# Correcting Survey Measurement Error Using Road Sensor Data

Jonas Klingwort[1,2] Bart Buelens[3] Joep Burger[2] Rainer Schnell[1]

[1]University of Duisburg-Essen

[2]Statistics Netherlands

[3]Flemish Institute for Technological Research

International Total Survey Error Workshop 2019
Bergamo, 10–12 June 2019
12.06.2019

Statistics Netherlands

UNIVERSITÄT DUISBURG ESSEN

# Introduction

- Non-probability based sensor data is becoming increasingly popular in official statistics.
- Rarely used due to their unknown data generating processes.
- Moreover, sensor data is often not collected for research purposes.
- Maximum information gain: linking survey, sensor and administrative data.
- Especially, when survey and sensor independently measure an identical target variable.
- Available data should be used in the production of official statistics.

# Research background

- Need to provide statistics faster/cheaper and reduce respondent burden.
- Unnecessary respondent burden if the information of interest is accessible from other datasets.
- Especially time-based diary surveys impose a heavy response burden.
- Those surveys yield low response rates and might be biased downwards due to "inaccurate reporting, nonreporting, and nonresponse" (Richardson et al. 1996).

# Research background

- Underreporting in validation studies up to 81% documented for transport, mobility, and travel surveys (Bricka/Bhat 2006).
- Majority of validation studies used mobile GPS devices (attached on vehicle or respondent).
- 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 ($n_{svy} = 33,817$)
- Mandatory time-based diary survey with high response rate of about 90%.
- Each vehicle is in the survey for one week. Respondents must report all trips and shipments on each day.
- We expect underreporting due to nonresponse and misreporting by falsely responding that the truck was not used.
- Category 'Truck not owned' excluded, because the validity of response cannot be verified.

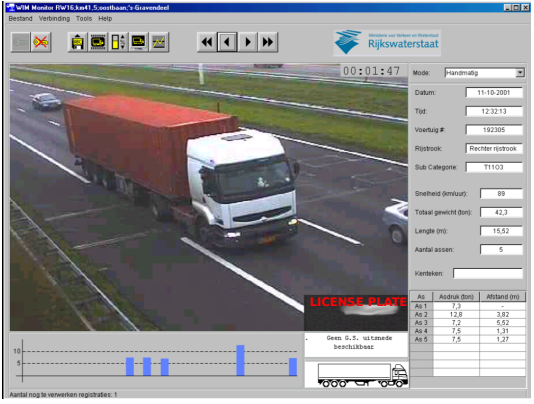
# Data – Survey response categories

| Response categories | n | % |
|---|---|---|
| truck used | 22,454 | 67 |
| truck not used | 5,304 | 15 |
| nonresponse | 3,597 | 10 |
| truck not owned | 2,462 | 7 |
| $\sum$ | 33,817 | 100 |

# Data – Sensor

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

# Road sensor network



# Method

- Capture-recapture techniques are used to estimate and adjust bias due to underreporting in transport survey estimates.
- Initially developed in ecology and biology to estimate (unknown) population sizes.
- Transferred to human populations: social and medical research, census undercount estimation, estimate unknown population sizes or prevalence of a disease.
- For this purpose, record-linkage on a micro-level using a unique identifier (license plate + timestamp) in survey, sensor, and administrative data is applied.
- Our contribution is using road sensor data and capture-recapture to correct survey point estimates.

# Method

- At least two lists/datasets need to be available (or one with duplicated entries).
- Survey and sensor observations are considered as a two occasion setup.
- Survey and Sensor data available, with overlap $Sensor \cap Survey$.

| | Survey | |
|---|---|---|
| Sensor | reported | not reported |
| recorded | Sensor ∩ Survey | Sensor only |
| not recorded | Survey only | – |

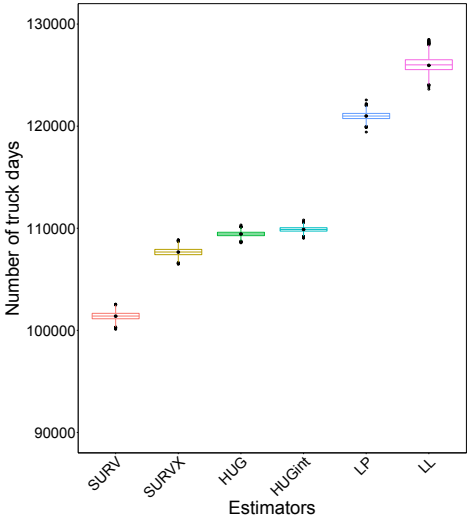
# Method

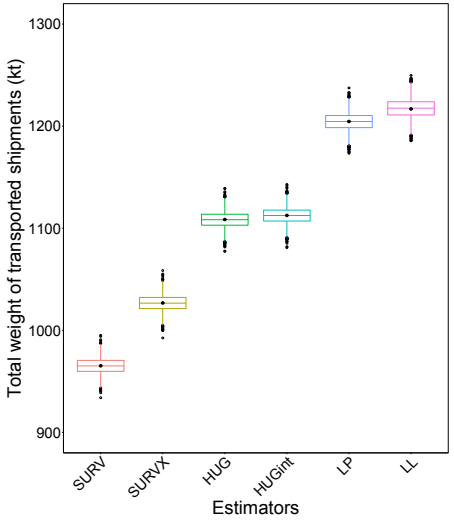
- Assumptions: independent data sets, closed population, elements belong to population, perfect linkage, homogeneous capture probabilities.
- Heterogeneity of the vehicles with respect to capture and recapture probabilities is modelled through logistic regression and log-linear models.
- Therefore, it should be irrelevant where and how many sensors are installed, given the assumptions hold.
- Six estimators for the two target variables 1. truck days ($D$) and 2. transported shipment weight ($W$).
- A truck day ($D$) is defined as a day 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 survey and sensor data, which can be different
  - $LL$: Assumes independent capture probabilities in the survey and sensor data; covariates used to model heterogeneity
- Stepwise selection procedure (based on BIC) to chose covariates to fit the logit and log-linear models.
- Sampling variance for all estimators is estimated by bootstrapping.

# Results


Number of truck days
130000
120000
110000
100000
90000
SURV
SURVX
HUG
HUGint
LP
LL
Estimators


Total weight of transported shipments (kt)
1300
1200
1100
1000
900
SURV
SURVX
HUG
HUGint
LP
LL
Estimators

# Results - Convergence of *LL* estimator

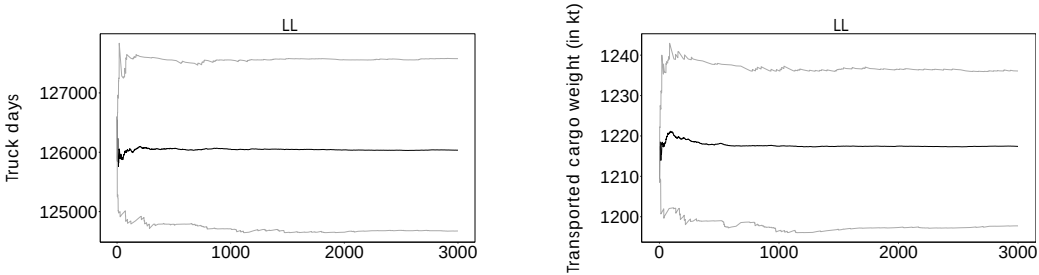
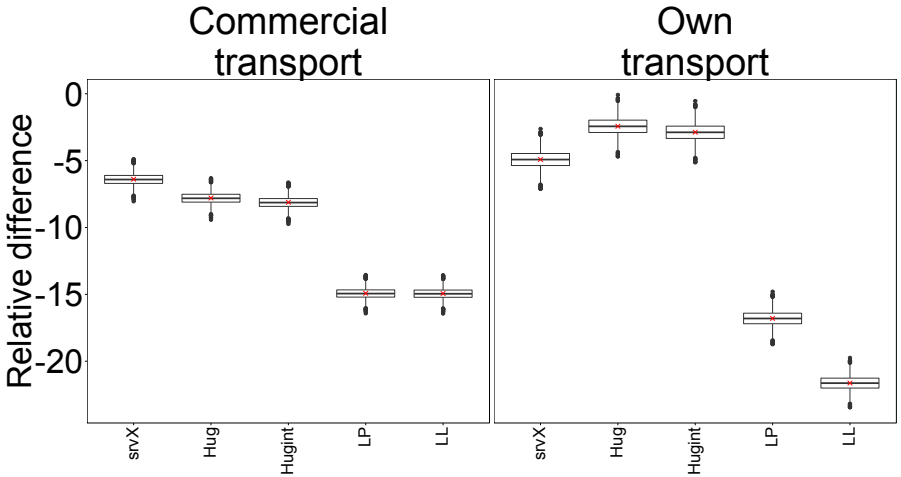



Figure: Convergence of the 3,000 bootstrap iterations for the number of truck days and the transported shipment weight

# Stratification

- All estimators are applied within strata.
- $D$ 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.
- Stratification revealed nearly the same amount of underreporting in $D$ and $W$.
- Therefore, only results for $D$ will be shown.

# Stratification: Type of transport ($D$)



# Simulation Study: Setup

- If vehicle $i$ has been on the road on day $j$ of its survey period the indicator $\delta_{i,j}^{svy}$ takes the value 1, and 0 otherwise. Therefore,

$$0 \leq \sum_{j=1}^{7} \delta^{svy}\{i,j\} \leq 7. \tag{1}$$

- Writing (1) as response pattern yields to
  1. 1111111 for the maximum of reported trips
  2. 0000000 for the minimum of reported trips

The following rules were used to simulate response errors:

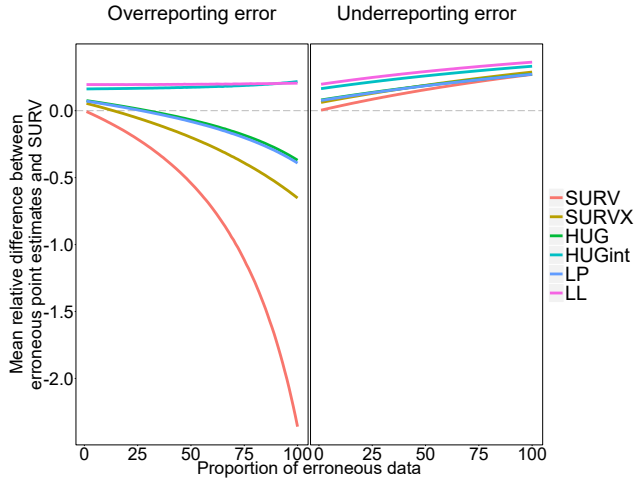
1. Overreporting error: If a 1 occurs sequentially $\geq 2$ in a response pattern, replace every 1 with 0. Keep the 1 which started the sequence.
2. Underreporting error: Starting from the first 1 in the pattern, all following 0 in the pattern are replaced by 1.

# Simulation Study: Setup

- Rules are simple but allow the simulation of a large number of errors by preserving subsets of original response patterns.
- Errors are only simulated for $D$, since error simulations for $W$ are difficult to implement.
- In the original survey data, 133 different response patterns occur, with simulated underreporting the number of response patterns is reduced to 38, and for the simulated overreporting only 12 different response patterns remain.

| Overreporting | | Underreporting | |
|---|---|---|---|
| Original | Simulated truth | Original | Simulated truth |
| 1111111 | 1000000 | 1000000 | 1111111 |
| 1000111 | 1000100 | 1000111 | 1111111 |
| 1101101 | 1001001 | 1101101 | 1111111 |
| 0011011 | 0010010 | 0011011 | 0011111 |

# Simulation Study: Results



# Summary

- All estimators yield larger estimates for truck days and transported shipment weight than the survey.
- We suggest using the log-linear estimator (based on the full likelihood, takes heterogeneity into account).
- The most likely amount of underestimation in the survey is 20% for truck days and 21% for the transported shipment weight.
- The overall trend in underestimation per strata is comparable to non-stratified analysis.
- The sensitivity assessment of the big data adjusted survey estimates towards response errors shows, that the recommended estimator $LL$ is robust against overreporting errors and sensitive to underreporting errors.

# Limitations

- OCR software does not recognise every license plate on the front and/or back of the vehicles (potential selection bias).
- Cameras not installed on all lanes, but reasonable to neglect fast lanes.
- Sensors record only one point of time from the entire journey, which might explain differences in the transported and reported shipment weight.

# Conclusion

- We demonstrated a method to use big data in official statistics to estimate bias in survey point estimates by combining survey, administrative, and sensor data with capture-recapture techniques.
- Such method requires a unique identifier or a unique combination of identifying attributes of micro units.
- 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.

# Questions / Discussion

- What critique, disadvantages or problems related to the demonstrated method do you discover?
- Do you know other or comparable application areas?
- Do you have ideas for comparable types of data to replicate analysis?

# 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.

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

Citro, Constance F. (2014): From Multiple Modes for Surveys to Multiple Data Sources for Estimates. In: *Survey Methodology* 40 (2): 137–161.

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.

Daas, Piet J. H./Marco J. Puts/Bart Buelens/Paul A. M. van den Hurk (2015): Big Data as a Source for Official Statistics. In: *Journal of Official Statistics* 31 (2): 249–262.

# References


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.

Lohr, Sharon L./Trivellore E. Raghunathan (2017): Combining Survey Data with Other Data Sources. In: *Statistical Science* 32 (2): 293–312.

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.

# References

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.