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

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

[1]University of Duisburg-Essen, [2]Statistics Netherlands, [3]VITO

## Project Goal

- Demonstrate a specific use of big data in official statistics for the estimation and adjustment of underreporting bias in survey point estimates.
- Assess the sensitivity of big data adjusted survey point estimates to response errors using a simulation study.

## Introduction

- The increasing relevance to implement big data in official statistics requires applications and empirical studies.
- Maximum information gain: linking survey, sensor and administrative data (Japec et al. 2015).
- Linking different datasets is especially valuable when survey and sensor independently measure an identical target variable.

## 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, yield low response rates (Krishnamurty 2008), and might be biased downwards due to "inaccurate reporting, nonreporting, and nonresponse" (Richardson et al. 1996).
- Permanently installed road sensors are used to estimate and adjust bias due to underreporting in transport survey estimates.

## Data

- Dutch Road Freight Transport Survey of 2015 ($\sim$ 35 thousand vehicles).
- Each vehicle is in the survey for one week. Respondents must report all trips and shipments on each day.
- Weigh-in motion road sensor data of 2015 ($\sim$ 36 million observations).
- Each station continuously measures the weight of passing trucks.
- Administrative data from the vehicle register and enterprise register.
- Linking by combination of license plate and day/quarter as unique identifier.

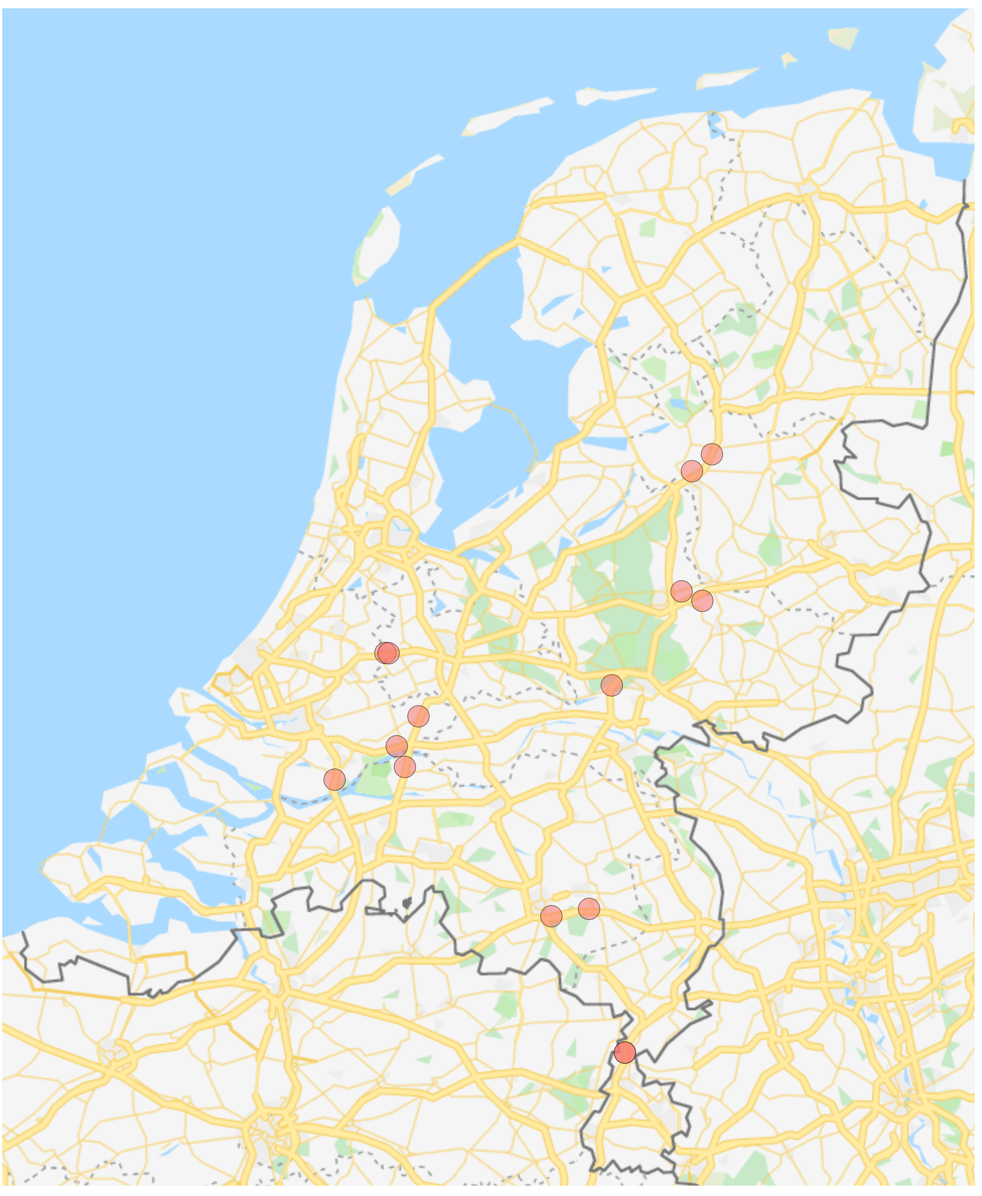

Fig. 1: Dutch Weigh-in motion road sensor network

## Methods

- Capture-recapture methods are used to estimate and adjust underreporting in the survey.
- Survey and sensor observations are considered as a two occasion capture setup.

| | Survey dataset | |
|---|---|---|
| Sensor dataset | included | not included |
| included | Sensor $\cap$ Survey | Sensor only |
| not included | Survey only | – |

- Heterogeneity of the vehicles with respect to capture and recapture probabilities is modelled 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 the two target variables truck days ($D$) and transported shipment weight ($W$) are applied, compared, and discussed.

## 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
- *HUGB*: 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; log-linear model

## Results

According to *LL*, underestimation in *SURV* is about 19%.

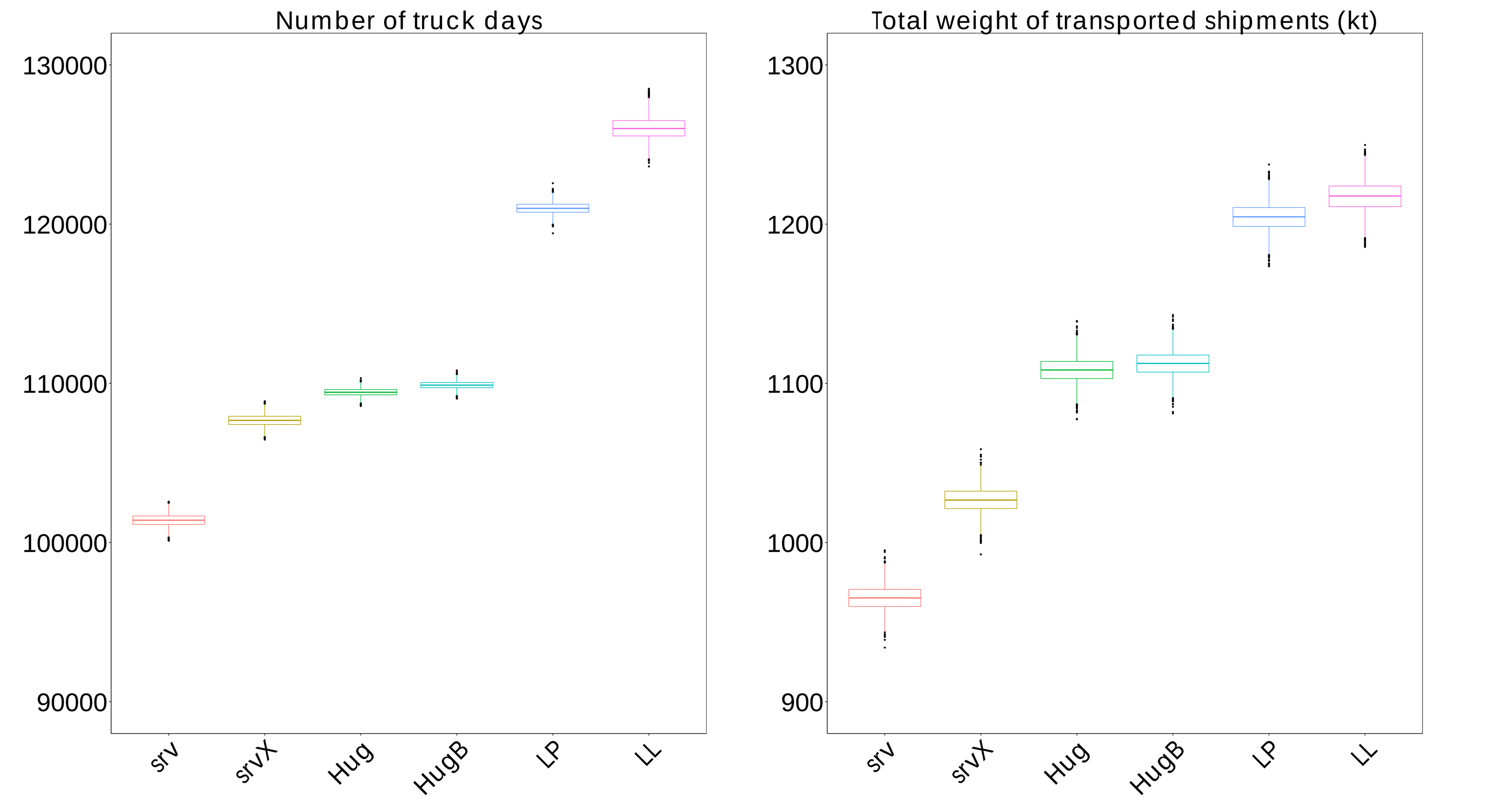


Fig. 2: Bootstrap estimates of the six estimators for truck days and transported shipment weights.

## Simulation study: Sensitivity of CRC estimates to response errors

Based on observed survey data two systematic response errors are simulated (maximum error).

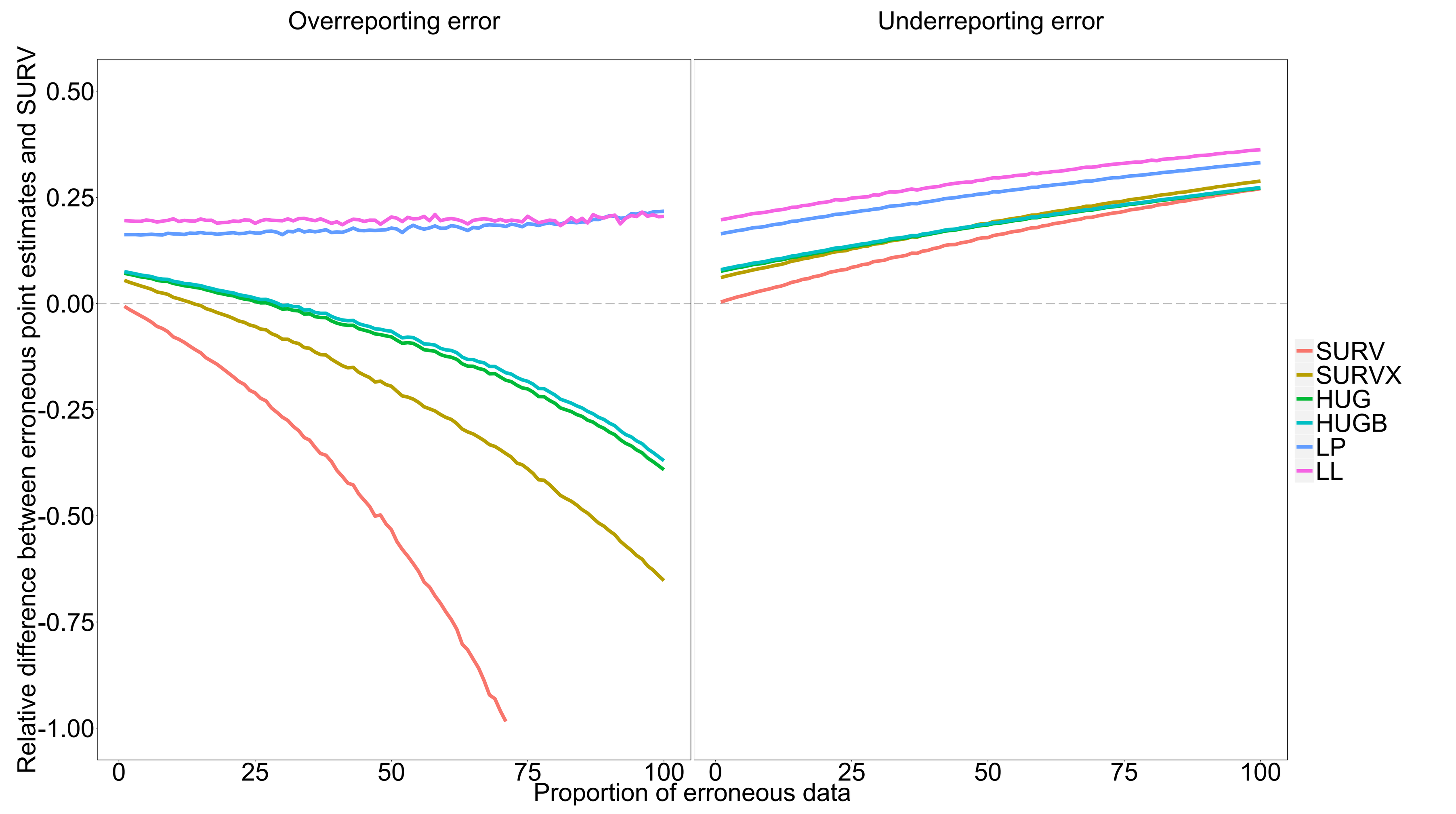


Fig. 3: Effect of response errors on point estimates for truck days and transported shipment weights.

## Conclusion

- The demonstrated method 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.
- The proposed combination of data sources and methods seem to produce reasonable estimates given the literature.
- 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.

## References

- Japec, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O'Neil & A. Usher (2015). Big Data in Survey Research: AAPOR Task Force Report. Public Opinion Quarterly 79.4, 839–880.
- Krishnamurty, P. (2008). Diary. Encyclopedia of Survey Research Methods. Ed. P. J. Lavrakas. Vol. 1. Thousand Oaks: Sage, 197–199.
- Miller, P. V. (2017). Is There a Future for Surveys? Public Opinion Quarterly 81.S1, 205–212.
- Richardson, A. J., E. S. Ampt & A. H. Meyburg (1996). Nonresponse Issues in Household Travel Surveys. Conference Proceedings 10: Household Travel Surveys–New Concepts and Research Needs. Ed. TRB National Research Council. Washington, 79–114.
- Schnell, Rainer (2015). Combining Surveys with Non-questionnaire Data: Overview and Introduction. Improving Survey Methods: Lessons Learned from Recent Research. Ed. U. Engel, B. Jann, P. Lynn, A. Scherpenzel, and P. Sturgis. New York: Routledge, 269–272.

**New Techniques and Technologies for Statistics (NTTS): Brussels, Belgium. 12.03.2019.** This presentation is granted by Förderverein Universität Duisburg Essen