Statistical Journal of the IAOS 37 (2021) 1289–1299
DOI 10.3233/SJI-210821
IOS Press

Transition from survey to sensor-enhanced official statistics: Road freight transport as an example

Jonas Klingwort[a,*], Joep Burger[a], Bart Buelens[a,1] and Rainer Schnell[b]
[a]*Statistics Netherlands (CBS), Research and Development, CBS-weg 11, Heerlen, The Netherlands*
[b]*University of Duisburg-Essen, Research Methodology Group, Forsthausweg 2, Duisburg, Germany*

Abstract. Capture-recapture (CRC) is currently considered a promising method to integrate big data in official statistics. We previously applied CRC to estimate road freight transport with survey data (as the first capture) and road sensor data (as the second capture), using license plate and time-stamp to identify re-captured vehicles. A considerable difference was found between the single-source, design-based survey estimate, and the multiple-source, model-based CRC estimate. One possible explanation is underreporting in the survey, which is conceivable given the response burden of diary questionnaires. In this paper, we explore alternative explanations by quantifying their effect on the estimated amount of underreporting. In particular, we study the effects of 1) reporting errors, including a mismatch between the reported day of loading and the measured day of driving, 2) measurement errors, including false positives and OCR failure, 3) considering vehicles reported not owned as nonresponse error instead of frame error, and 4) response mode. We conclude that alternative hypotheses are unlikely to fully explain the difference between the survey estimate and the CRC estimate. Underreporting, therefore, remains a likely explanation, illustrating the power of combining survey and sensor data.

Keywords: Capture-recapture, measurement error, total survey error, underreporting, Weigh-in-Motion

1. Introduction

Sensor data is becoming increasingly important in official statistics because such data has some advantages over sample survey data: no declining response rates, no response burden, and continuous measurements in real-time [1–4]. Sensor data is currently rarely used to produce direct estimates due to the often unknown data generating process. However, sensor data can be used to assess underreporting bias in survey point estimates by linking them to survey data and applying capture-recapture (CRC) techniques. [5] illustrated this method using road sensor data and diary survey data to estimate road freight transport. Their CRC-estimates based on the combined sources were higher than the design-based, survey-only estimates. One possible explanation for the difference is underreporting in the survey, which is conceivable given the high response burden, especially in diary questionnaires.

In this paper, we study potential alternative explanations from which practical implications and guidelines for survey agencies and transportation management can be drawn. First, owners of vehicles in the sample are requested to fill out the day of loading, whereas the sensors measure the day of driving. Second, sensors measure vehicles that do not have to be reported (false positives), or Optical Character Recognition (OCR) failure prevents linkage of sensor records to survey reports. Third, vehicles reported not owned were considered a frame error but may alternatively be considered nonresponse. Finally, if underreporting would be the main explanation for the difference, it should be apparent only

[*] Corresponding author: Jonas Klingwort, Statistics Netherlands (CBS), Research and Development, CBS-weg 11, 6401 CZ Heerlen. Tel.: +31 455707070; E-mail: j.klingwort@cbs.nl.

[1] Current address: Flemish Institute for Technological Research (VITO), Boeretang 200, Mol, Belgium.

in the manual response modes but not in the automated response mode.

We see great potential for the use of CRC in assessing survey underreporting, as several European countries conduct road freight transport surveys [6,7] and have installed a Weigh-in-Motion road sensor network [8,9]. Our study provides a profound understanding of the CRC methodology and strengthens the trustworthiness in combining new data sources with established methodology in official statistics. Such empirical studies are of high practical importance for ultimately uptaking these methods in statistical production processes leading to a transition from experimental to official statistics [10,11].

2. Research background

Time-based diary surveys impose a heavy response burden. To reduce the response burden, respondents may respond inaccurately or not at all. Accordingly, such surveys often yield low response rates and downward-biased estimates [12,13]. Diary surveys on transport and mobility suffer from these drawbacks, since the target variables rely on accurate and complete responses [14]. New data sources for validation and estimation could improve official statistics based on diary surveys.

For studies on underreporting in transport and mobility surveys, units have been equipped with GPS receivers or mobile phones [15]. [16] reported for the first GPS household travel survey (1997, USA) underreporting in trip rates up to 31%. In the California Statewide Household Travel Survey, [17] reported rates of missed trips between up to 42%. In a comparative study, [18] documented levels of trip underreporting between 11% and 81% in GPS-based travel surveys in the USA. [19] reported only 7% underreporting for the Sydney Household Travel Survey (2004). The results of prior studies in the field of transport, travel, and mobility, show evidence for underreporting bias, although the amount varies both within and between the reported findings. Issues commonly occurring in these studies include technical problems with GPS devices and varying data quality between instrument types [20]. Further, problems due to switch-offs, delays, battery issues, or the device not being carried and difficulties matching recorded and reported data were reported by [21,22].

Here, instead of using GPS receivers, permanently installed road sensors are used to validate and adjust an underreporting bias in survey point estimates. Therefore, respondent-related issues such as described above can be neglected. This methodology has been first suggested by [5]. In this article, we report on further research to better understand differences in freight transport estimates with and without road sensor data.

3. Data

3.1. Survey data

The Dutch Road Freight and Transport Survey is a sample survey, conducted by Statistics Netherlands (CBS). A central aim of the survey is to estimate the total transported shipment weight (W) transported by Dutch commercial vehicles. In this paper, the total number of vehicle days (D) is considered an additional, simpler target variable (no measurement error correction required, see Subsection 3.2). One vehicle day is defined as a day that a vehicle has been on the road in the Netherlands.

The target population is the Dutch commercial vehicle fleet, excluding military, agricultural, and vehicles older than 25 years. Only vehicles with a weight of at least 3.5 t (empty vehicle weight + loading capacity) are taken into consideration [23]. The target population consists of about 135 thousand license plates and is updated quarterly.

Each quarter a stratified random sample was drawn totaling 33,817 unique vehicle-week combinations in 2015. Vehicle owners are legally required to report the days on which the vehicle was loaded and the corresponding shipment weight for one week. No report is required if the vehicle was driving all day without loading or unloading or if it was not driving for transport purposes. The effect of measuring these false positives by road sensors is one of the research questions addressed in this paper.

Of the 34 thousand vehicle weeks, 22,454 vehicles (66%) were reported used on at least one day during the assigned week, 5304 vehicles (16%) were reported not used during the assigned week, 2462 vehicles (7%) were reported not owned, and 3597 (11%) was nonresponse.

The option to report that the vehicle has not been used reduces the respondent's burden considerably since only small parts of the questionnaire have to be answered. That is the expected major cause of underreporting. Another way of responding with minimal burden is to report only a single day. The CRC approach allows assessing the severity of these two reporting errors.

Vehicles reported not owned are treated as nonresponse in the regular published official statistics. For the CRC study, such responses are defined as frame errors and excluded from analysis since the validation of such responses is not possible (only quarterly updates of the vehicle register, complexity in holding companies, vehicle rental/leasing). This decision was made to avoid false-positive links. An assumption made here is that all sample units classified as nonresponse own the vehicle, i.e., that if the vehicle is not owned this is always reported. The effect of this decision is one of the research questions addressed in this paper.

Owners of sampled license plates can respond by completing an internet questionnaire, a paper questionnaire, or by providing a data export (XML) from a software-based journey planning system commonly used by large haulers. The paper questionnaire is only available upon request and is used only by a minor fraction of the respondents. Generally paper questionnaires cause lower data quality compared to technology-based response modes [24]. We will refer to the internet and paper questionnaires as the manual response modes, and the planning system as the automated response mode. In this paper, we test the hypothesis that underreporting occurs only in the manual response modes.

3.2. Sensor data

The Weigh-in-Motion road sensor network's purpose is to detect and enforce penalties on overloaded trucks, tractors, and other heavy transport vehicles [25]. In 2015, nine sensor systems were operating in both directions on Dutch highways, resulting in 18 measurement points (Fig. 1).

When a vehicle passes a station, it is weighted, classified and a photograph of its front license plate is taken, which allows linkage to additional register information. A vehicle's individual axle weights are measured, summing to the total weight. Upward-biased axle measurements (over 16.1 t) were corrected using conditional mean imputation, i.e., replaced by the average across the axle measurements below 16.1 t per truck, based on the guidelines of [26] and expert information from the road administration. Transported shipment weight is calculated by subtracting the registered empty weight from the measured total weight. Negative values were trimmed to 0.

The empty trailer weight could not always be linked. The rear license plate was not recognized by the OCR software (11,340), or the trailer was not registered in the vehicle register (5980). If the front and rear license plates are identical, no trailer was pulled, and the trailer weight was set to 0. For the remaining missing values, conditional mean imputation was applied.

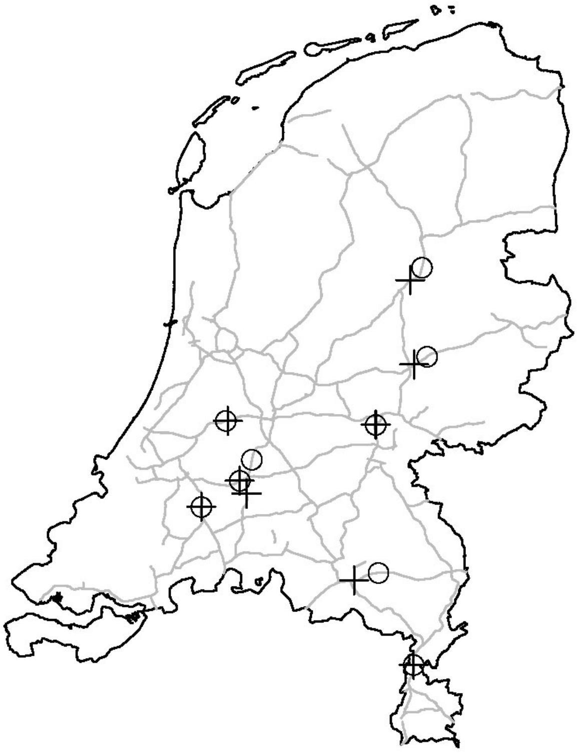

Fig. 1. Road sensor network on Dutch highways consisting of nine different systems with 18 measurement stations. Crosses and circles indicate the two stations of the nine WiM systems.

Table 1
Number of vehicle days D reported in survey and measured by sensors

D		Sensor		
		Measured	Not measured	$\sum$
Survey	Reported	34,284	60,522	94,806
	Not reported	9,727	?	?
	$\sum$	44,011	?	?

3.3. Linking survey and sensor data

After matching survey and sensor data by license plate and day as unique key, contingency tables can be constructed for D (Table 1) and W (Table 2). The total reported in the survey is the unweighted survey estimate. The quantities D and W for vehicles that were neither reported in the survey nor measured by the sensors are unknown, and are estimated by the CRC estimator.

Table 2
Transported shipment weight W (kt) reported in survey and measured by sensors

W (kt)		Sensor		
		Measured	Not measured	$\sum$
Survey	Reported	591	879	1470
	Not reported	139	?	?
	$\sum$	730	?	?

3.4. CRC assumptions

CRC was initially developed to estimate the unknown size of animal populations but has been applied to human populations [27] and is based on the following assumptions: independent datasets, closed population, homogeneous capture probabilities, all elements belong to the target population, and perfect linkage of datasets.

The first assumption of independent datasets and the second assumption of a closed population are fulfilled. The probability that a sampled vehicle is reported in the survey is independent of the probability that it is measured by a sensor. By equating the study population to the survey sample, the population is closed by definition.

Third, the response or capture probabilities for the elements should be homogeneous for at least one data source [28,29]. This assumption is met by modeling capture probabilities as a function of auxiliary information (see Subsection 4.3).

Fourth, elements in the data sources should belong to the population of interest. This assumption is met because sampled vehicles belong to the population by definition and vehicles detected by the sensors are filtered by linking their license plate to the register. Vehicles might not belong to the target population on specific days and in specific situations, but we consider these violations of the fifth assumption of perfect linkage.

The fifth assumption addresses the perfect linkage of the elements. It is the strength of the WiM system that vehicles can be linked by license plate. However, this assumption may occasionally be violated in both data sources for different reasons. Vehicle owners might report too few, too many or the wrong dates and they have to report the day of loading rather than the day of driving. Sensors also measure vehicles that are driving but not transporting, and OCR failure prevents linkage. These potential violations of the fifth assumption are addressed in this paper.

3.5. Register data

To model capture probabilities, variables from the vehicle and business registers are used. The vehicle register provides both technical and non-technical vehicle features. The business register provides 5 features about the vehicle owner.

Register data are linked at micro-level using license plate and quarter as unique key. For some sample units, no register information could be linked. For variables with rather small proportions of missing values, the missing values were replaced with the most common category (mode imputation). Otherwise, the category 'Unknown' was assigned. Continuous variables were then categorized based on their unweighted response quantiles.

4. Methods

4.1. Definitions and notations

Underreporting is estimated by comparing survey estimates corrected for selective nonresponse with CRC estimates corrected for both selective nonresponse and measurement error.

We define the indicator $\delta_{ij}^{svy} = 1$ if vehicle i was reported used in the survey on day j and 0 otherwise. We define $\delta_{ij}^{sen} = 1$ if vehicle i was measured at least once by a sensor on day j and 0 otherwise. We define Ξ_{ij}^{svy} and Ξ_{ij}^{sen} as the shipment weight of vehicle i on day j reported in the survey and measured by the sensors, respectively. If multiple weights per day were reported or measured, the maximum was used. Measures such as mean or median would be feasible as well, but by choosing the maximum, the estimated amount of underreporting is considered a conservative estimate.

Underreporting in the survey is estimated as the relative difference RD between a survey estimate $\widehat{Y}^{SVY}$ and a CRC estimate $\widehat{Y}^{CRC}$, using the CRC estimate as benchmark:

$$RD = \frac{\widehat{Y}^{SVY}}{\widehat{Y}^{CRC}} - 1. \tag{1}$$

4.2. Survey estimators

The survey estimators for the total number of truck days D and transported shipment weight W in a sample-week are post-stratification estimates, taking into account the sampling design and correcting for selective nonresponse:

$$\widehat{D}^{SVY} = \sum_{i=1}^{r} \left(w_i \sum_{j=1}^{7} \delta_{ij}^{svy} \right), \tag{2}$$

$$\widehat{W}^{SVY} = \sum_{i=1}^{r} \left(w_i \sum_{j=1}^{7} \delta_{ij}^{svy} \Xi_{ij} \right), \tag{3}$$

with w_i the survey weight for vehicle i, and r the number of respondents. The w_i is based on the initial post-stratification weight w_i^+ [23]:

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

where N_h^+ is the total number of vehicles in post-stratum h including vehicles reported not owned and r_h the number of respondents in post-stratum h excluding vehicles reported not owned (treating reported not owned as nonresponse). The factor 13 scales up from week to quarter. Here, we treat reported not owned as frame error and scale up the response only to the sample. The w_i^+ were therefore rescaled to:

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

so that $\sum_{i=1}^{r} w_i = n$, excluding vehicles not owned, with n being the sample size. The effects of excluding vehicles reported not owned are reported in Subsection 4.8.

4.3. Capture-recapture estimators

Log-linear models for population size estimation in closed populations were introduced by [30]. They can be written as a generalized linear model, where count y_i in cell i of the 2×2 contingency table (see Tables 1 and 2) is assumed to follow a Poisson distribution:

$$y_i \sim \text{Poisson}(\lambda_i), \tag{5}$$

where λ_i is the Poisson parameter, which is modeled on a logarithmic scale so that it is always positive:

$$\log \lambda_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2},$$

where X_{i1} is an indicator for being reported in the survey (1 for cells $(1,1)$ and $(1,2)$, 0 otherwise), X_{i2} an indicator for being measured by a sensor (1 for cells $(1,1)$ and $(2,1)$, 0 otherwise) and β the parameters to be estimated. Since the count in cell $(2,2)$ is unknown and the target of the study, the model has no interaction term $X_{i1}X_{i2}$ and capture probabilities in one source are assumed independent from those in the other source (a realistic assumption as discussed in Subsection 3.4). After fitting the model on the three observed cells, the unobserved cell $(2,2)$ can be estimated by $e^{\hat{\beta}_0}$.

The model is extended with auxiliary information from the register, to model heterogeneity in the capture probabilities in both sources:

$$\log \boldsymbol{\lambda} = \boldsymbol{X\beta}, \tag{6}$$

where $\log \boldsymbol{\lambda}$ is an m-dimensional column vector with cell counts, $\boldsymbol{X}$ an $m \times p$-dimensional matrix with covariates and $\boldsymbol{\beta}$ a p-dimensional column parameter vector. The first column in $\boldsymbol{X}$ is a vector of 1s to estimate the intercept, the second column the indicator for being reported in the survey, and the third column the indicator for being measured by a sensor ($m = p = 3$ if no covariates are added). Each k-class categorical covariate adds $2(k-1)$ columns for first-order interactions with the survey and sensor indicators and increases the number of rows m by a factor k. In practice, the number of rows will be limited as empty cells ($y_i = 0$) are ignored. The independence assumption is now conditioned on the covariates.

The CRC estimator for target variable Y (D or W) is the sum of the expected counts in the m cells observed in the survey, by the sensors or both, and the estimated counts observed neither in the survey nor by the sensors:

$$\widehat{Y}^{CRC} = \mathbf{1} e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}} + e^{\hat{\beta}_0}, \tag{7}$$

where $\mathbf{1}$ is a row vector of 1s of dimension m, for summing.

4.4. Model selection

A stepwise selection procedure based on the Bayesian Information Criterion (BIC) was used to choose an optimal model for the CRC estimator [31]. Only main effects of the variables were considered to avoid sparse contingency tables for the log-linear model. The model selection for the log-linear model is based on a logit-model selection, to cover the full information of the covariates. Using the approach proposed by [32,33], two independent logit-models were applied: using δ_{ij}^{svy} and δ_{ij}^{sen} as dependent variables predicting the capture probabilities for each source independently. For the log-linear model, the five variables with the most explanatory power in the two logit-models were combined. If necessary, the count distribution of the chosen variable was used to categorize each of the variables into: 'low count', 'medium count', and 'high count'. The chosen variables are maximum mass of pulled trailer (5 categories), province of vehicle owner (12 categories), classification of economic activity (2 categories), year of manufacture, vehicle type (2 categories), size of the vehicle fleet (5 categories), and power (kW). In addition, the number of operating stations and a weekend indicator were included in the models (both 2 categories).

4.5. Precision estimation

To estimate the precision of $\widehat{Y}^{SVY}$ and $\widehat{Y}^{CRC}$, percentile bootstrapping was applied. $B = 3000$ samples were drawn with replacement, each of size equal to the original linked dataset. The 2.5th and 97.5 percentiles of the bootstrap estimates were used to estimate the 95% confidence interval. Potential bias was checked by comparing the bootstrap mean and the estimate based on the original data.

4.6. Reporting errors

Survey respondents must fill out the day of loading instead of the day of driving. When comparing survey and sensor data, this may be considered a reporting error by design. Two systematic reporting errors were simulated: underreporting and overreporting error. The simulation study was restricted to target variable D. For W, the reported shipment weights would have to be distributed over more or less days, which would require assumptions that are hard to verify.

If vehicle i was reported used on day j, the indicator takes the value 1, and 0 otherwise.

It follows,

$$0 \leqslant \sum_{j=1}^{7} \delta_{i,j}^{svy} \leqslant 7.$$

The response pattern can be written as a string of 7 indicators $\delta_{i,j}^{svy}$, giving a range from 0000000 (not used) to 1111111 (used on every day of the week). The following rules were used to simulate the errors:

1. Underreporting error: replace each trailing 0 with a 1. For instance, 0010000 becomes 0011111. This new pattern attempts to correct for the questionnaire asking about the day of loading instead of the day of driving, and for respondents pooling multiple days of loading to the first day.
2. Overreporting error: replace each trailing 1 with a 0. For instance, 0111010 becomes 0100010. This new pattern attempts to correct for overreporting, which is not our main concern but a control that should show the opposite effect of underreporting.

The response pattern was replaced for 1% to 100% of the respondents, repeating each sample 100 times. Changing 100% of the data implies all reported response patterns being erroneous, which is very unlikely. The results are considered an approximation of an upper error limit.

4.7. Measurement errors

Errors in the sensor observations were simulated to evaluate the effect of linking false positives (FP) and OCR failure. In contrast to the 'reporting error' simulation, both D and W were considered since shipment weights do not have to be distributed over days. False positives arise when vehicles are measured that do not drive for transport purposes (e.g., empty journeys, journeys for maintenance, or fueling) or when the reported day of loading is not the measured day of driving. The effect of FP was simulated by removing units from m_{21}, i.e., the number of vehicle days or transported shipment weight that was not reported in the survey but was measured by the sensors. The numbers and weights reported in the survey (m_{11} and m_{12}) were left unchanged.

As already reported in Subsection 3.2, about 30% of the sensor observations had no recognized license plate. The effect of OCR failure was simulated by moving units from measured to not measured by the sensors, independent from what was reported in the survey. As a result, both m_{11} and m_{21} were decreased by the same percentage, i.e., the number of vehicle days or transported shipment weight measured by the sensors. To leave unchanged what was reported in the survey, the amount reported in the survey but not measured by the sensors, m_{12}, was increased by the same amount that m_{11} was decreased.

The fraction FP and OCR failure was simulated from 1% to 100%. For every simulated step, observations were randomly dropped. To estimate the sampling error, each step was repeated 100 times. Changing 100% of the sensor observations in both setups implies the data source being completely unreliable, which is very unlikely.

4.8. Reported not owned

Vehicles reported not owned were treated as frame error and excluded because we do not know whether or not they can be measured by the sensors. If they have been scrapped or exported, the sensors cannot measure them. The decision to exclude these vehicles implies that all sample units classified as nonresponse are assumed to own the vehicle, i.e., that if the vehicle is not owned, then it is assumed that this is always reported. On the other hand, vehicles reported not owned are treated as nonresponse error in the published official statistics. To compare these two visions, we re-estimated the amount of underreporting when treating vehicles reported not owned as nonresponse error. This means that the weights of the respondents were rescaled to

$$w_i = w_i^+ \frac{n^+}{\sum_{i=1}^{r} w_i^+} \tag{8}$$

Table 3
Survey estimates, CRC estimates and the estimated underreporting (%) of the number of vehicle days D and transported shipment weight W (kt)

Estimator	Point estimate	Bootstrap mean	Bootstrap standard error	Bootstrap 95% CI	Estimated underreporting (%)	Bootstrap 95% CI
$\widehat{D}^{SVY}$	102,273	102,266	408	[101,474, 103,059]	-18.4	[-19.2, -17.6]
$\widehat{D}^{CRC}$	125,327	125,350	619	[124,125, 126,572]		
$\widehat{W}^{SVY}$	1499	1499	9.4	[1481, 1518]	-18.3	[-19.3, -17.4]
$\widehat{W}^{CRC}$	1835	1835	11	[1815, 1857]		

so that $\sum_{i=1}^{r} w_i = n^+$ including vehicles not owned (compare Eq. (4)). This increases m_{21} by 23% from 9727 to 11,946 vehicle days or by 22% from 139 to 169 kt. Note that this is not the opposite of the FP simulation (Subsection 4.7), because the vehicles reported not owned are not a random sample.

4.9. *Response mode*

To test the hypothesis that the difference between survey estimates and CRC estimates is due to manual underreporting in the survey, we stratified both estimators by response mode. Two strata were formed: manual (internet and paper questionnaires) and automated response mode (journey planning system). The weights w_i were rescaled to sum up to the size of the corresponding strata.

5. Results

5.1. *Total*

According to the CRC estimator, the weighted survey estimator underestimates both D and W by about 18% (Table 3). The nonresponse correction by the survey estimator adds about 8% to the unweighted D (94,806 in Table 1) and about 2% to the unweighted W (1470 kt in Table 2). Since the CRC estimators correct for both nonresponse and measurement error, the most likely explanation is that the underestimation by the survey estimators is due to underreporting. The remaining part of the paper is dedicated to alternative explanations.

If the mean or median were used in case of multiple recordings/reports per day (see Subsection 4.1), the underestimation by $\widehat{W}^{SVY}$ would increase from 18% to 25% or 27%, respectively.

The fit of the selected log-linear model is moderate. Using a likelihood ratio test comparing deviances showed that the selected model had a better fit than the null model. Using Pearson correlation coefficient, the linear relationship between the actual and predicted frequencies for D is moderate with $r = 0.7$ andstrong for W with $r = 0.8$. For additional details on the model fit see [34].

5.2. *Reporting errors*

If the reported day of loading would result in days of driving on all remaining days of the reporting week, both the survey estimate and the CRC estimate for D would increase (underreporting error in Fig. 2). The absolute relative difference would decrease from 18.4 (Table 3) to at least 10.1 if the underreporting error was simulated in all vehicles. The survey estimate increases linearly with the proportion of vehicles for which this underreporting error would be corrected. The increase in the CRC estimate, however, levels off. The relative difference, therefore, remains fairly constant at low proportions. Thus, the mismatch between the reported day of loading and the measured day of driving can only partially explain the relative difference, and only if it occurs in a high proportion of vehicles.

In the simulation correcting for overreporting errors, the survey estimate would decrease linearly with the proportion of vehicles for which this error would be corrected (overreporting error in Fig. 2). The CRC estimate, however, is fairly robust and only steeply increases at high proportions. As a result, the absolute relative difference only increases. Obviously, overreporting cannot explain the relative difference.

5.3. *Measurement errors*

The relative difference between the survey estimate and the CRC estimate gradually vanishes as more units are considered false positives (FP in Fig. 3). Recall that in this simulation, units were removed that were not reported in the survey but were measured by the sensors (cell $(2, 1)$ in Tables 1 and 2). A similar pattern appears for both the number of vehicle days D and the transported shipment weight W. If all events measured by the sensors would have been reported in the survey, the CRC estimates reduce to the unweighted survey estimates (Tables 1 and 2) and the relative difference

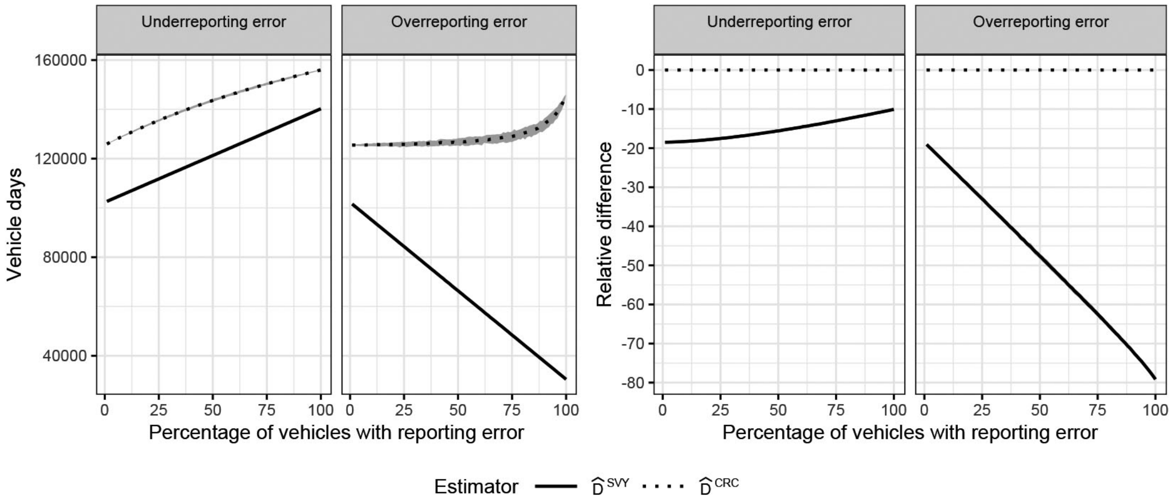

Fig. 2. Effect of simulated percentage of vehicles with reporting errors on the survey and CRC estimate of the number of vehicle days D (left) and their relative difference (right). Shown are bootstrapped means and 95% percentile confidence intervals.

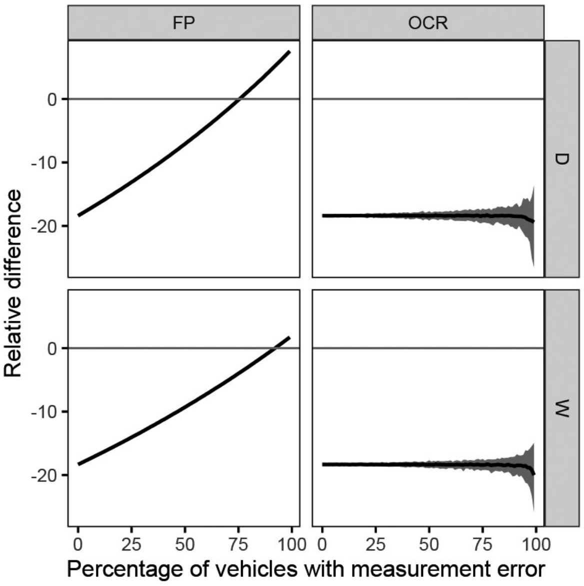

Fig. 3. Effect of simulated percentage of FP (left panels) and OCR failure (right panels) on the estimated relative difference between the survey and CRC estimate of the number of vehicle days D (top panels) and transported shipment weight W (bottom panels). Shown are bootstrapped means and 95% percentile confidence intervals.

reduces to the nonresponse error. Assuming that the weighting model is effective, a positive relative difference (above the null line) can, therefore, be attributed to nonresponse errors and a negative relative difference (below the null line) to measurement error. The weighted survey estimate and the CRC estimate would agree if 76% (D) to 93% (W) of the units in cell $(2, 1)$ would be considered false positives. If a better weighting model could correct for more selective nonresponse, the null line would cross at a lower, presumably more realistic percentage.

Thus, false positives can explain the relative difference between survey and CRC estimate, but the estimated underreporting only disappears if the sensors would measure mostly vehicles that do not have to be reported in the survey, or if the weighting model is not effective.

The relative difference between the survey estimate and the CRC estimate remains the same as more OCR failures are simulated (OCR in Fig. 3). Only the precision is compromised. Recall that in this simulation, units were moved from measured to not measured by the sensors, irrespective of whether or not they were reported in the survey (Subsection 4.7). Thus, OCR failure cannot explain the relative difference. This can also be shown analytically for a simple CRC estimator, such as the Lincoln-Petersen estimator [34].

5.4. *Reported not owned*

If the 2462 units reported not owned in the survey are considered a nonresponse error instead of a frame error, the nonresponse correction by the survey estimator increases from 8% (Subsection 5.1) to 16% for D (110,300) and from 2% to 8% for W (1616 kt). The CRC estimate, however, also increases. As a result, the absolute relative difference between the survey estimate and the CRC estimate decreases only moderately from 18.4% to 16.0% for D and from 18.3% to 15.3% for W (Fig. 4). Thus, treating vehicles reported not owned as

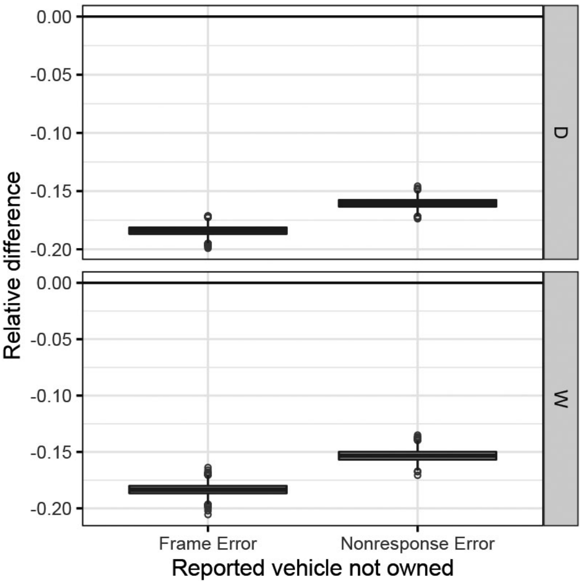

Fig. 4. Effect of treating vehicles reported not owned as frame error or nonresponse error on the estimated relative difference between the survey and CRC estimate of the number of vehicle days D (top panel) and transported shipment weight W (bottom panel). Shown are boxplots of bootstrapped estimates.

frame error is not a plausible explanation for the relative difference.

In contrast to reporting errors and measurement errors, no simulation is required because the number of vehicles reported not owned measured by the sensors is known (2219 vehicle days or 30.2 kt). If we nevertheless simulate that none of them is detected by the sensors, the absolute relative difference would decrease to about 12% (not shown). However, if a sensor detects none, it makes more sense to treat them as frame error. If simulating all being detected, the relative difference would not change.

5.5. Response mode

The relative difference between the survey and CRC estimate is considerably larger in the manual response modes than in the automated response mode (Fig. 5). This finding supports our hypothesis that the difference is caused by underreporting in the survey. The relative difference is, however, still negative in the automated response mode. Thus, according to this analysis, about 13%-point (D) to 11%-point (W) of the relative difference between the survey and CRC estimate based on the manual response modes can be attributed to underreporting. The remaining 6%-point (D) to 8%-point (W) may be underreporting in the journey planning system or may have other unknown causes.

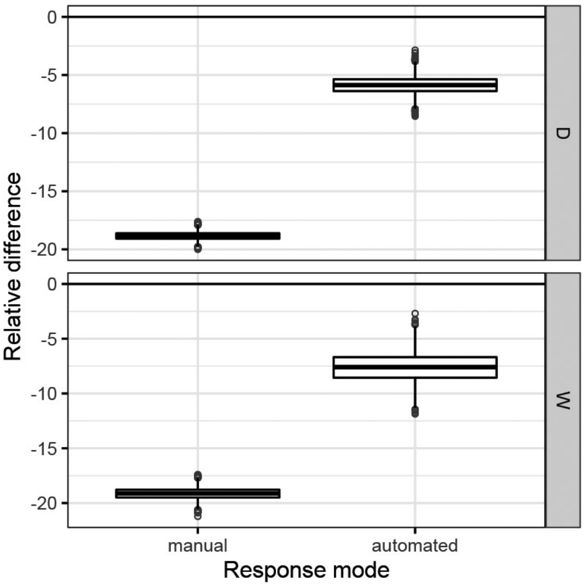

Fig. 5. Effect of response mode on the estimated relative difference between the survey and CRC estimate of the number of vehicle days D (top panel) and transported shipment weight W (bottom panel). Shown are boxplots of bootstrapped estimates.

6. Discussion and conclusion

With the rare opportunity to link survey, sensor and register data using a deterministic key, we presented a comprehensive study to better understand the use of sensor data and capture-recapture in official statistics. We studied several sources of error, potentially biasing log-linear CRC estimates on road freight transport. As evident from the literature, we also showed that different levels of underreporting can be found within a study on underreporting, depending on study design. However, even in demonstrated implausible scenarios, the estimated amount of underreporting did not decrease below 10%. We explored four alternatives to underreporting as an explanation for a difference between the survey and CRC estimates of road freight and transport.

First, vehicle owners are asked to report the day of loading, whereas the sensors measure the day of driving. To study the effect of this mismatch, we simulated that the vehicle was driving on all days following the reported day of loading. Although this decreased the estimated difference between the survey and CRC estimate, the difference would not drop below about 10%. The opposite, overreporting error corrected for by collapsing multiple days of loading to the first reported day, only increased the difference.

Second, putative underreporting by survey respondents could be due to over-detection by sensors. Detected vehicles may not have to be reported, for instance, because they are empty or drive for maintenance. Unlikely high proportions of false positives can explain the difference completely, but the difference remains substantial at more reasonable, albeit unknown rates. The difference can be shown to be robust against linkage errors and sensor failure, although precision is compromised.

Third, vehicles reported not owned were considered as a frame error, assuming they have been scrapped or exported. Treating them as nonresponse error only explained about 2% to 3%-point of the relative difference between the survey and CRC estimates. Moreover, the less are measured by the sensors, the more the relative difference can be explained by treating them as nonresponse error. However, if it is more likely they have been scrapped or exported, they should be treated as frame error.

Fourth, if the difference would be caused by underreporting in the survey, this would only be apparent in the manual response modes (web and paper) and not in the automatic response mode. Estimating the difference by response mode indeed showed that the difference is much lower in the automated than in the manual modes. The remaining difference in automated mode, however, suggests that 11%–13% can be attributed to underreporting. The remaining difference in automated mode could, for instance, be a nonresponse error not corrected for by post-stratification. The CRC model selection supports this hypothesis by choosing variables currently not included in the survey post-strata. On the other hand, it could still be underreporting, as the data is entered manually into this system by humans. Finally, this finding might be confounded by variables such as company size. As small companies do not use the automated mode, it is not possible to control for such a confounding variable.

Limitations of this study are first, that we were not able to estimate the amount of false positives links. This error source is considered of high importance since the CRC estimates are highly sensitive to such error. Second, multiple explanations were not studied simultaneously but individually. Regarding generalizability, future research could apply the CRC estimates and simulations to other systems, or other data generated in different years of the system described here.

This study also provides approaches to discussions in the area of social and political implications, especially on freight transport management. Such research is the basis for additional research to derive recommendations for transport (infrastructure) management, economic/financial implications, emission, and climate change.

We conclude that sensor data, in combination with the CRC estimator, provides a valid tool to assess underreporting in survey questionnaires. We consider the contribution of this study a useful reference for official statistics, survey agencies, transportation management, and statisticians if survey, sensor, and register data can be linked, and CRC can be applied.

References

[1] Buelens B, Boonstra H, van den Brakel J, Daas P. Shifting Paradigms in Official Statistics: From Design-based to Model-based to Algorithmic Inference. Statistics Netherlands (CBS), Discussion Paper (201218), The Hague/Heerlen. 2012.

[2] Alwin DF. Reflections on thirty years of methodology and the next thirty. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique. 2013; 120(1): 28–37. doi: 10.1177/0759106313497855.

[3] Couper MP. Is the sky falling? New technology, changing media, and the future of surveys. Survey Research Methods. 2013; 7(3): 145–156. doi: 10.18148/srm/2013.v7i3.5751.

[4] Rao JNK, Fuller WA. Sample survey theory and methods: Past, present, and future directions. Survey Methodology. 2017; 43(2): 145–160.

[5] Klingwort J, Buelens B, Schnell R. Capture-recapture Techniques for Transport Survey Estimate Adjustment Using Road Sensor Data. Social Science Computer Review. 2019; 39(4): 527–542. doi: 10.1177/0894439319874684.

[6] Eurostat. Methodologies Used in Surveys of Road Freight Transport in Member States, EFTA and Candidate Countries Revised, 2017 Edition. Luxembourg: Publications Office of the European Union; 2018. doi: 10.2785/262283.

[7] Eurostat. Road Freight Transport Statistics: Statistics Explained. Luxembourg: Publications Office of the European Union; 2020.

[8] Jacob B, OBrien EJ. Weigh-in-motion: Recent Developments in Europe. In: 4th International Conference on Weigh-in-Motion-ICWIM4. Taiwan; 2005. pp. 1–11.

[9] Jacob B, van Loo H. Weigh-in-motion for Enforcement in Europe. In: Proceedings of the 10th International Symposium on Heavy Vehicle Transportation Technology. Paris; 2008. pp. 15–24.

[10] De Broe S, Meijers R, Ten Bosch O, Buelens B, Priem A, Laevens B, et al. From experimental to official statistics: The case of solar energy. Statistical Journal of the IAOS. 2019; 35(3): 371–385.

[11] Carciotto A, Signore M. Improving Relevance: Istat Experience on Experimental Statistics. Statistical Journal of the IAOS. 2021; 37: 593–601. doi: 10.3233/SJI-200764.

[12] Richardson AJ, Ampt ES, Meyburg AH. Nonresponse Issues in Household Travel Surveys. In: TRB National Research Council, editor. Conference Proceedings 10: Household Travel Surveys-New Concepts and Research Needs. Washington; 1996. pp. 79–114.

[13] Krishnamurty P. Diary. In: Lavrakas PJ, ed. Encyclopedia of Survey Research Methods. vol. 1. Thousand Oaks: Sage; 2008. pp. 197–199.

[14] Meyburg AH, Rahman S. The Challenges of Freight and Commercial Transport Surveys. In: Stopher P, Jones P, eds. Transport Survey Quality and Innovation. Bingley: Emerald Group Publishing Limited; 2003. pp. 443–454.

[15] Wang Z, He SY, Leung Y. Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society. 2018; 11: 141–155. doi: 10.1016/j.tbs.2017.02.005.

[16] Pearson D. 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]; 2001.

[17] Wolf J, Oliveira M, Thompson M. Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. Journal of the Transportation Research Board. 2003; 1854: 189–198. doi: 10.3141/1854-21.

[18] Bricka S, Bhat C. Comparative analysis of global positioning system-based and travel survey-based data. Transportation Research Record: Journal of the Transportation Research Board. 2006; 1972: 9–20. doi: 10.1177/0361198106197200102.

[19] Stopher PR, Greaves SP. Household travel surveys: Where are we going? Transportation Research. 2007; Part A(41): 367–381. doi: 10.1016/j.tra.2006.09.005.

[20] Sun QC, Odolinski R, Xia JC, Foster J, Falkmer T, Lee H. Validating the efficacy of GPS tracking vehicle movement for driving behaviour assessment. Travel Behaviour and Society. 2017; 6: 32–43. doi: 10.1016/j.tbs.2016.05.001.

[21] Bricka S, Sen S, Paleti R, Bhat CR. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation Research. 2012; Part C(21): 67–88. doi: 10.1016/j.trc.2011.09.005.

[22] Shen L, Stopher P. Review of GPS travel survey and GPS data-processing methods. Transport Reviews. 2014; 34(3): 316–334. doi: 10.1080/01441647.2014.903530.

[23] Centraal Bureau voor de Statistiek. Basisbestanden Goederenwegvervoer 2015 [CBS publicaties, The Hague/Heerlen]; 2017.

[24] Roddis S, Winter S, Zhao F, Kutadinata R. Respondent preferences in travel survey design: An initial comparison of narrative, structured and technology-based travel survey instruments. Travel Behaviour and Society. 2019; 16: 1–12. doi: 10.1016/j.tbs.2019.03.003.

[25] Federal Highway Administration. Effective Use of Weigh-in-motion Data: The Netherlands Case Study [Office of International Programs. FHWA/US DOT (HPIP). Publication No. FHWA-PL-07-028 HPIP/10-07(3.5)EW]; 2007.

[26] Enright B, OBrien EJ. Cleaning Weigh-in-motion Data: Techniques and Recommendations [Dublin Institute of Technology & University College Dublin]; 2011.

[27] International Working Group for Disease Monitoring and Forecasting. Capture-recapture and Multiple Record Systems Estimation I: History and Theoretical Development. American Journal of Epidemiology. 1995; 142(10): 1047–1058. doi: 10.1093/oxfordjournals.aje.a117558.

[28] Zwane EN, van der Heijden PGM. Semiparametric models for capture-recapture studies with covariates. Computational Statistics & Data Analysis. 2004; 47(4): 729–743. doi: 10.1016/j.csda.2003.11.010.

[29] van der Heijden PGM, Cruyff M, Whittaker J, Bakker BFM, Smith PA. Dual and Multiple System Estimation: Fully Observed and Incomplete Covariates. In: Böhning D, van der Heijden PGM, Bunge J, eds. Capture-recapture Methods for the Social and Medical Sciences. Boca Raton: CRC; 2017. pp. 213–227. doi: 10.4324/9781315151939.

[30] Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika. 1972; 59(3): 591–603. doi: 10.2307/2334810.

[31] Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research. 2004; 33(2): 261–304. doi: 10.1177/0049124104268644.

[32] Huggins RM. On the statistical analysis of capture experiments. Biometrika. 1989; 1(76): 133–140. doi: 10.2307/2336377.

[33] Alho JM. Logistic regression in capture-recapture models. Biometrics. 1990; 46: 623–635. doi: 10.2307/2532083.

[34] Klingwort J. Correcting Survey Measurement Error With Big Data from Road Sensors Through Capture-recapture. Doctoral Thesis. Faculty of Social Sciences. University of Duisburg-Essen; 2020. doi: 10.17185/duepublico/72081.