

# Implementing Big Data in Official Statistics: Capture-recapture Techniques to Adjust for Underreporting in Transport Surveys Using Sensor Data

## **Keywords:**

*admin data, record linkage, log-linear models, logit-models, Lincoln-Petersen*

## **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 [1]. Up to now, big data is rarely used in statistical production due to its unknown data generating process [2]. However, in the long-term, using big data in official statistics is unavoidable. 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. More specifically, the different problems of surveys and big data might be minimized if the survey and the sensors measure the same target variable and the resulting micro data can be combined 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 to validate, estimate and adjust a bias due to underreporting in the target variables of the survey.

## **2 RESEARCH BACKGROUND**

The number of surveys conducted has increased over the last decades [3], while the nonresponse rates are increasing [4]. In particular, diary surveys imposes heavy response burden and yield very low response rates [5]. 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 [6]. 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 [7]. Instead of using mobile GPS devices we use permanently installed road sensors to validate and adjust survey estimates. Hereby, the problems caused by respondents impact on sensors are avoided.

## **3 DATA**

The Road freight transport survey of the Netherlands (2015) consists of approximately 35.000 trucks sampled from the national vehicle register. A central objective of the mandatory diary survey is to collect data on the shipments weights transported by the trucks. Therefore, truck owners must report the days on which the truck was used and the corresponding shipment weight.

The sensor data used is collected by the weigh-in motion road sensor network operated by the national road administration of the Netherlands. 18 stations on Dutch highways record trucks while they pass the station. While passing, the vehicle's weight is

measured. In 2015 approximately 36 million trucks were registered by the 18 stations. In addition to weight, a timestamp and a photograph of the front/rear license plates of the truck are also recorded. Using the combination of license plate and day, the trucks in the survey and in the sensor data can be linked one-by-one. Data from the national vehicle and enterprise data is linked to the sensor data. This results in a dataset consisting of survey responses, sensor data, technical truck specifications and enterprise characteristics from the registers. 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.

## 4 METHODS

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 1).

	Survey response	
	Sensor detections	reported      not reported
recorded	$\text{Sensor} \cap \text{Survey}$	Sensor only
not recorded	Survey only	–

Table 1: Subsets of linked survey and sensor datasets.

The empty cell contains zero elements, which were either not reported in the survey or recorded at a sensor station. Using capture-recapture techniques the number in the empty cell is estimated. We estimate two target variables of the survey: the number of truck days ( $D$ ) and the corresponding transported shipment weights ( $W$ ). One truck day is defined as a day that a truck has been on the road in the Netherlands. Six different estimators are applied: two survey estimators ( $SURV$ ,  $SURVX$ ), two conditional likelihood estimators ( $HUG$ ,  $HUG_{int}$ ), and two unconditional likelihood estimators ( $LP$ ,  $LL$ ). The estimator  $SURV$  is the post-stratified survey estimator and the estimator  $SURVX$  is a naive extended survey estimator. The conditional likelihood estimators are conditioned on the captured elements. Here, heterogeneity in capture probabilities is modelled using covariates and a logistic regression. The unconditional likelihood estimator  $LP$  assumes homogeneous capture probabilities within the datasets but different in between the datasets. The estimator  $LL$  assumes independent capture probabilities in the survey and sensor data and uses covariates to model heterogeneity. The most likely amount is estimated for the survey and furthermore in specific subgroups by stratification on administrative variables.

## 5 RESULTS

Figure 1 shows the six different estimates by estimator and the sampling variance for all estimators estimated by bootstrapping (3000 bootstrap samples). According to  $SURVX$  the amount of underestimation for  $D$  and  $W$  is about 6%. The unconditional likelihood estimators yield about 10% underestimation in the survey for  $D$  and about 16% for  $W$ . The small difference between  $HUG$  and  $HUG_{int}$  show little effects of the covariates. According to  $LP$  the amount of underestimation for  $D$  is 19% and for  $W$

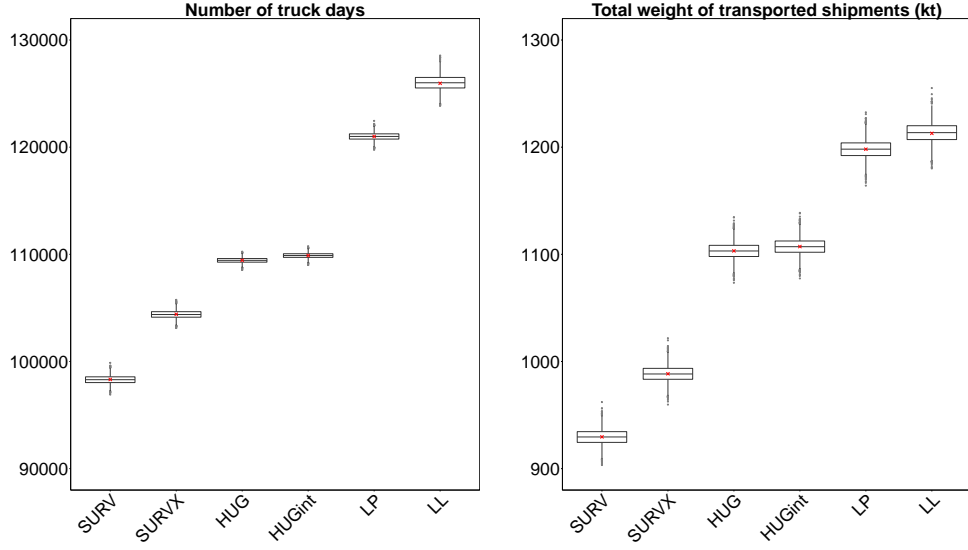


Figure 1: Effect of estimator on bootstrap estimates of truck days (left panel) and transported shipment weights (right panel).

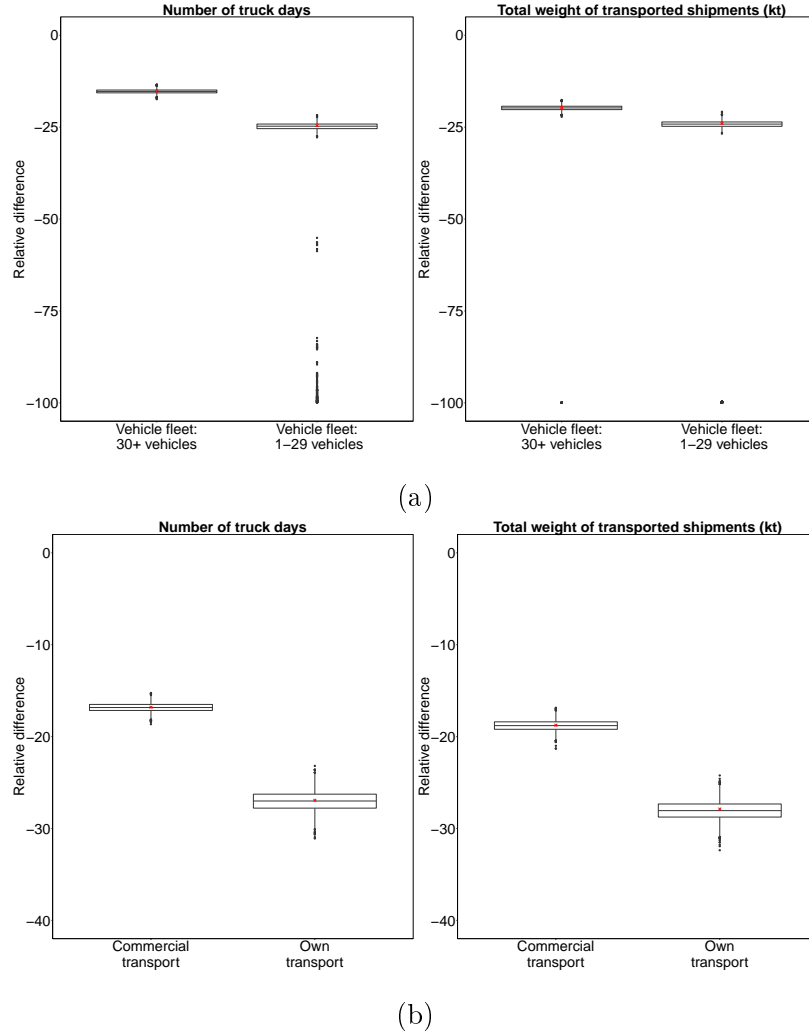


Figure 2: Relative difference between  $SURV$  and  $LL$  for size of vehicle fleet (a) and type of transport (b).

22%. In both target variables  $D$  and  $W$  the most likely amount of underestimation according to  $LL$  is about 22% for  $D$  and 23% for  $W$ .

The large difference between the unconditional likelihood estimators shows the effect

of modelling heterogeneity using covariates. We recommend relying on the estimates of  $LL$  since they are based on the full likelihood and take heterogeneity in the capture probabilities into account. Therefore, we show in Figures 2a and 2b the relative difference between  $SURV$  and  $LL$  based on the stratified analysis. For smaller vehicle fleets and non-commercial transport even higher amounts of underestimation, up to 28% are found for  $D$  and 25% for  $W$  respectively.

The proposed combination of data sources and methods seem to produce reasonable estimates given the literature on underestimation bias in transportation surveys. However, since the sensors are not randomly distributed, the road sensor data might also be biased. Moreover, the OCR software does not recognise every single license plate on the front and/or back of the vehicles, 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 the object of ongoing research.

## 6 CONCLUSION

We demonstrated a specific use of big data in official statistics for the estimation of underreporting bias. We combined survey, sensor, and administrative micro data using capture-recapture techniques. 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 on a micro-level using a unique identifier.

## REFERENCES

- [1] Piet J. H. Daas, Marco J. Puts, Bart Buelens, and Paul A. M. van den Hurk. Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262, 2015.
- [2] Bart Buelens, Piet Daas, Joep Burger, Marco Puts, and Jan van den Brakel. Selectivity of big data. *CBS Discussion Paper*, (2014–11), 2014.
- [3] Eleanor Singer. Reflections on surveys’ past and future. *Journal of Survey Statistics and Methodology*, 4(4):463–475, 2016.
- [4] Bruce D. Meyer, Wallace K. Mok, and James X. Sullivan. Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226, 2015.
- [5] Parvati Krishnamurty. Diary. In Paul J. Lavrakas, editor, *Encyclopedia of Survey Research Methods*, volume 1, pages 197–199. Sage, Thousand Oaks, 2008.
- [6] Stacey Bricka and Chandra Bhat. Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Record: Journal of the Transportation Research Board*, 1972:9–20, 2006.
- [7] Stacey Bricka, Sudeshna Sen, Rajesh Paleti, and Chandra R. Bhat. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies*, 21(1):67–88, 2012.