Capture-recapture Techniques for Transport Survey Estimate Adjustment Using Road Sensor Data

Jonas Klingwort[1] Bart Buelens[2] Rainer Schnell[1]

[1]University of Duisburg-Essen

[2]Statistics Netherlands

BigSurv 2018
Barcelona, 25–27 October 2018
26.10.2018

UNIVERSITÄT DUISBURG ESSEN

Statistics Netherlands

Introduction

- Non-probability based sensor data sources are becoming increasingly popular in social science research and official statistics.
- Maximum information gain: linking survey, sensor and administrative data (Shlomo/Goldstein 2015; Japec et al. 2015).
- Especially, when a survey and a sensor independently measure an identical target variable.
- Sensor data is most often not collected for research purposes (Connelly et al. 2016).
- Nevertheless, sensor data information could be used for research purposes.

Research background

- Unnecessary response burden if the information of interest is accessible from other datasets (Miller 2017; Schnell 2015).
- Especially time-based diary surveys impose a heavy burden.
- Such surveys yield low response rates (Krishnamurty 2008) and might be biased downwards due to “inaccurate reporting, nonreporting, and nonresponse” (Richardson et al. 1996).
- Up to 81% of underreporting in validation studies documented by Bricka/Bhat (2006).
- We use permanently installed road sensors to estimate and adjust bias due to underreporting in transport survey estimates.

Data – Survey

- Road Freight Transport Survey of the Netherlands 2015 ($n_{svy} = 34,828$ vehicles).
- Mandatory time-based diary survey with response rate about 90%.
- Each vehicle is in the survey for one week. Respondents must report all trips and shipments on each day.
- It is expected to find cases of underreporting due to nonresponse and misreporting by falsely responding that the truck was not used.

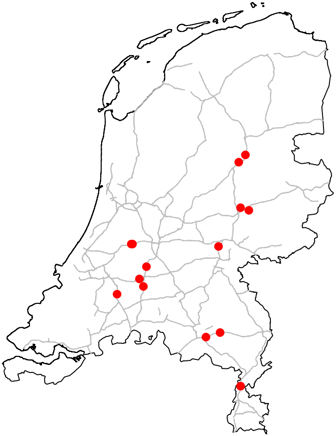
Response categories	n	%
truck used	23,461	67.4
truck not used	5,304	15.2
nonresponse	3,601	10.3
truck not owned	2,462	7.1
$\sum$	34,828	100%

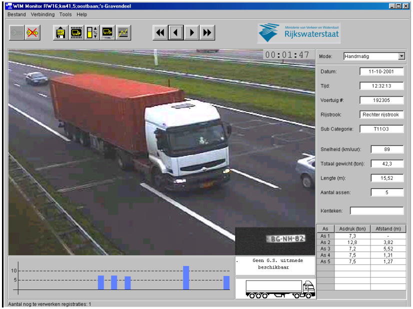
Table: Survey response categories

Data – Sensor

- Weigh-in motion road sensor data of 2015 ($n_{wim} = 35,669,347$).
- Dynamic measurement of the weight for each passing truck.
- Measurements: photograph of front/rear license plate, total weight, axles pressure, and truck classification.
- Weight of entire unit (truck, trailer, and shipment) measured.
- Result of subtracting truck and trailer weights from entire unit corresponds to the transported weight, which is equal to the definition of reported weight in the survey.

Road sensor network





Data – Administrative Data

1. The Dutch vehicle register provides information on technical truck characteristics.
2. The Dutch enterprise register provides information on characteristics of the truck owners.

Linking the datasets:

- Survey and Sensor: Linking by combination of license plate and day as unique identifier.
- Matched data set: Linking by combination of license plate and quarter as unique identifier.

Capture-Recapture Method for this Setting

- Capture-recapture methods are used to to estimate and adjust underreporting in the survey.
- Survey (A) and sensor (B) observations are considered as a two occasion capture setup.
- Three quantities are derived: $A \setminus B$, $B \setminus A$, and $A \cap B$.
- $A \setminus B$ is the first capture occasion (survey-only), $B \setminus A$ is the second capture occasion (sensor-only), and $A \cap B$ are the elements captured twice.

	Survey response	
Sensor detections	reported	not reported
recorded	A ∩ B	B
not recorded	A	–

Table: Quantities of linked survey and sensor datasets.

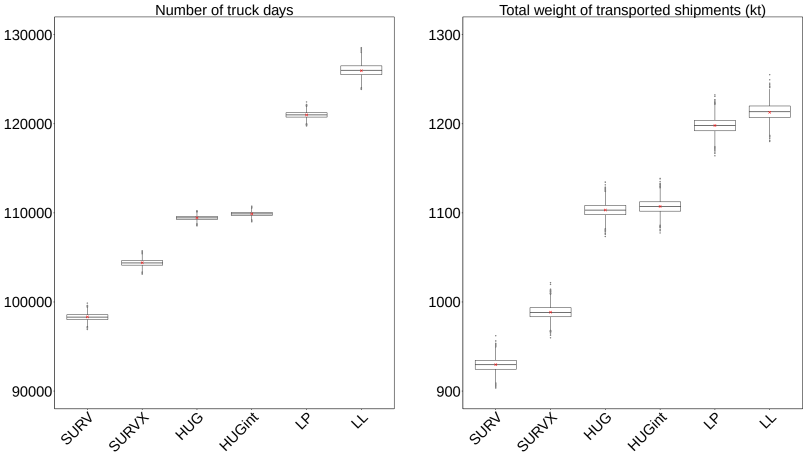
Definitions and Assumptions

- Heterogeneity of the vehicles with respect to capture and recapture probabilities is modeled through logistic regression and log-linear models.
- Assumptions: independent data sets, closed population, elements belong to population, perfect linkage, homogeneous capture probabilities.
- Six estimators for truck days (D) and transported shipment weights (W) are applied, compared, and discussed.
- One truck day is defined as a day that a truck has been on the road in the Netherlands.

Estimators

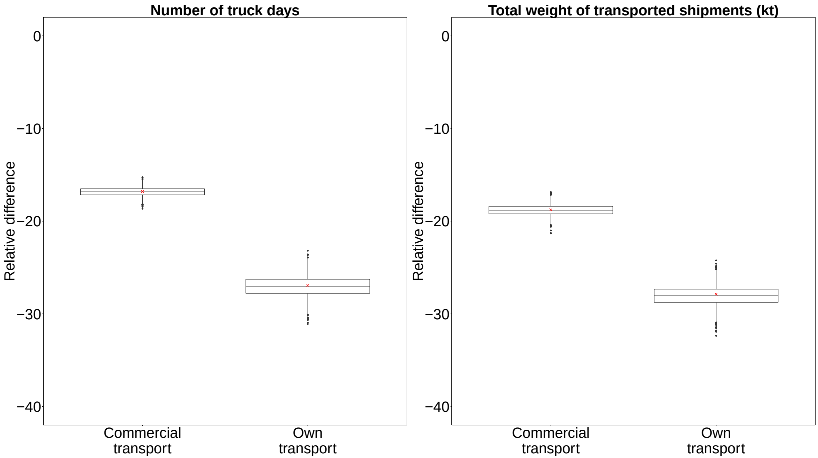
- Survey Estimators:
  - *SURV*: Post-stratified survey estimator
  - *SURVX*: Naive extended survey estimator
- Conditional likelihood estimators
  - *HUG*: Conditioned on the captured elements; heterogeneity in capture probabilities modelled using covariates; logistic regression
  - HUG_{int}: intercept model
- Full likelihood estimators:
  - *LP*: Homogeneous capture probabilities in A and B; uses $A \setminus B$, $B \setminus A$, and $A \cap B$
  - *LL*: Assumes independent capture probabilities in A and B; Covariates used to model heterogeneity
- Stepwise selection procedure (based on BIC) to chose covariates to fit the logit and log-linear models.
- Bootstrap variance estimates for all estimators were computed.

Results

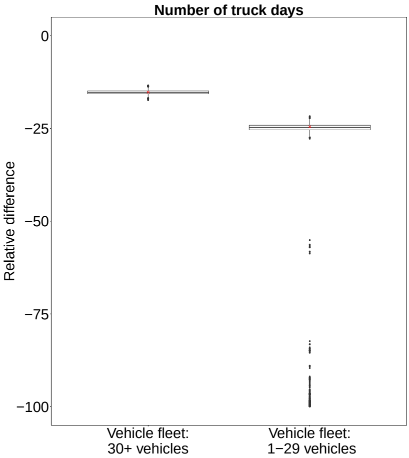


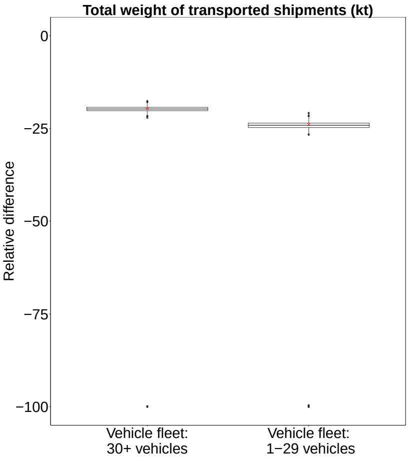
Number of truck days
130000
120000
110000
100000
90000
SURV
SURVX
HUG
HUGint
LP
LL
Total weight of transported shipments (kt)
1300
1200
1100
1000
900
SURV
SURVX
HUG
HUGint
LP
LL

Results – Type of transport



Results – Size of vehicle fleet





Summary

- All estimators yield larger estimates for truck days and transported shipment weights than the survey.
- Recommendation to rely on the log-linear model (based on the full likelihood, takes heterogeneity into account).
- Most likely amount of underestimation in the survey up to 22% for truck days and 23% for the transported shipment weight.
- In comparison to results in the literature, we observed a moderate underestimation in the survey.
- Stratification showed larger amounts of underestimation in the survey for
    - own transport ($\hat{D} = 27\%$, $\hat{W} = 28\%$)
    - and smaller vehicle fleets ($\hat{D} = 25\%$, $\hat{W} = 24\%$).

Conclusion

- We demonstrated a method to use big data in official statistics to estimate bias in survey estimates by combining survey, administrative, and sensor data using capture-recapture.
- With this technique, we quantified the survey underestimation and adjusted the survey estimate.
- The capture-recapture technique for survey adjustment introduced here can be applied whenever survey, administrative, and sensor data (or any other external big data source) can be linked on a micro-level using a unique identifier.

References

Bricka, Stacey/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: 9–20.

Connelly, Roxanne/Christopher J. Playford/Vernon Gayle/Chris Dibben (2016): The Role of Administrative Data in the Big Data Revolution in Social Science Research. In: *Social Science Research* 59 (Supplement C): 1–12.

Japec, Lilli/Frauke Kreuter/Marcus Berg/Paul Biemer/Paul Decker/Cliff Lampe/Julia Lane/Cathy O'Neil/Abe Usher (2015): Big Data in Survey Research: AAPOR Task Force Report. In: *Public Opinion Quarterly* 79 (4): 839–880.

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

References

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

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

Schnell, Rainer (2015): "Combining Surveys with Non-questionnaire Data: Overview and Introduction". In: *Improving Survey Methods: Lessons Learned from Recent Research*. Ed. by Uwe Engel/Ben Jann/Peter Lynn/Annette Scherpenzel/Patrick Sturgis. New York: Routledge: 269–272.

Shlomo, Natalie/Harvey Goldstein (2015): Editorial: Big Data in Social Research. In: *Journal of the Royal Statistical Society, Series A* 178 (4): 787–790.