days on which transport occurred and the corresponding transported shipment weights. The corrections are based on an application of CRC techniques. These techniques were initially developed in ecology and biology to estimate (unknown) population sizes. The RFTS and the WIM observations are considered as two capture occasions. The BVR and BR provide covariates to model heterogeneity in the capture probabilities both for RFTS and WIM. This application is a new example of multisource estimation in official statistics.

2 Research Background

The number of surveys conducted has increased over the last decades (Singer, 2016), while the nonresponse rates are increasing, too (Meyer et al., 2015). Furthermore, surveys put an unnecessary burden on the respondent if the information of interest is accessible from other

[1)] This workingpaper is partially based on Klingwort et al. (2019a).

data sets (Schnell, 2015; Miller, 2017). Especially, time-based diary surveys collect data on specified time intervals and impose a heavy response burden. To reduce the effort of reporting, respondents may omit spells or may not respond at all. Correspondingly, those surveys yield low response rates (Krishnamurty, 2008). Accordingly, survey estimates might be biased downward due to 'inaccurate reporting, nonreporting, and nonresponse' (Richardson et al., 1996). As will be shown, even in the case of a mandatory survey with a high response rate, these problems cannot be neglected.

The RFTS is a mandatory time-based diary survey collecting data from truck owners on road freight transport activities in a specified time interval. In the past, transport, mobility, and travel surveys were already subject to validation studies. For this purpose, GPS sensor data from portable devices have been analyzed with a geographic information system (GIS). However, these studies have important limitations. If no external sensor data can be linked to the survey, the potential respondents must participate in a supplementary survey. This additional burden results in low participation rates (Bricka & Bhat, 2006). Additional issues arise from the fact that the data collection devices are connected to the survey unit (vehicle or person). In practice, GPS devices cause problems due to intended or unintended switch off, delays due to standby mode, battery issues, or the device not being carried.

Furthermore, the use of GPS devices in surveys is not suitable for all population members, for example, the elderly or retired (Bricka et al., 2012). Finally, signal loss, signal noise, and matching of GPS and survey data complicate accurate measurements (Shen & Stopher, 2014). Instead, data based on local, permanently installed road sensors are used in our research to validate and adjust survey estimates using CRC techniques. As a result of this, the problems caused by respondent behavior, as discussed above, are avoided. However, the road sensors used have different drawbacks (see section 3.2). To the best of our knowledge, road sensor data have not been used for correcting surveys before.

2.1 Underreporting in Transport, Mobility, and Travel Surveys

In 1986, Hassounah et al. (1993) documented underreporting rates for a large-scale transportation survey in the United States varying regionally from 2.6% to 46.8%. Due to the technical absence of GPS data, this study used cordon counts to estimate underreporting. In GPS validation studies, vehicles were equipped with GPS devices to track movements. In the first GPS household travel survey (1997, United States), Pearson (2001) reported underreporting in trip rates of 12.4% and 31.1%. The discrepancy is due to the definition of dwell times. These findings were confirmed by Wolf et al. (2003), who reported rates of missed trips up to 42% in the Californian Household Travel Survey. Bricka and Bhat (2006) summarized the levels of underreporting in GPS surveys in the United States and reported even higher rates up to 81%. Stopher et al. (2007) reported contrary results for the Sydney Household Travel Survey (2004), where only 7.4% of trips were missed. However, all non-recorded GPS trips due to technical issues were excluded. In recent studies, Bohte and Maat (2009) concluded that GPS-/GIS-based results from 2007 are comparable to results from the 2006 Dutch Travel Survey. In contrast, Wolf et al. (2013) reported for a regional household travel survey in the United States (2010/2011) that GPS-based results showed higher trip rates.

Summarizing the results for transport, mobility, and travel surveys, there seem to be contrary results, but underreporting in reported trips is likely. Therefore, the use of sensor data to assess survey data quality and to validate and adjust biased survey estimates seems to be a promising method.

3 Data

3.1 Survey Data

The RFTS is conducted by Statistics Netherlands and is based on Eurostat (2016) guidelines. A central objective of the mandatory survey is to collect data on the weight of the shipments transported by Dutch trucks. Therefore, truck owners must report the days on which the truck was used and the corresponding shipment weight. No report is required if the truck was not used for transport purposes. The target population is the Dutch commercial vehicle fleet, excluding military, agricultural, and commercial vehicles older than 25 years. Furthermore, only vehicles with a weight of at least 3.5 tons and at least 2 tons of load capacity are taken into consideration. The sample is stratified by six variables (the type of transport, type of vehicle, industry class, load capacity, age of the vehicle, size of vehicle fleet), resulting in 74 strata. For each quarter of 2015, a separate sample is drawn and invited to the survey. A sampling unit consists of a truck license plate and a specific week for which reporting is required. Hence, a truck can be sampled more than once in 2015, but with different survey periods.

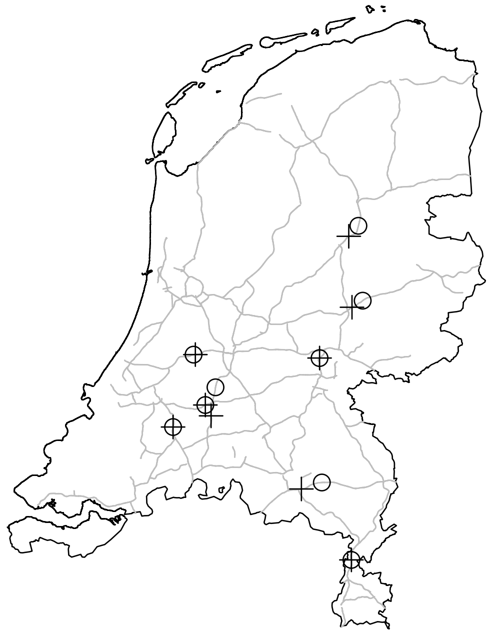
The RFTS is conducted using internet interviewing, postal interviewing, and querying of software-based journey planning systems. Large haulers use the latter. Especially, small companies receive a paper questionnaire. All other haulers and truck owners are contacted by a postal letter and are invited to participate in a web survey.

The sample consists of 33,817 unique vehicle–week combinations. Of these, 3,597 cases are classified as nonresponse, resulting in a response rate of 89.4%. The answer categories regarding truck-related activities are the following: truck used (22,454), truck not used (5,304), and truck not owned (2,462). The latter case is excluded from the analysis because the validity of the response cannot be verified. This decision is due to quarterly updates of the BVR, complexity in holding companies, vehicle rental, and vehicle leasing. However, in the case of this response, the survey agency asks for the contact information of the current owner. If this information is available, the current owner is contacted, and the new response replaces the initial response. The answer category that the truck has not been used reduces the respondent's burden considerably since only small parts of the questionnaire must be answered. Nevertheless, choosing this response fulfills the obligation to participate in the mandatory survey. It is expected to find cases of underreporting due to nonresponse and misreporting by falsely responding that the truck was not used. It is not possible to assess measurement errors due to the respondent reporting wrong dates the truck was used in the survey period, or the respondent reporting the wrong weight of the transported shipment.

3.2 Weigh-in Motion Road Sensor Data

The Dutch national road administration operates the WIM road sensor network. The purpose of this system is to detect overloaded trucks using a dynamic measurement while trucks pass the station. If there is suspicion of overloading, the truck is taken to a traffic checkpoint, and a static weighting is done. In 2015, there were nine operating WIM systems. Each system measured both directions (see Figure 3.1).

This network installation results in 18 measurement points. For the analysis, the recorded variables are the date, front/rear license plate, total weight, axles pressure, and automated truck



3.1 WIM road sensor network on Dutch highways. Circles and crosses indicate the 18 sensor systems on Dutch highways.

classification. In 2015, a total of 35,669,347 trucks pass-by were recorded, of which 24,825,019 had a front license plate recognized by the WIM software system. Of those eligible, 3,733,064 records matched a truck from the survey using the front license plate as match variable, and 44,011 of the recorded trucks matched a truck in its corresponding survey period using the combination of the time stamp (day) and front license plate as match variable. For each truck, its axle weights are measured, and the total weight corresponds to the sum of the individual axle weights. Based on Enright and OBrien (2011) and expert information from the road administration, a conditional mean imputation was applied to the measured axles weights to correct for measurement errors. A deterministic error correction rule is used in this study. If the measured weight of an axle is greater than 20 tons, the weight of this axle is imputed by the average weight of the remaining axles. If the weight of more than one axle exceeds 20 tons, the average value of the remaining axles with a weight of fewer than 20 tons is used here, too. Due to 1,629 trucks having no axle weights stored in the data (the total weight is available), this rule could not be applied to these cases. Sensitivity analysis of the deterministic correction showed that choosing the threshold too small ($\leq$ 15 tons) leads to a downward bias in the distribution of the measured weight. For these cases and cases where trucks were driving outside the speed interval [60, 120 km/h], the total weight was predicted using the technical characteristics from the vehicle register (described in paragraph 'Register data'). Therefore, a stepwise model selection procedure based on the Bayesian information criterion (BIC) was applied (using the R function 'stepAIC' of the MASS package by Venables and Ripley (2002)) to select a model for a linear regression ($r^2_{adj.} = 0.54$).

In 17,321 of the 44,011 matched trucks (see the Results section), the trailer weight could not be linked to the WIM. This was due to the license plate not being recognized by the WIM software system (11,341) or the trailer not being registered in the vehicle register (5,980). The missing trailer weight was imputed with the mean of the empty trailer weight, conditional on the automated classification of the truck, and its loading capacity. Since the total weight measured includes the entire unit (truck, trailer, and shipment), the truck and trailer weights were subtracted using the weight information from the BVR. The resulting value corresponds to the transported weight, which is equal to the definition of reported weight in the RFTS. The calculation of the transported shipment weight resulted in 3,945 negative values. These values were trimmed to 0 to not bias the estimate of the transported weight downward. Finally, an overall proportional bias correction was applied, calibrating the WIM-measured shipment weights to those reported in the RFTS. The correction factor was obtained from the subset of vehicles that were observed both in the RFTS and by WIM. This resulted in a down-scaling of the WIM shipment weights by approximately 14%.

3.3 Vehicle and Business Register Data

The BVR and BR provide additional administrative data with information about technical truck characteristics and specifics of the truck owners. Both data sets were linked one by one to the RFTS and WIM data using the combination of license plate and the annual quarter as match variable.

The BVR contains the covariates, truck equipment class, type of fuel, number of wheels, number of cylinders, horsepower, emission class, maximum mass of truck, mass of empty truck, maximum mass of trailer, loading capacity, number of axles, width of truck, length of truck, leasing status, status of owner (person or company), province in which the owner is located, year of manufacture, and vehicle classification.

The covariates provided by the BR are the classification of economic activity (NACE), commercial or own transport, classification of company size, the size of the vehicle fleet, and the total fleet loading capacity. The variables of the BVR and BR will be used within the model selection to find appropriate covariates for the CRC models. Observations with missing administrative data were excluded from the analysis. This decision explains the difference between the 44,011 matches and the 43,775 truck days in table 4.1).

4 Methods

4.1 Definitions and Notation

We define the indicator $\delta_{i,j}^{svy}$, which takes the value 1 if vehicle i has been on the road on day j of its survey period according to the survey response, and the value 0 otherwise. Similarly, we define $\delta_{i,j}^{sen}$ to be an indicator equal to 1 if vehicle i is recorded by a WIM station on day j and equal to 0 otherwise. We define $\Theta_{i,j}$ to be the weight of the shipment carried by truck i on day j. If $\delta_{i,j}^{svy} = 1$ we use the sum of reported shipment weights in the survey, otherwise if $\delta_{i,j}^{sen} = 1$ we use the WIM shipment measurements as described in section 3.2. A vehicle can be captured by WIM sensors multiple times a day, in that case, the maximum of the weights measured at these occasions is taken. If the vehicle is recorded only once, simply the weight measured on that occasion is used.

4.2 Survey Estimates

The regular, published statistics from the RFTS are post-stratification estimates. Survey weights are calculated, taking the survey design into account and correcting for selective nonresponse (Centraal Bureau voor de Statistiek, 2017). The regular survey estimators for the total of D and W are defined as $\widehat{D}^{SURV}$ and $\widehat{W}^{SURV}$ and estimated by

$$\hat{D}^{SURV} = \sum_{i=1}^{r} \left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \right) \tag{1}$$

with w_i, the survey weight for vehicle i and r the number of respondents. The post-stratification estimator for the total transported weight is given by

$$\hat{W}^{SURV} = \sum_{i=1}^{r} \left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \Theta_{i,j} \right). \tag{2}$$

The w_i are based on the initial post-stratification weights w_i^+ (technical name "Ophoogfactor") that correct for selective nonresponse

$$w_{i\in h}^{+} = 13 \frac{N_h^+}{r_h}, \tag{3}$$

where N_h^+ is the number of vehicles in post-stratum h including vehicles not owned and r_h the number of respondents in post-stratum h excluding vehicles reported not owned. The factor 13 scales up from one survey reporting week to the quarter. These initial post-stratification weights have been developed and before. The weighting itself is not part of this research. Since vehicles not owned are excluded from the analysis and the sample is defined as the study population, the w_i^+ were scaled to

$$w_i = w_i^+ \frac{n}{\sum_{i=1}^{r} w_i^+}, \tag{4}$$

so that $\sum_{i=1}^{r} = n$, with n being the number of vehicles in the sample, excluding vehicles not owned. Bootstrap estimates will be used for comparison with CRC methods. This is due to the truck days and shipment weights being clustered by trucks and not being independent and identically distributed.

In addition, we complement the survey observations with WIM observations resulting in an extended survey estimator,

$$\hat{D}^{SURVX} = \sum_{i=1}^{r} \left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \vee \delta_{i,j}^{sen} \right) \tag{5}$$

$$\hat{W}^{SURVX} = \sum_{i=1}^{r} \Big(w_i \sum_{j=1}^{7} (\delta_{i,j}^{svy} \vee \delta_{i,j}^{sen})\, \Theta_{i,j} \Big). \tag{6}$$

4.3 Capture-recapture Techniques

Capture-recapture techniques (CRC) were originally developed to estimate the unknown size of an animal population (International Working Group for Disease Monitoring and Forecasting, 1995). These techniques were transferred to human populations and are frequently used in social and medical research to address undercounts in censuses, to estimate unknown population sizes, or to estimate the incidence of a disease (Böhning et al., 2017). The biological procedure, using traps to (re)-capture animals, is replaced by using at least two datasets containing elements of the target population. With two datasets A and B available, in the first capture occasion elements are captured and marked. On the second occasion, elements get recaptured. The overlap of both capture occasions are the elements captured twice. Transferred to the present study, the first capture occasion is the RFTS, where trucks are considered as being captured and marked on specific days in the survey period ($n_1 = \sum_{i,j} \delta_{i,j}^{svy}$). The second capture occasion is the WIM data, where ($n_2 = \sum_{i,j} \delta_{i,j}^{sen}$) trucks are recorded on specific days in the survey period, of which ($m_2 = \sum_{i,j} \delta_{i,j}^{svy} \wedge \delta_{i,j}^{sen}$) are recaptured.

4.4 Capture-recapture Assumptions

In the present study the population is assumed to be closed. There are no elements entering or leaving the population, making the unknown population size a constant. This assumption is justified, since a sample is observed and that sample does not change (no vehicles leave or enter the sample after it was drawn). Furthermore, all elements used in this analysis belong to the population, as only Dutch trucks are in the RFTS and the Dutch trucks in WIM can be identified by their license plate. The assumption of perfect linkage is met for trucks recognized by the OCR software as they can be linked one-by-one with their unique identifier of license plate and date. Since the OCR software failed occasionally to properly recognize a license plate, there might be more trucks which have been recorded in the survey period. Moreover, the inclusion of a truck being in the RFTS is independent of the same truck being recorded by a WIM station (Chao et al., 2001). Finally, the capture probabilities for the elements should be homogeneous. However, it is sufficient if the homogeneity in capture probabilities is given for at least one dataset (Van der Heijden et al., 2017). In the present study the capture probabilities in the RFTS and WIM are modeled using covariates. Since the RFTS is a random sample survey, homogeneity conditional on the stratification variables can be assumed. By contrast, the inclusion of trucks in WIM is non-probabilistic.

4.5 Lincoln-Petersen Estimator

The Lincoln-Petersen estimator (Lincoln, 1935; Petersen, 1893), also known as dual-system estimator (Wolter, 1986), assumes homogeneous capture probabilities for all elements within every dataset. The Lincoln-Petersen estimator estimates the population sizes of (D) and (W) by

$$\hat{D}^{LP} = \frac{\sum_{i,j}\delta_{i,j}^{svy}\sum_{i,j}\delta_{i,j}^{sen}}{\sum_{i,j}\delta_{i,j}^{svy}\wedge\delta_{i,j}^{sen}}, \tag{7a}$$

$$\hat{W}^{LP} = \frac{(\sum_{i,j}\delta_{i,j}^{svy}\Theta_{i,j})(\sum_{i,j}\delta_{i,j}^{sen}\Theta_{i,j})}{\sum_{i,j}(\delta_{i,j}^{svy}\wedge\delta_{i,j}^{sen})\,\Theta_{i,j}}. \tag{7b}$$

The estimator $\hat{W}^{LP}$ considers the transported shipment weights on each truck day rather than the observation counts. In this research, bootstrap variance and confidence interval estimates are used to account for the dependency between vehicle and weight of the transported shipment.

4.6 Logit Model

Huggins (1989) and Alho (1990) proposed a likelihood approach, which is conditioned on the captured elements, to model heterogeneity in capture probabilities using covariates. The capture probabilities for each element on each occasion are modeled using a linear logistic model. Covariates are used to model P_{ij}^{s} and P_{ij}^{w}, which are the capture probabilities for the RFTS and WIM, respectively.

$$\hat{D}^{HUG} = \sum_{i,j}\frac{1}{\hat{\psi}_{ij}} \tag{8a}$$

$$\hat{W}^{HUG} = \sum_{i,j}\frac{1}{\hat{\psi}_{ij}}\,\Theta_{i,j} \tag{8b}$$

with

$$\hat{\psi}_{ij} = 1 - (1-\hat{P}_{ij}^{svy})(1-\hat{P}_{ij}^{sen}) \tag{9}$$

the estimated probability to be captured at least once, and $\hat{P}_{ij}^{svy}$ and $\hat{P}_{ij}^{sen}$ the model predictions of the capture probabilities in the RFTS and WIM respectively.

4.7 Log-Linear Model

Fienberg (1972) introduced Log-linear models for population size estimation in closed populations. Two datasets A and B form a $A \times B$ contingency table with a cell representing the counts of elements that are never observed. The count of this cell can be estimated by fitting a log-linear model to the incomplete contingency table.

The following is based on Coumans et al. (2017) and uses the notation of log-linear models by Bishop et al. (1975):

$$log\ m_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_x^X + \lambda_{ax}^{AX} + \lambda_{bx}^{BX} + \lambda_y^Y + \lambda_{ay}^{AY} + \lambda_{by}^{BY}. \tag{10}$$

The parameters λ_{ax}^{AX}, λ_{bx}^{BX}, λ_{ay}^{AY}, and λ_{by}^{BY} are the interaction terms between the datasets A and B and the covariates X and Y, respectively. The independence assumption between the datasets A and B is now conditioned on the covariates. For every level of the included covariates, a subpopulation size is estimated, which in sum gives the total population size. This method is used to estimate $\hat{D}^{LL}$ and $\hat{W}^{LL}$. The latter estimator is obtained using the weight of the transported shipment as the dependent variable in the models rather than cell counts.

4.8 Model Selection

To select appropriate covariates to fit the logit and log-linear models, a stepwise selection procedure based on the Bayesian Information Criterion (BIC) is used (using the stepAIC function of the MASS-package by Venables and Ripley (2002)). To cover the full information of the covariates, the model selection is based on the logit model, since the log-linear model only allows for categorical variables. All of the selected variables were used in the logit-model, whereas only five variables with the most explanatory power of both models were used in the log-linear model. For that purpose, the continuous covariates were categorized based on their quantiles.

Using $\delta_{i,j}^{svy}$ as dependent variable, the finally chosen independent variables for the logit-model were classification of economic activity (NACE), classification of company size, total fleet loading capacity, number of wheels, horsepower, maximum mass of truck, mass of empty truck, maximum mass of trailer, status of owner (person or company), and province in which the owner is located.

Using $\delta_{i,j}^{sen}$ as dependent variable, the finally chosen independent variables for the logit-model were classification of economic activity (NACE), commercial or own transport, classification of company size, size of the vehicle fleet, total fleet loading capacity, truck equipment class, type of fuel, horsepower, mass of empty truck, maximum mass of trailer, number of axles, width of truck, length of truck, status of owner (person or company), province in which the owner is located, year of manufacture, and vehicle classification.

Accordingly, the variables selected for the log-linear model were classification of economic activity (NACE), commercial or own transport, classification of company size, size of the vehicle fleet, total fleet loading capacity, number of wheels, and horsepower.

4.9 Variance Estimation

Bootstrapping is typically used to obtain variance estimates of model-based methods (Efron, 1979; Efron & Tibshirani, 1994). For consistency and comparability, we computed bootstrap variance estimates for all estimators discussed. As mentioned earlier, this accounts for the cluster effects in the data, due to the trucks being the sampling units and not the truck days. Hence, there are more truck days then sampling units. Further, the weight of the transported shipment is clustered in trucks.

Bootstrap samples are obtained from the original RFTS sample of trucks by simple random sampling with replacement. A bootstrap data set for estimation purposes consists of all records, both RFTS and WIM, that are available for the vehicles in the bootstrap sample. When the same vehicle is drawn more than once, all its associated records are repeated the number of times the vehicle occurs in the bootstrap sample.

The mean of the bootstrap distribution is computed to ascertain that the bootstrap procedure is unbiased. The variance of the bootstrap distribution is used as a variance estimate. The 0.025% and 0.975% quantiles of the bootstrap distribution are used to estimate the lower and upper boundaries of the 95% confidence intervals.

4.10 Linking Survey and Sensor Data

Table 4.1 shows the truck days of the matched RFTS and WIM. There were 94,338 truck days reported in the RFTS. 43,775 truck days were captured in the WIM, of which 34,131 were reported in the RFTS. On 9,644 days, there were no reported trips in the RFTS, but trucks were recorded at a WIM station. On 60,207 days, there were reported trips in the RFTS, but nothing was captured in the WIM.

4.1 Captures of Truck Days in RFTS and WIM

D	Sensor		
Survey	recorded	not recorded	$\sum$
reported	34,131	60,207	94,338
not reported	9,644	?	?
$\sum$	43,775	?	?

Table 4.2 shows the matched data sets as well, but the cells include the transported shipment weight in kilotons (kt) on the reported truck days. In the RFTS, 953.71 kt were reported. In the WIM, 475.96 kt were captured in the WIM, of which 376.83 kt were reported in the RFTS. In the WIM, 99.13 kt were measured but were not reported in the RFTS. In the RFTS, 576.88 kt were reported, which were not captured in the WIM.

4.2 Captures of Transported Shipment Weight (in kt) in RFTS and WIM

W (in kt)	Sensor		
Survey	recorded	not recorded	$\sum$
reported	376,83	576,88	953,71
not reported	99,13	?	?
$\sum$	475,96	?	?

5 Results

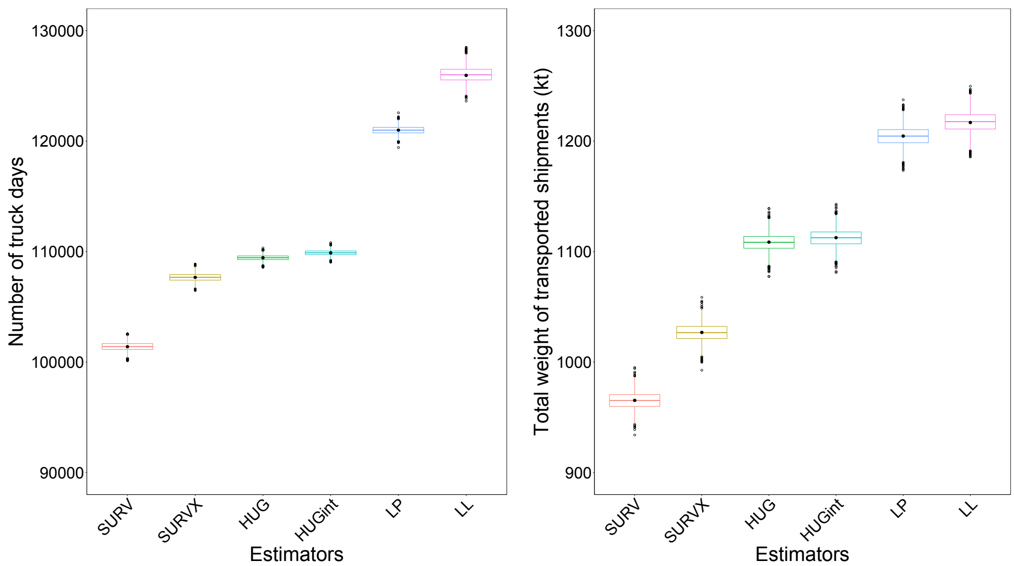
All estimators show considerable amounts of underreporting for truck days (Table 5.1; Figure 5.1, left panel) and transported shipment weight (Table 5.1; Figure 5.1, right panel). Figure 5.1 shows the six different estimates by estimator and the sampling variance for all estimators estimated by bootstrapping (3,000 bootstrap samples). The estimated amount of underreporting varies between the estimators, but a similar pattern in the amount of underestimation is found for both two target variables.

5.1 RFTS and CRC estimates for D and W (in kt), bootstrapped mean, standard error, confidence interval, and underestimation

Estimator	Point estimate	Bootstrap mean	Bootstrap standard error	Bootstrap confidence interval	Estimated underestimation (in %)
$\hat{D}^{SURV}$	101,390	101,399	395.96	[100,643; 102,197]	–
$\hat{D}^{SURVX}$	107,666	107,672	380.66	[106,923; 108,441]	5.83
$\hat{D}^{HUG}$	109,439	109,440	244.73	[108,975; 109,926]	7.35
$\hat{D}^{HUG_{int}}$	109,882	109,885	246.86	[109,412; 110,376]	7.73
$\hat{D}^{LP}$	120,994	120,996	363.75	[120,304; 121,723]	16.2
$\hat{D}^{LL}$	125,954	126,034	737.46	[124,673; 127,577]	19.5
$\hat{W}^{SURV}$	965.30	965.23	8.20	[949.33; 965.30]	–
$\hat{W}^{SURVX}$	1,026.83	1,026.69	8.37	[1,009.94; 1,043.53]	5.99
$\hat{W}^{HUG}$	1,108.58	1,108.36	8.32	[1,091.65; 1,124.37]	12.92
$\hat{W}^{HUG_{int}}$	1,112.59	1,112.40	8.34	[1,095.52; 1,128.38]	13.24
$\hat{W}^{LP}$	1,204.60	1,204.38	9.14	[1,185.83; 1,221.89]	19.87
$\hat{W}^{LL}$	1,216.85	1,217.40	9.74	[1,197.73; 1,236.08]	20.67

The amount of underestimation for the extended survey estimator $\hat{D}^{SURVX}$ is about 6%. Both conditional likelihood estimators $\hat{D}^{HUG}$ and $\hat{D}^{HUG_{int}}$ show amounts of underestimation about 7%. Using covariates to did not have a substantial effect on the estimate. The full likelihood estimators $\hat{D}^{LP}$ and $\hat{D}^{LL}$ show about 16% and 20% underestimation, respectively. In the case of the full likelihood estimators, the covariates had a more substantial effect on the estimates. The reported standard errors are likely too small and would probably be larger using the specific formulas of the estimators.

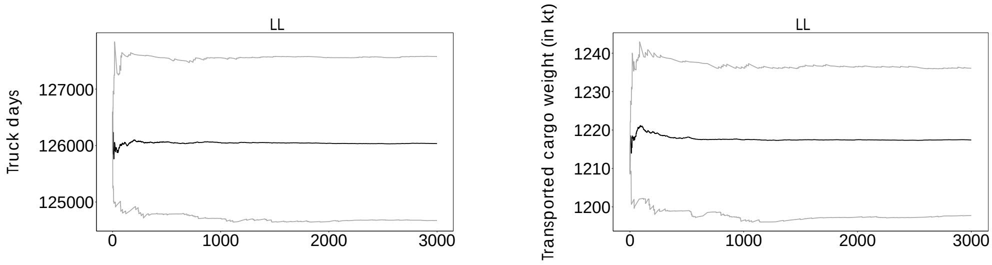
The naive extended survey estimator $\hat{W}^{SURV}$ also shows an amount of underestimation in the RFTS from about 6%. Both conditional likelihood estimators $\hat{W}^{HUG}$ and $\hat{W}^{HUG_{int}}$ show about 13% underreporting in the RFTS. Again, the covariates did not have a large effect on the conditional likelihood estimates. The largest amount of underestimation in the RFTS is obtained by the full likelihood estimators $\hat{W}^{LP}$ and $\hat{W}^{LL}$. Here, the amount of underestimation is about 20% ($\hat{W}^{LP}$) and 21% ($\hat{W}^{LL}$).



5.1 Bootstrap estimates of the six estimators for D and W (in kt). The solid black dot within each box shows the point estimates based on the original data

5.1 Bootstrap Diagnostics

Bootstrap diagnostic showed a similar overall pattern for both target variables and all applied estimators, for the point estimator as well as for the upper and lower limits of the confidence interval. This also holds for the estimators with the largest standard error $\hat{D}^{LL}$ as well as for the estimator with the smallest standard error $\hat{D}^{HUG}$ and equally for $\hat{W}^{LL}$ and $\hat{W}^{SURV}$. Figure 5.2 shows the convergence of $\hat{D}^{LL}$ and $\hat{W}^{LL}$. The point estimate is stable after 1,000 bootstrap iterations. The upper and lower limits of the bootstrapped confidence intervals limits take more iterations to converge.



5.2 Convergence of the 3,000 bootstrap iterations for the number of truck days and the transported shipment weight

5.2 Stratification

Motivated by the work from Darroch (1961), Plante et al. (1998), all estimators are applied in a stratified manner. This approach aimed not only to answer content-related questions but rather to evaluate the flexibility as well as the limitations of the applied estimators. The RFTS respondents and nonrespondents constitute the population under study. For stratification, the population under study is divided into H strata. Within each stratum, $\hat{D}_h$, $\hat{W}_h$, and the most likely amount of underreporting will be estimated. The stratified CRC estimates showed partially slightly larger amounts of underestimation in the RFTS, for example, for smaller companies and vehicles driving not for commercial purposes. However, the stratified approach also revealed the limitations of the CRC estimators when strata are small or only few captures are present within strata (Klingwort et al., 2018).

5.3 Simulation

It is not possible to assess measurement errors due to the respondent reporting wrong dates the truck was used in the survey period, or the respondent reporting the wrong weight of the transported shipment. Such errors are likely due to satisficing, digit-preference, and memory errors (Krosnick, 1991; Schnell, 2019b). The RFTS responses constitute a substantial part of the unique identifier to link the survey and sensor data. Accordingly, the CRC estimates rely among the WIM observations on the RFTS responses. Hence, errors in the reported truck days influence the obtained estimates for $\hat{D}$ and $\hat{W}$. To evaluate the robustness and sensitivity of the obtained estimators towards systematic response errors, a simulation study was conducted. Two systematic respondent-driven error patterns were simulated. One is considered as an underreporting and the other as an overreporting error. The considered overreporting error assumes too much being reported in the RFTS, and the underreporting error assumes too few being reported in the RFTS. Klingwort et al. (2019b) showed that both unconditional likelihood estimators are robust against overreporting errors and sensitive to underreporting errors.

5.4 Summary

All applied CRC estimators show considerable amounts of underreporting in the RFTS. The levels are not surprising compared with the findings of other validation studies on underreporting in transport, mobility, and travel surveys (see section 2.1). It is recommended relying on the log-linear model for estimating truck days and transported shipment weight. First, it is based on the full likelihood, whereas the logit models are conditional likelihood approaches. Second, it takes heterogeneity in the trucks related to capture and recapture probabilities into account, whereas the Lincoln–Petersen estimator assumes homogeneity.

6 Discussion

Linking sensor data to surveys and applying capture-recapture techniques is a promising method to estimate underreporting in questionnaires. Note that the sensor data are used to complement rather than replace the survey data. Implementing the method in the Dutch Road Freight Transport Survey is, however, hampered by several issues. First, the sensors do not recognize every license plate on the front and/or back of the vehicles. If this failure is selective, there could be a selection bias in the sensor data. This selection bias could affect the amount of underreporting both up and downwards, depending on the selectivity of failure. Second, some trucks were used in the analysis for which no report in the RFTS would have been required, for example, when driving for maintenance, fuel, signposting, etc. Third, there is a trend to move from road sensors to on-board sensors, which compromises the quality and accessibility of the WIM sensor data in the future.

7 Conclusion

The application demonstrates that road sensor data can potentially be used to assess underreporting in the Road Freight Transport Survey. More generally, big data can complement survey data in the production of official statistics to estimate bias in survey point estimates by combining survey, administrative, and sensor data with capture-recapture techniques. The method presented here applies to any validation study and not only to transport surveys, as long as different sources can be linked on a micro-level using a unique identifier.

The presented application is currently not considered to be implemented in connection with the Road Freight Transport Survey. Nonetheless, there are possibilities for further research. For example, a replication study with the same data for other years or the same method using different data sources. For example, finding other big data sources in which signals are recorded that could validate survey questions. Questions about facts or actual behavior would qualify sooner than questions about opinions and attitudes. Moreover, a simulation study would be useful to investigate how sensitive the method is to linkage errors.

References

Alho, J. M. (1990). “Logistic Regression in Capture-recapture Methods”. In: *Biometrics* 46, pages 623–635.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.

Böhning, Dankmar, Peter G. M. van der Heijden, and John Bunge, editors (2017). *Capture-recapture Methods for the Social and Medical Sciences*. Boca Raton: CRC.

Bohte, Wendy and Kees Maat (2009). “Deriving and Validating Trip Purposes and Travel Modes for Multi-day GPS-based Travel Surveys: A Large-scale Application in the Netherlands”. In: *Transportation Research Part C: Emerging Technologies* 17.3, pages 285–297.

Bricka, Stacey and Chandra Bhat (2006). “Comparative Analysis of Global Positioning System-based and Travel Survey-based Data”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1972, pages 9–20.

Bricka, Stacey et al. (2012). “An Analysis of the Factors Influencing Differences in Survey-reported and GPS-recorded trips”. In: *Transportation Research Part C: Emerging Technologies* 21.1, pages 67–88.

Buelens, Bart (2012). *Shifting Paradigms in Official Statistics: From Design-based to Model-based to Algorithmic Inference*. CBS Discussion Paper 2012–18.

Buelens, Bart et al. (2014). *Selectivity of Big Data*. CBS Discussion Paper 2014–11.

Centraal Bureau voor de Statistiek (2017). *Basisbestanden Goederenwegvervoer 2015*. CBS publicaties.

Chao, Anne et al. (2001). “The Applications of Capture-recapture models to Epidemiological Data”. In: *Statistics in Medicine* 20, pages 3123–3157.

Citro, Constance F. (2014). “From Multiple Modes for Surveys to Multiple Data Sources for Estimates”. In: *Survey Methodology* 40.2, pages 137–161.

Connelly, Roxanne et al. (2016). “The Role of Administrative Data in the Big Data Revolution in Social Science Research”. In: *Social Science Research* 59. Supplement C, pages 1–12.

Coumans, A. M. et al. (2017). “Estimating Homelessness in the Netherlands Using a Capture-recapture Approach”. In: *Social Indicators Research* 130.1, pages 189–212.

Daas, Piet J. H. et al. (2015). “Big Data as a Source for Official Statistics”. In: *Journal of Official Statistics* 31.2, pages 249–262.

Darroch, J. N. (1961). “The Two-sample Capture-recapture Census When Tagging and Sampling Are Stratified”. In: *Biometrika* 48.3/4, pages 241–260.

De Waal, Ton, Arnout van Delden, and Sander Scholtus (2017). *Multi-source Statistics: Basic Situations and Methods*. CBS Discussion Paper 2017 | 12 2017–12.

Efron, Bradley (1979). “Bootstrap Methods: Another Look at the Jackknife”. In: *Annals of Statistics* 7.1, pages 1–26.

Efron, Bradley and Robert Tibshirani (1994). *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC.

Enright, Bernard and Eugene J. OBrien (2011). *Cleaning Weigh-in-motion Data: Techniques and Recommendations*. Dublin Institute of Technology & University College Dublin.

Eurostat (2016). *Road Freight Transport Methodology: Manuals and Guidelines, 2016 edition*.

Fienberg, Stephen E. (1972). “The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables”. In: *Biometrika* 59.3, pages 591–603.

Ganguly, Auroop R. et al., editors (2009). *Knowledge Discovery From Sensor Data*. Boca Raton: CRC Press.

Hassounah, Mazen I., Loy-Sai Cheah, and Gerald N. Steuart (1993). “Underreporting of Trips in Telephone Interview Travel Surveys”. In: *Transportation Research Record* 1412, pages 90–94.

Huggins, R. M. (1989). “On the Statistical Analysis of Capture Experiments”. In: *Biometrika* 1.76, pages 133–140.

International Working Group for Disease Monitoring and Forecasting (1995). “Capture-recapture and Multiple Record Systems Estimation. Part I. History and Theoretical Development”. In: *American Journal of Epidemiology* 142.10, pages 1047–1058.

Japec, Lilli et al. (2015). “Big Data in Survey Research: AAPOR Task Force Report”. In: *Public Opinion Quarterly* 79.4, pages 839–880.

Klingwort, Jonas, Bart Buelens, and Rainer Schnell (2018). *Transport Survey Estimate Adjustment by Permanently Installed Highway-sensors Using Capture-recapture Techniques*. International Methodology Symposium 2018; Ottawa, Canada, 07.11.2018.

Klingwort, Jonas, Bart Buelens, and Rainer Schnell (2019a). "Capture-recapture Techniques for Transport Survey Estimate Adjustment Using Road Sensor Data". In: *Social Science Computer Review* (online first).

Klingwort, Jonas et al. (2019b). *Implementing Big Data in Official Statistics: Capture-recapture Techniques to Adjust for Underreporting in Transport Surveys Using Sensor Data*. New Techniques and Technologies for Statistics (NTTS): Brussels, Belgium. 12.03.2019.

Krishnamurty, Parvati (2008). "Diary". In: *Encyclopedia of Survey Research Methods*. Edited by Paul J. Lavrakas. Volume 1. Thousand Oaks: Sage, pages 197–199.

Krosnick, Jon A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys". In: *Applied Cognitive Psychology* 5.3, pages 213–236.

Lincoln, Frederick C. (1935). *The Waterfowl Flyways of North America*. Washington: United States Department of Agriculture.

Lohr, Sharon L. and Trivellore E. Raghunathan (2017). "Combining Survey Data with Other Data Sources". In: *Statistical Science* 32.2, pages 293–312.

Meyer, Bruce D., Wallace K. Mok, and James X. Sullivan (2015). "Household Surveys in Crisis". In: *Journal of Economic Perspectives* 29.4, pages 199–226.

Miller, Peter V. (2017). "Is There a Future for Surveys?" In: *Public Opinion Quarterly* 81.S1, pages 205–212.

Pearson, D. (2001). *Global Positioning System (GPS) and Travel Surveys: Results from the 1997 Austin Household Survey*. Paper presented at the Eighth Conference on the Application of Transportation Planning Methods. Corpus Christi, Texas.

Petersen, C. G. Joh. (1893). *On the Biology of Our Flat-fishes*. Kjøbenhaven: The Danish Biological Station.

Plante, Nathalie, Louis-Paul Rivest, and Gilles Tremblay (1998). "Stratified Capture-Recapture Estimation of the Size of a Closed Population". In: *Biometrics* 54.1, pages 47–60.

Richardson, A. J., E. S. Ampt, and A. H. Meyburg (1996). "Nonresponse Issues in Household Travel Surveys". In: *Conference Proceedings 10: Household Travel Surveys–New Concepts and Research Needs*. Edited by TRB National Research Council. Washington, pages 79–114.

Schnell, Rainer (2015). "Combining Surveys with Non-questionnaire Data: Overview and Introduction". In: *Improving Survey Methods: Lessons Learned from Recent Research*. Edited by Uwe Engel et al. New York: Routledge, pages 269–272.

Schnell, Rainer (2016). "Record Linkage". In: *The SAGE Handbook of Survey Methodology*. Edited by Christof Wolf et al. Thousand Oaks: Sage, pages 662–669.

Schnell, Rainer (2019a). "'Big Data' aus wissenschaftssoziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt". In: *Erklärende Soziologie und soziale Praxis*. Edited by Daniel Baron, Oliver Arránz Becker, and Daniel Lois. Wiesbaden: Springer, pages 101–125.

Schnell, Rainer (2019b). *Survey Interviews: Methoden standardisierter Befragungen*. 2nd edition. Wiesbaden: Sringer VS.

Shen, Li and Peter Stopher (2014). "Review of GPS Travel Survey and GPS Data-processing Methods". In: *Transport Reviews* 34.3, pages 316–334.

Shlomo, Natalie and Harvey Goldstein (2015). "Editorial: Big Data in Social Research". In: *Journal of the Royal Statistical Society, Series A* 178.4, pages 787–790.

Stopher, Peter, Camden FitzGerald, and Min Xu (2007). "Assessing the Accuracy of the Sydney Household Travel Survey with GPS". In: *Transportation* 37, pages 723–741.

Van der Heijden, Peter G. M. et al. (2017). "Dual and Multiple System Estimation: Fully Observed and Incomplete Covariates". In: *Capture-recapture Methods for the Social and Medical Sciences*. Edited by Dankmar Böhning, Peter G. M. van der Heijden, and John Bunge. Boca Raton: CRC, pages 213–227.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. 4th edition. New York: Springer.

Wolf, Jean, Marcelo Oliveira, and Miriam Thompson (2003). "Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-enhanced Household Travel Survey". In: *Journal of the Transportation Research Board* 1854, pages 189–198.

Wolf, Jean et al. (2013). “A Case Study: Multiple Data Collection Methods and the NY/NJ/CT Regional Travel Survey”. In: *Transport Survey Methods: Best Practice for Decision Making*. Edited by Johanna Zmud et al. EmeraldInsight, pages 71–90.
Wolter, Kirk M. (1986). “Some Coverage Error Models for Census Data”. In: *Journal of the American Statistical Association* 81.394, pages 337–346.

Colophon

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

Prepress
Statistics Netherlands

Design
Edenspiekermann

Enquiries
Telephone: +31 88 570 70 70
Via contact form: www.cbs.nl/infoservice