Proceedings of Statistics Canada Symposium 2018
Growth in Statistical Information: Challenges and Benefits

Transport Survey Estimate Adjustment by Permanently Installed Highway-sensors Using Capture-recapture Techniques

Jonas Klingwort, Bart Buelens, and Rainer Schnell[1]

Abstract

The integration of sensor data in official statistics is in particular valuable if it can be linked with survey and administrative data. Such datasets of the Netherlands are linked in this application one-to-one using a unique identifier to quantify and adjust underreporting in survey point estimates. The survey sample consists of registered truck owners who report trips and shipment weight. The sensors measure continuously every passing truck on certain highway stations. Capture-recapture techniques are used to estimate underreporting in the survey. Heterogeneity in capture and recapture probabilities is modeled through logistic regression and log-linear models. Results show the approach being promising in terms of validating and adjusting survey point estimates using external sensor data.

Key Words: Big Data, Record linkage, Data Validation, Underreporting, Multisource estimation, Weigh-in-motion

1. Introduction

Producing unbiased estimates in official statistics based on survey data becomes more difficult and expensive. Accordingly, research on methods using big data for the production of official statistics is currently increasing (Daas et al. 2015). Up to now, big data is rarely used in statistical production due to its unknown data generating process (Buelens et al. 2014). However, in the long-term, using big data in official statistics is necessary (Lohr and Raghunathan 2017). Therefore, instead of using single big data sources, research on combining different probability and non-probability based datasets is a promising approach to use big data in official statistics (Shlomo and Goldstein 2015). More specifically, the different problems of surveys and big data might be minimized if a survey and a sensor (collecting big data) measure an identical target variable and the resulting micro-data can be linked with a unique identifier. Using this principle, we link survey, sensor, and administrative data for transport statistics. Using the linked dataset, we apply capture-recapture techniques (CRC) to validate, estimate and adjust a bias in survey point estimates due to underreporting in the target variables of the survey.

2. Research background

The number of surveys conducted has increased over the last decades (Singer 2016), but at the same time the nonresponse rates have increased, too (Meyer et al. 2015). In particular, diary surveys impose a heavy response burden and yield very low response rates (Krishnamurty 2008). In the past, mobility and transport diary surveys have been validated and adjusted using GPS data. It has been shown that these surveys are often downward biased due to underreporting, varying between 2.6% (Hassounah et al. 1993) and 81% (Bricka and Bhat 2006). Those studies used mobile GPS devices attached to vehicles or respondents. 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 (Bricka, Sen, et al. 2012; Shen and Stopher 2014). Instead of using mobile GPS devices we use permanently installed road sensors to validate and adjust survey estimates.

[1]Jonas Klingwort, University of Duisburg-Essen & Statistics Netherlands, Forsthausweg 2, Germany, 47057 Duisburg & CBS-weg 11, Netherlands, 6412 EX Heerlen, jonas.klingwort@uni-due.de; Bart Buelens, VITO NV, Boeretang 200, Belgium, 2400 MOL; Rainer Schnell, University of Duisburg-Essen, Forsthausweg 2, Germany, 47057 Duisburg

3. Data

The target population of the Road Freight Transport Survey of the Netherlands (2015) is the Dutch commercial vehicle fleet, excluding military, agricultural and commercial vehicles older than 25 years (≥3.5*t* weight, ≥2*t* loading capacity). The sample consists of 33,817 trucks sampled from the national vehicle register. A central objective of the mandatory diary survey is to collect data on the shipment's weights transported by the trucks. Therefore, truck owners must report the days on which the truck was used and the corresponding shipment weight. 3,597 cases are classified as nonresponse. The answer categories regarding truck-related activities are: truck used (22,454), truck not used (5,304), and truck not owned (2,462). The latter case is defined as technical-nonresponse and is excluded from the analysis because the validity of the response cannot be verified. Underreporting is expected due to nonresponse and misreporting by falsely responding that the truck was not used. The sensor data is collected by the weigh-in-motion road sensor network (WIM) operated by the national road administration of the Netherlands consisting of 18 measurement stations. While passing, the vehicle's weight is measured. The sensors do not cover all highways in the Netherlands, though are installed at locations with a high traffic volume and at logistical hubs. In 2015 there were 35,669,347 trucks recorded of which, using the unique combination of license plate and day as identifier, 44,011 could be linked one-to-one to reported trips in the survey. Data quality checks and cleaning were applied following the guidelines developed by Enright and OBrien (2011). Corrections of measurement errors on the axle weights were applied using a conditional mean imputation. Using a deterministic error correction rule, the weight of an axle is imputed by the average weight of the remaining axles if the measured weight is greater than 20t. If the weight of more than one axle exceeds 20t, the average value of the remaining axles with a weight of less than 20t is used here, too. Predictive modeling using a linear regression ($r^2 = adj.r^2 = 0.54$) was applied to correct the weights for trucks driving outside the recommended speed interval [60;120] km/h. In 17,321 of the 44,011 matched trucks, no trailer could be linked to the truck. For 11,341 cases the OCR detection failed and in 5,980 the trailer was not listed in the register. The missing weights were imputed with the mean of the empty trailer weight, conditional on the automated classification of the truck and its loading capacity. The Dutch vehicle and enterprise register are linked to the data on a micro-level using the combination of license plate and annual quarter as match variable. Since the sensors measure the weight of the entire unit (truck, trailer, and shipment) the truck and trailer weights were subtracted using information from the vehicle register. The resulting value is the transported shipment weight, which corresponds to the definition of reported weight in the survey. In 3,945 cases negative shipment weights resulted, which were set to 0. Finally, an overall proportional bias correction was applied, calibrating the sensor measured shipment weights to those reported in the survey. The correction factor was obtained from the subset of vehicles that were observed both in the survey and by the sensors. This resulted in a downscaling of the sensor shipment weights by approximately 14%. Observations with missing register data was excluded from analysis (which explains the difference between the 44,011 matches and the 43,775 matches in Table 4.2-1).

4. Methods

Let the indicator $\delta_{i,j}^{svy}$ be 1 if vehicle i has been on the road on day j of its survey period according to the survey response, and 0 otherwise. Let $\delta_{i,j}^{wim}$ be an indicator equal to 1 if vehicle i is recorded by a sensor station on day j and equal to 0 otherwise. $\Theta_{i,j}$ is defined as the shipment weight carried by truck i on day j. If $\delta_{i,j}^{svy} = 1$ the sum of reported shipment weights in the survey is used, otherwise if $\delta_{i,j}^{wim} = 1$ the sensor shipment measurements are used. If a vehicle was recorded by the sensors multiple times a day, the maximum of the weights measured at these occasions is taken. Two target variables are considered: the total number of truck days (D) and the total transported shipment weight (W). One truck day is defined as a day that a truck has been on the road in the Netherlands. The regular survey statistics are post-stratification estimates, with the weights computed to take the survey design into account and to correct for selective nonresponse. The total of D and W are estimated by $\widehat{D}^{SURV} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy}\right)$ and $\widehat{W}^{SURV} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \Theta_{i,j}\right)$. The sensor observations are simply added to the survey observations resulting and estimated by $\widehat{D}^{SURVX} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \vee \delta_{i,j}^{wim}\right)$ and $\widehat{W}^{SURVX} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7}\left(\delta_{i,j}^{svy} \vee \delta_{i,j}^{wim}\right) \Theta_{i,j}\right)$. This is a basic way to include the sensor data and to provide a lower bound on the CRC estimators. Linking survey and sensor data results in three subsets of units: Elements in the survey only, in the sensor data only or in both datasets (Table 4.2-1). The empty cell represents the trucks and trips respectively, which were not reported in the survey and not recorded by a sensor. In the present study, the first capture occasion is the survey where trucks are considered as being captured and marked on specific days in the survey period ($\sum_{i,j} \delta_{i,j}^{svy}$). The second capture occasion is the sensor where $\left(\sum_{i,j} \delta_{i,j}^{wim}\right)$ are captured in total, of which $\left(\sum_{i,j} \delta_{i,j}^{svy} \wedge \delta_{i,j}^{wim}\right)$ are recaptured.The Lincoln-Petersen estimator (Lincoln 1935; Petersen 1893) uses

the quantities of those subsets to estimate the population sizes (D) and (W) by $\widehat{D}^{LP} = \frac{n_1 n_2}{m_2}$ and $\widehat{W}^{LP} = \frac{\left(\sum_{i,j} \delta_{i,j}^{svy} \Theta_{i,j}\right)\left(\sum_{i,j} \delta_{i,j}^{wim} \Theta_{i,j}\right)}{\sum_{i,j}\left(\delta_{i,j}^{svy} \wedge \delta_{i,j}^{wim}\right)\Theta_{i,j}}$.

The likelihood approach proposed by Huggins (1989) and Alho (1990) models heterogeneity in capture probabilities using covariates conditioned on the captured elements. A logistic model is used to model capture probabilities for each element on each occasion. Hence, covariates are used to model the capture probabilities $\hat{P}_{ij}^{s}$ and $\hat{P}_{ij}^{w}$, which are the capture probabilities for the survey and sensor, respectively. The Horvitz-Thompson estimator (Horvitz and Thompson 1952) is used to estimate D and W by $\widehat{D}^{HUG} = \sum_{i,j} \frac{1}{\hat{\psi}_{ij}}$ and $\widehat{W}^{HUG} = \sum_{i,j} \frac{\Theta_{i,j}}{\hat{\psi}_{ij}}$, with $\hat{\psi}_{ij} = 1 - \left(1 - \hat{P}_{ij}^{s}\right)\left(1 - \hat{P}_{ij}^{w}\right)$. The estimator HUG_{int} is the intercept only model. Fienberg (1972) introduced log-linear models for population size estimation in closed populations. To model heterogeneity in the capture probabilities in the survey (A) and sensor (B), any number of available covariates can be included in the model. Given the covariate X, the two-way contingency table is expanded to a four-way contingency table $\log m_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_x^X + \lambda_{ax}^{AX} + \lambda_{bx}^{BX}$. For every level of the included covariates, a sub-population size is estimated which in sum gives the total population size. This method is used to estimate $\widehat{D}^{LL}$ and $\widehat{W}^{LL}$. Using CRC techniques, the number in the empty cell is estimated. Two target variables of the survey are estimated: the number of truck days (D) and the corresponding transported shipment weights (W).

In addition, all estimators are applied in a stratified manner. Since suspected underreporting by nonresponse and misreporting in the RFTS is the subject of this study, the RFTS (non)-respondents constitute the study population. The number of vehicles under study is N. Therefore, the indicators $\delta_{i,j}^{svy}$ and $\delta_{i,j}^{wim}$ estimate D and W are divided into S strata, with N_s sampling units in stratum s. Within each stratum $\widehat{D}_s$ and $\widehat{W}_s$ are estimated. Strata are based on covariates in the models (see section 4.1). Within each stratum the most likely amount of underreporting will be estimated.

4.1 Model selection and variance estimation

Covariates to fit the logit and log-linear models are selected by a stepwise selection procedure (based on BIC). Since the log-linear model only allows for categorical variables and to retain the full information of the covariates, the model selection is based on the logit model. In the log-linear model, the five variables with the most predictive power in the two logit models were combined. For that purpose, the continuous covariates were categorized based on their quantiles. Using $\delta_{i,j}^{svy}$ as the dependent variable in the logit model, the selected covariates 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 $\delta_{i,j}^{wim}$ as the dependent variable, the following covariates were selected: 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. 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.

Since the trucks being the sampling units and not the truck days, bootstrapping was used to account for this cluster effect in the data (there are more truck days then sampling units). Furthermore, the shipment weight is clustered in trucks and not i.i.d. Simple random sampling with replacement was used to draw bootstrap samples. One bootstrap sample for estimation purposes consists of all elements, both survey and sensor, that are available for the vehicles in the bootstrap sample. The mean of the bootstrap distribution is computed to ascertain that the bootstrap procedure is unbiased. The 0.025% and 0.975% quantiles of the bootstrap distribution are used to estimate the boundaries of the 95% confidence intervals.

4.2 Linkage of survey and sensor data

In table 4.2-1 the results of linking the survey and sensor data are shown. The left panel of the table shows 94,338 truck days being reported in the survey. The sensors recorded 43,775 truck days of which 34,131 were reported in the survey. 9,644 truck days were recorded by the sensors which were not reported in the survey. The sensors did not record 60,207 truck days which were reported in the survey. The right panel shows the transported shipment weight in kilotons (kt) on the reported truck days.

Table 4.2-1
Captures of truck days (D) and transported shipment weight (W) in the survey and sensors.

D	Survey			W	Survey		
Sensor	reported	not reported	$\sum$	Sensor	reported	not reported	$\sum$
recorded	34,131	9,644	43,775	recorded	376,83	99,13	475,96
not recorded	60,207	–	60,207	not recorded	576,88	–	576,88
$\sum$	94,338	9,644	103,982	$\sum$	953,71	99,13	1052,84

953,71 kt were reported in the survey. 475,96 kt were recorded by the sensors, of which 376,83 kt were reported in the survey. Additional 99,13 kt were recorded by the sensors which were not reported in the survey. The sensors did not record 576,88 kt which were reported in the survey.

5. Results

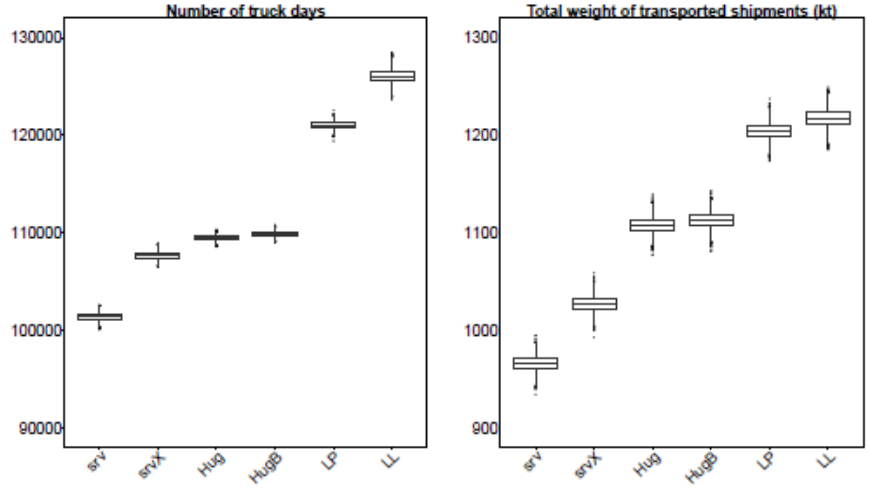
Table 5-1 shows the survey and CRC estimates for D and W. According to $SURVX$ the amount of underestimation for D and W is about 6%. The estimators HUG and HUG_{int} yield about 7% underestimation for D and about 13% for W. The estimator LP yields about 16% underestimation for D and 19% for W. In both target variables D and W the most likely amount of underestimation according to LL is about 19% for D and 20% for W.

Table 5-1
Survey and CRC estimates for D and W, bootstrapped variance, standard error and confidence interval.

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
$\hat{W}^{SURV}$	965,3	965,23	8,20	[949,33; 981,40]	–
$\hat{W}^{SURVX}$	1026,83	1026,69	8,37	[1009,94; 1043,53]	5.99
$\hat{W}^{HUG}$	1108,58	1108,36	8,32	[1091,65; 1124,37]	12.92
$\hat{W}^{HUG_{int}}$	1112,59	1112,40	8,34	[1095,52; 1128,38]	13.24
$\hat{W}^{LP}$	1204,60	1204,38	9,14	[1185,83; 1221,89]	19,87
$\hat{W}^{LL}$	1216,85	1217,40	9,74	[1197,73; 1236,08]	20.67

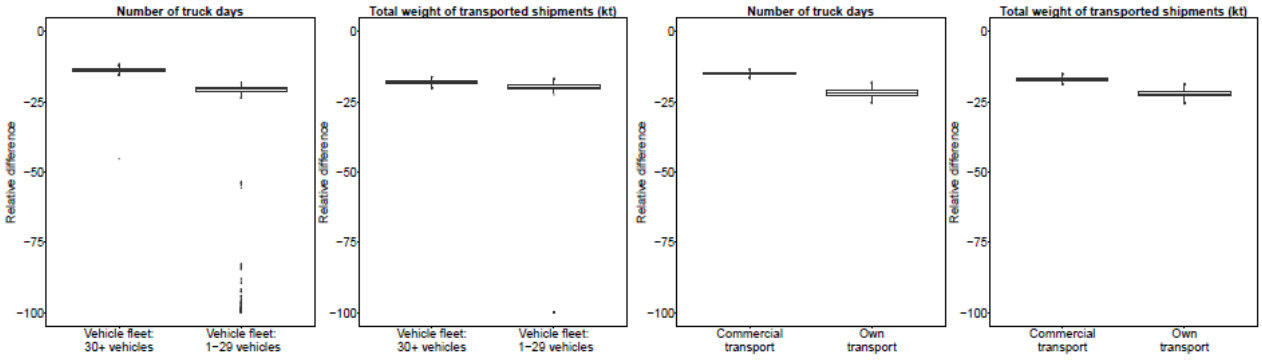
Figure 5-1 shows the six different estimators and the bootstrapped sampling variance (based on 3,000 bootstrap samples). The six different point estimates nearly match the median and are therefore not shown. In contrast to the conditional likelihood estimators, the larger difference between the unconditional likelihood estimators shows a stronger effect of modeling heterogeneity using covariates. It is recommended to rely on the estimates of LL since they are based on the full likelihood and take heterogeneity in the capture probabilities into account.

Figure 5-1
Effect of estimator on bootstrap estimates of truck days and transported shipment weights.



Therefore, for the stratified analysis of D and W the relative difference between $SURV$ and LL is shown in Figure 5-2. Again, point estimates nearly match the median and are therefore not shown. For smaller vehicle fleets (1–29 vehicles) the amount of underestimation for D is 20% and for W 19%. Larger vehicle fleets (30 + vehicles) show 13% underestimation for D and 18% for W. Commercial transport shows 15% underestimation for D and 17% for W. The most likely amount of underestimation for own transport is 22% both for D and W.

Figure 5-2
Stratification by size of vehicle fleet and type of transport, showing the effect of LL on bootstrap estimates and the relative difference between $SURV$ and LL.



6. Conclusion

We demonstrated a specific use of big data in official statistics for the estimation and adjustment of underreporting bias. Using CRC techniques, survey, sensor, and administrative micro-data were linked. The proposed combination of data sources and methods seem to produce reasonable estimates given the literature. The method presented here is applicable to any validation study where survey, administrative, and sensor data (or any other external big data source) can be linked at a micro-level using a unique identifier. However, since the sensors are not randomly distributed, the sensor data might be biased. Moreover, the OCR software does not recognize every front and/or back license plate and the resulting mismatches may influence the results. Finally, imputations methods were used to estimate missing sensor measurements. A systematic study of the effects of these problems on results is object of ongoing research.

References

Alho, J. M. (1990). Logistic regression in capture-recapture methods. *Biometrics*, 46, 623–635.

Bricka, S., & Bhat, C. (2006). Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 9–20.

Bricka, S., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies*, 21(1), 67–88.

Buelens, B., Daas, P., Burger, J., Puts, M., & van den Brakel, J. (2014). Selectivity of big data. *CBS Discussion Paper*, (2014–11).

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

Enright, B., & OBrien, E. J. (2011). Cleaning weigh-in-motion data: Techniques and recommendations. Techinal Report, Dublin Institute of Technology & University College Dublin.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59(3), 591–603.

Hassounah, M. I., Cheah, L.-S., & Steuart, G. N. (1993). Underreporting of trips in telephone interview travel surveys. *Transportation Research Record*, 1412, 90–94.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 1(76), 133–140.

Krishnamurty, P. (2008). Diary. In P. J. Lavrakas (Editor), *Encyclopedia of survey research methods* (Volume 1, Pages 197–199). Thousand Oaks: Sage.

Lincoln, F. C. (1935). *The waterfowl flyways of north america*. Washington: United States Department of Agriculture.

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293–312.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4), 199–226.

Petersen, C. G. J. (1893). On the biology of our flat-fishes. Kjøbenhaven: The Danish Biological Station.

Shen, L., & Stopher, P. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 34(3), 316–334.

Shlomo, N., & Goldstein, H. (2015). Editorial: Big data in social research. *Journal of the Royal Statistical Society, Series A*, 178(4), 787–790.

Singer, E. (2016). Reflections on surveys' past and future. *Journal of Survey Statistics and Methodology*, 4(4), 463–475.