Understanding the Difference in Freight Transport Estimates With and Without Road Sensor Data

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

[1]Statistics Netherlands

[2]University of Duisburg-Essen

[3]Flemish Institute for Technological Research

BigSurv20
Session ‘Official statistics in the era of big data’
13 November 2020 (virtually)

Center for Big Data Statistics

UNIVERSITÄT DUISBURG ESSEN

Introduction

- Diary surveys may suffer from underreporting due to high response burden.
- Linking sensor with survey data and applying capture-recapture (CRC) estimators provide an estimate of the amount of underreporting.
- To what extent can other factors explain the difference between survey and CRC estimates?
- Our study strengthens the trustworthiness in combining new data sources with established methodology in official statistics, which could ultimately lead to the uptake of these methods in statistical production processes.

Data

- Dutch Road Freight Transport Survey of 2015 ($\sim$ 35 thousand vehicles).
- Each vehicle is in the survey for one week. Vehicle owners must report day of loading and all trips on each day (used, not used, not owned, nonresponse).
- Weigh-in-Motion road sensor data of 2015 ($\sim$ 36 million observations).
- 18 stations continuously measure the weight of passing trucks.
- Auxiliary information from vehicle and enterprise registers.
- Linking by unique key (license plate and day).
- Target variables: total number of vehicle days (D), total transported shipment weight (W).

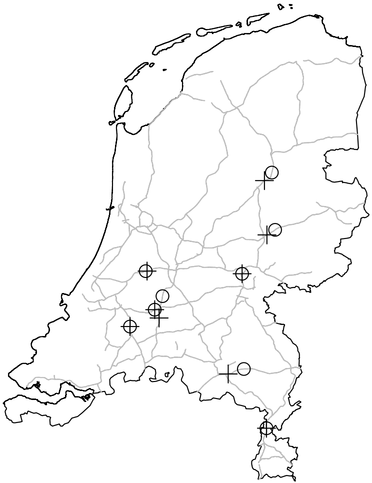


Figure: Weigh-in-Motion road sensor network

Data linkage

Table: Number of vehicle days D reported in survey and measured by sensors

D		Sensor		
		measured	not measured	$\sum$
Survey	reported	34,284	60,522	94,806
	not reported	9,727	?	?
	$\sum$	44,011	?	?

Table: Transported shipment weight W (kt) reported in survey and measured by sensors

W (kt)		Sensor		
		measured	not measured	$\sum$
Survey	reported	591	879	1470
	not reported	139	?	?
	$\sum$	730	?	?

Methods

- Log-linear CRC estimator
- BIC-based model selection
- Bootstrapped variance estimation
- Potential alternative explanations:
    - Reporting errors: simulate questionnaire-based errors
    - Measurement errors: simulate false positives and OCR failure
    - Reported not owned: treat as frame error or nonresponse error
    - Response mode: compare manual (Web/Paper) with automated (XML)

Methods – Reporting errors

- Survey respondents must fill out the day of loading instead of the day of driving (reporting error by design).
- The response pattern can be written as a string giving a range from 0000000 (not used) to 1111111 (used on every day of the week).
- Underreporting error: replace each trailing 0 with a 1. For instance, 0010000 becomes 0011111.
  - Attempts to correct for day of loading vs. day of driving, and for pooling multiple days to the first day.
- Overreporting error: replace each trailing 1 with a 0. For instance, 0111010 becomes 0100010.
  - Not main concern but a control to show the opposite effect of underreporting.

Methods – Measurement errors

- False positives: not driving for transport purposes; removing units from m_{21}.
- OCR failure: sensors do not recognize license plate (30%); m_{11} and m_{21} were decreased by the same percentage, m_{12}, was increased by the same amount that m_{11} was decreased.

False positives

D		Sensor	
		measured	not measured
Survey	reported	m_{11}	m_{12}
	not reported	m_{21} ($\downarrow$)	m_{22}

OCR failure

W (kt)		Sensor	
		measured	not measured
Survey	reported	m_{11} ($\downarrow$)	m_{12} ($\uparrow$)
	not reported	m_{21} ($\downarrow$)	m_{22}

Methods – Reported not owned

- Vehicles reported not owned were treated as frame error and excluded
- Unknown whether or not they can be measured by the sensors (i.e., scrapped or exported).
- Implies that all sample units classified as nonresponse are assumed to own the vehicle, i.e., that if the vehicle is not owned, then it is assumed that this is always reported.
- Vehicles reported not owned are treated as nonresponse error in published official statistics.
- To compare these two visions, we re-estimated the amount of underreporting when treating vehicles reported not owned as nonresponse error.

Methods – Response mode

- Test the hypothesis that the difference between survey estimates and CRC estimates is due to manual underreporting in the survey.
- Stratified both estimators by response mode: manual (internet and paper questionnaires) and automated response mode (journey planning system).

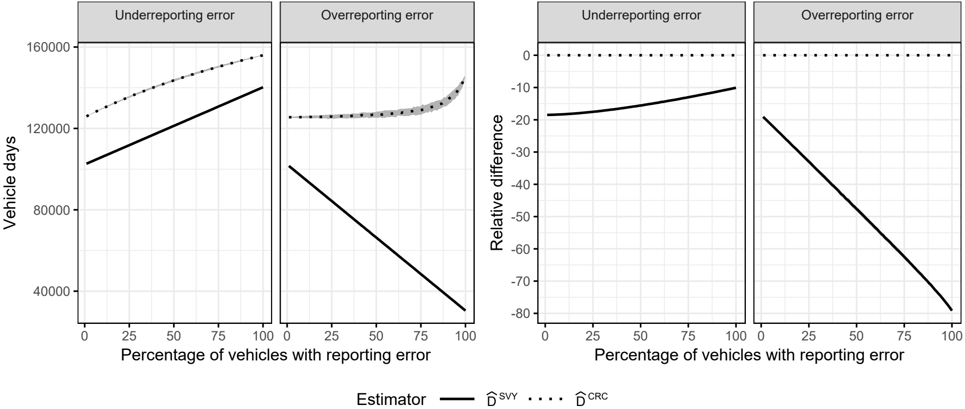
Results – Total underreporting

Table: Survey estimates, CRC estimates and the estimated underreporting (%) of the number of vehicle days D and transported shipment weight W (kt)

Estimator	Point estimate	Bootstrap mean	Bootstrap standard error	Bootstrap 95% CI	Estimated underreporting (%)	Bootstrap 95% CI
$\widehat{D}^{SVY}$	102,273	102,266	408	[101,474, 103,059]	-18.4	[-19.2, -17.6]
$\widehat{D}^{CRC}$	125,327	125,350	619	[124,125, 126,572]		
$\widehat{W}^{SVY}$	1499	1499	9.4	[1481, 1518]	-18.3	[-19.3, -17.4]
$\widehat{W}^{CRC}$	1835	1835	11	[1815, 1857]		

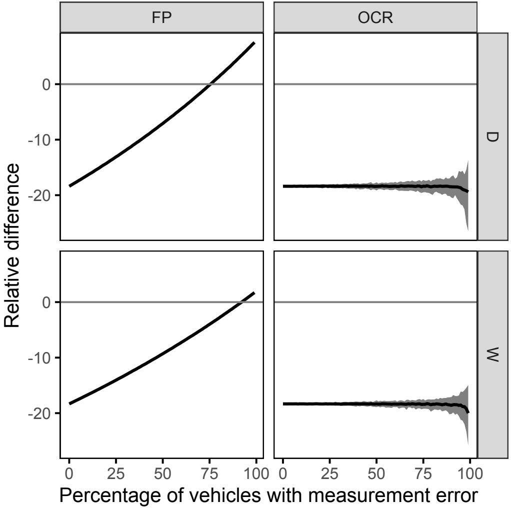
Results – Reporting errors

- Even if all vehicle owners would only report the first day of loading, the survey estimate would still be 10% lower than the CRC estimate.
- Overreporting would only increase the relative difference.



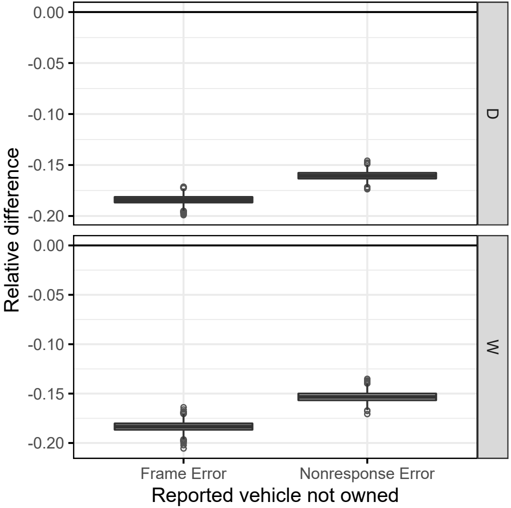
Results – Measurement errors

- The relative difference decreases with the proportion of false positives (left panels).
- The relative difference is robust against linkage errors and sensor failure (right panels).



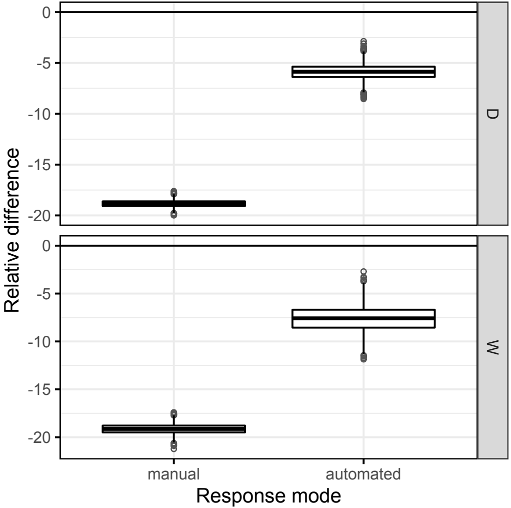
Results – Reported not owned

- Treating vehicles reported not owned as nonresponse error instead of as frame error only explained about 2% to 3%-point of the relative difference.



Results – Response mode

- The relative difference is much lower in the automated than in the manual modes.
- The difference in the automated mode, however, suggests that factors other than underreporting play a role.



Conclusion

- We conclude that sensor data, in combination with the CRC estimator, provides a valid tool to assess underreporting in survey questionnaires.
- We consider this study a useful reference for statisticians in the fields of transport research and official statistics if survey, sensor, and register data can be linked, and CRC can be applied.
- Some alternative explanations can jeopardize this methodology and need to be considered closely, in particular false positives.
- Limitations: unable to estimate the amount of false positives links and multiple explanations were not studied simultaneously but separately.

References

- Klingwort, Jonas/Bart Buelens/Rainer Schnell (2019): Capture–Recapture Techniques for Transport Survey Estimate Adjustment Using Permanently Installed Highway–Sensors. In: Social Science Computer Review special issue on 'Big Data and Survey Science'. doi:10.1177/0894439319874684.
- Klingwort, Jonas (2020): Correcting Survey Measurement Error With Big Data from Road Sensors Through Capture–recapture. PhD thesis. University of Duisburg–Essen. doi:10.17185/duepublico/72081.
- Klingwort, Jonas/Joep Burger/Bart Buelens/Rainer Schnell: Understanding the Difference in Freight Transport Estimates with and Without Road Sensor Data. In: Travel Behaviour and Society (under review).