

Effect of classification errors on accuracy of business statistics

Arnout van Delden, Sander Scholtus and
Joep Burger.

EESW 9 sept. 2015

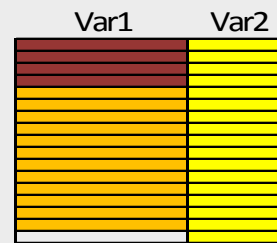


Statistics
Netherlands

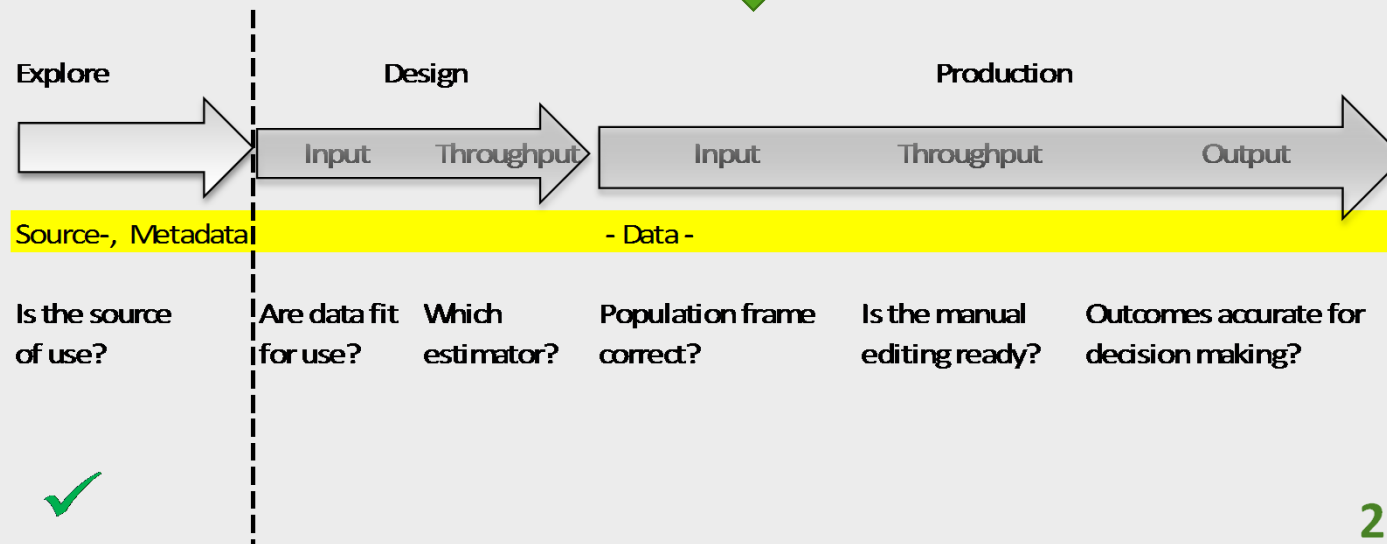
Combining sources

Combining data sources

- Sample surveys
- Administrative data
- Big data



Quality?

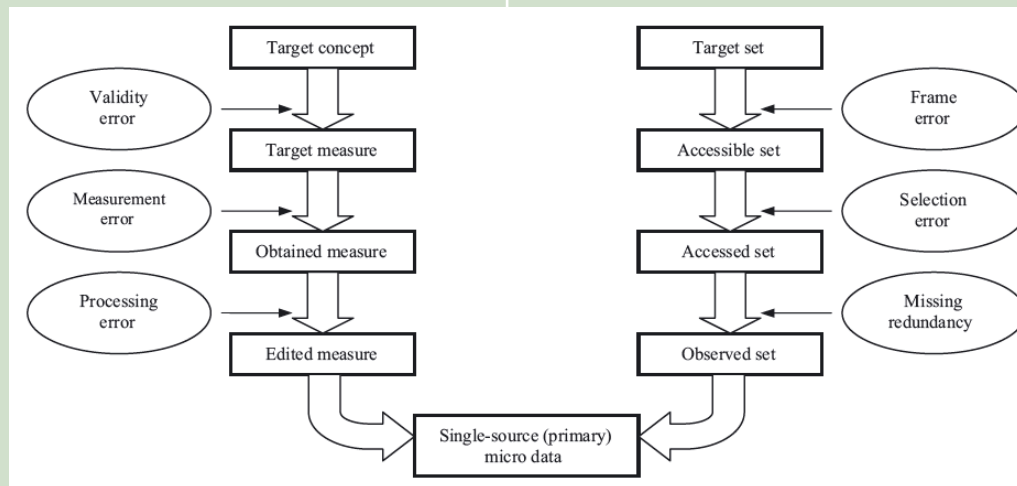


Errors?

Measurement side

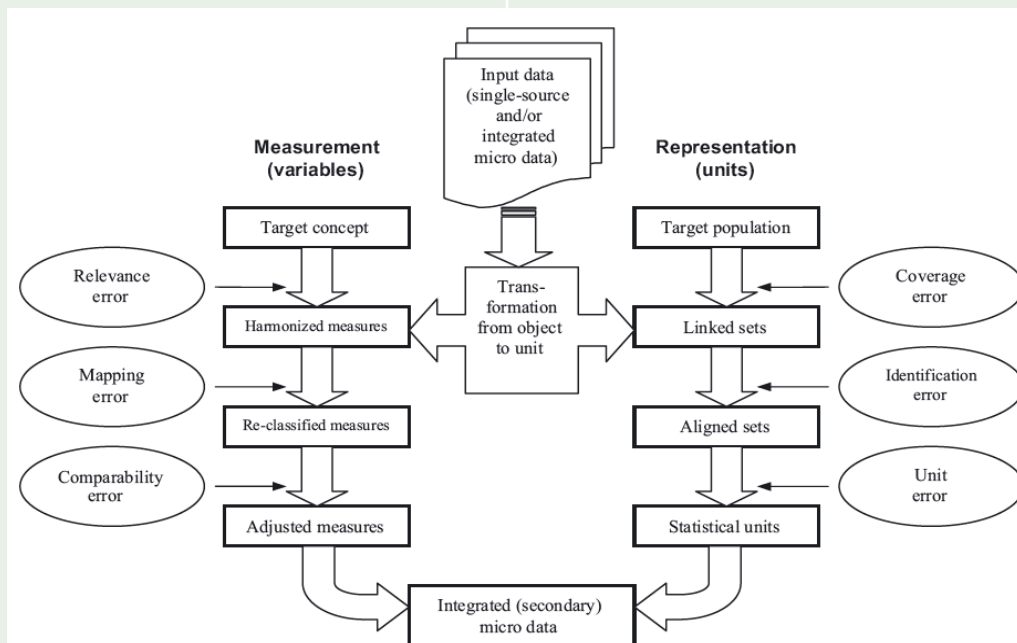
Representation side

Single source



Combined sources

Zhang (2012) & Bakker (2011)



Case study in Car Trade

Quarterly turnover estimates

- census survey (complex units)
- value added tax data (simple units)

Output

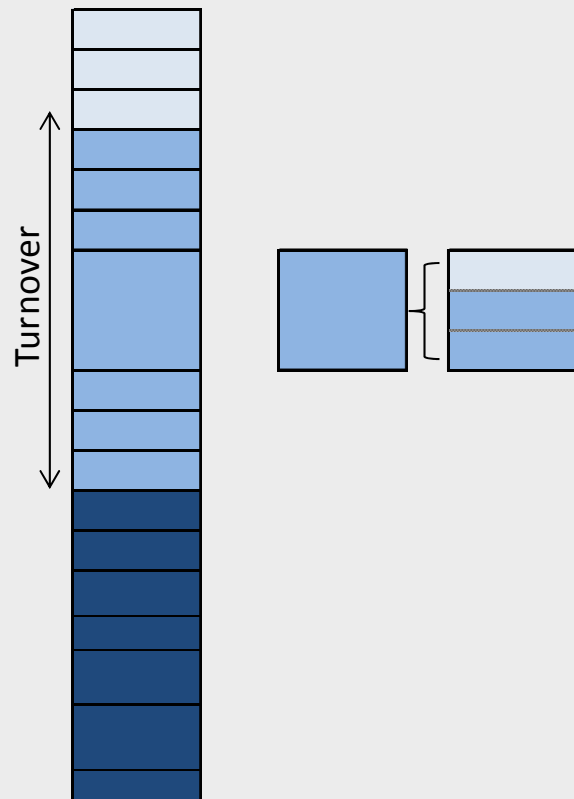
- estimates used by STS, SBS, NA
- consider only level estimates
- nine NACE groupings (base cells) to produce all output
45112 (sale and repair of passenger cars) , , 45194 (sale and repair of caravans)

Effect of classification errors

- assume: no other errors
- we target only at the correct *main activity* per statistical unit.

NACE code of statistical units

Correct



Legend

NACE A
NACE B
NACE C

SU

LU
LU
LU

Observed (in BR)

True value



Coding error CoC



Not reported change



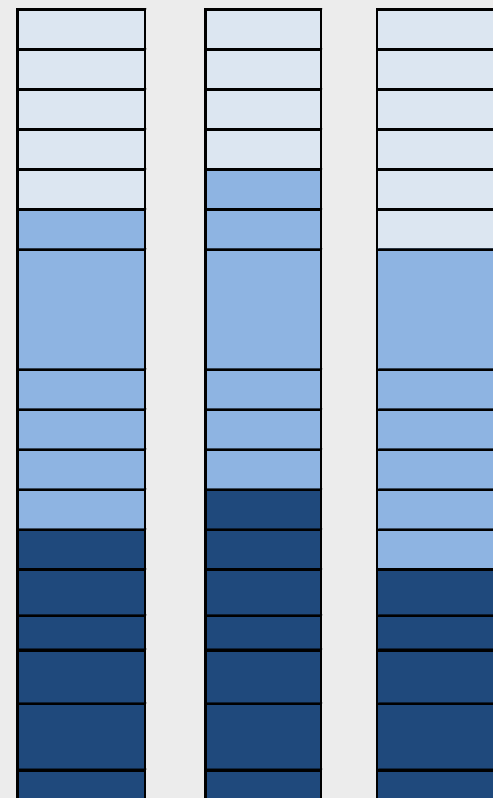
Wrong code reported



Wrong delineation



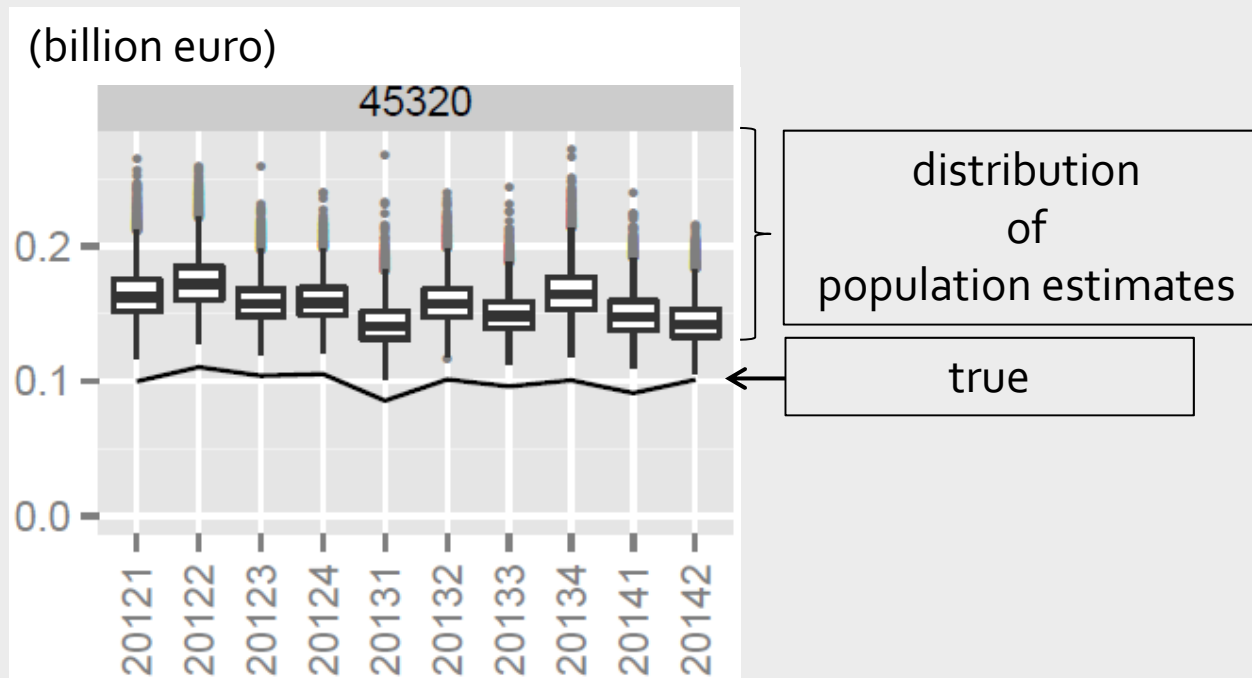
Possible outcomes



5 / 19



Effect of classification errors



$$\text{True: } \text{RMSE}_y = \sqrt{\text{Var}_y + \text{Bias}_y^2}$$

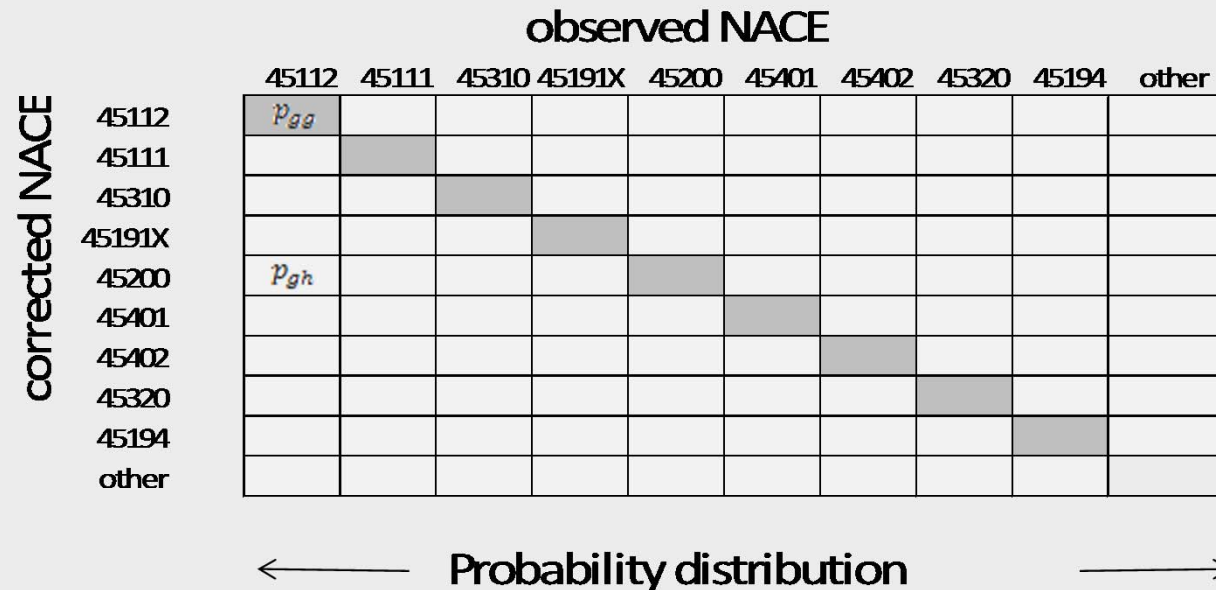
Three issues

How to ...

1. Determine the size of the error type(s)
2. Model error and compute the accuracy
3. Control the accuracy in production



Distribution of (NACE) Codes (unit level)



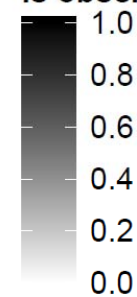
Estimate

1. Sample 25 SU per observed NACE code. Two experts determined correct NACE code
2. Yearly transitions of BR
3. Expert knowledge: grouping of transitions

Diagonal elements of transition matrix

Probability class	45112	45111	45310	45191X	45200	45401	45402	45320	45194
Supplement	1	1	1	1	1	1	1	1	1
Most complex; 5+	1	1	1	1	1	1	1	1	1
Most complex; 4	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Most complex; 0-3	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Complex; 6+	1	1	1	1	1	1	1	1	1
Complex; 5	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Complex; 0-4	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Simple; 5+	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Simple; 4	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Simple; 0-3; 3+ LU	0.88	0.02	0.75	0.53	0.74	0.15	0.73	0.16	0.52
Simple; 0-3; 1-2 LU	0.97	0.1	0.93	0.84	0.93	0.44	0.92	0.48	0.83

Probability that true industry code is observed

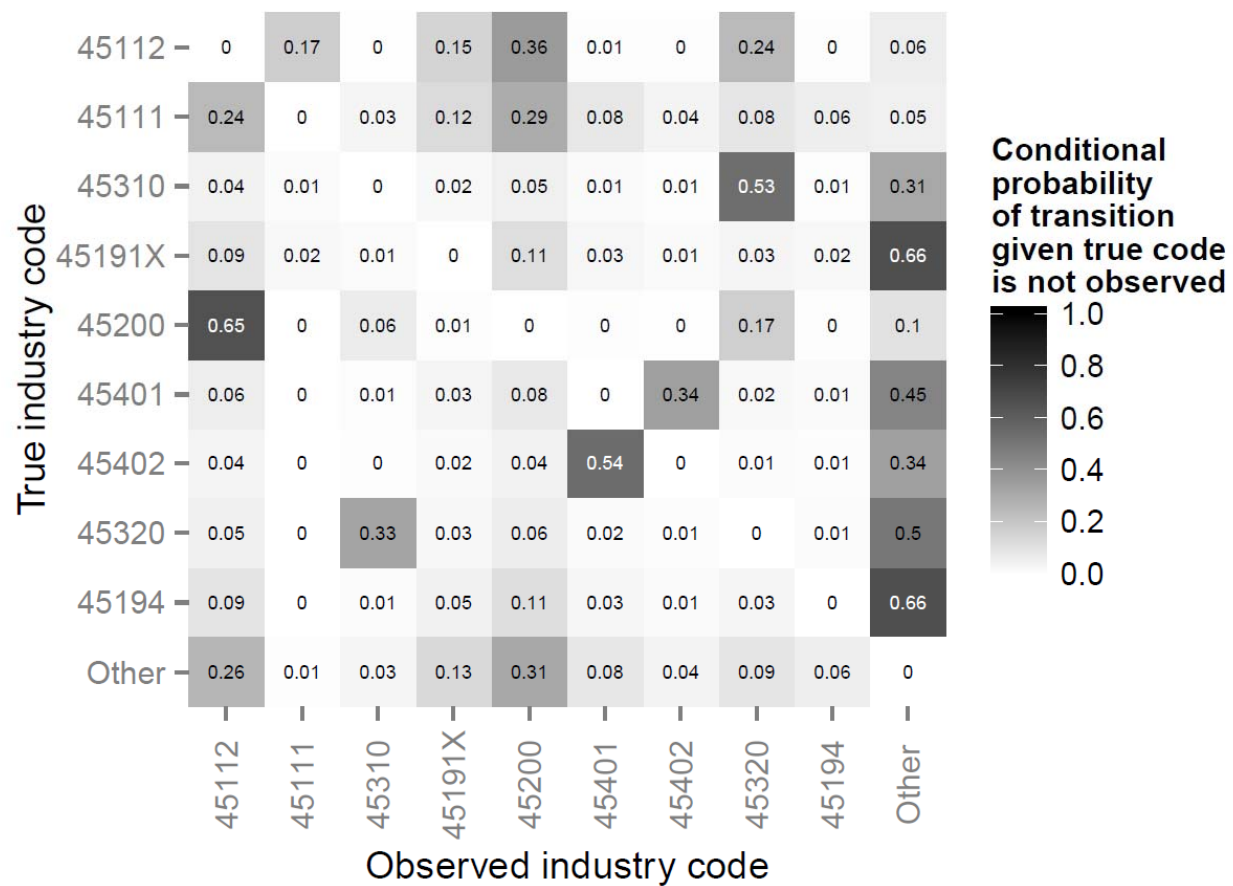


Logit model

Tested background variables:

- NACE code
- # LU
- Size class
- Receives CBS-questionnaire?
- Legal form

Off-diagonal elements



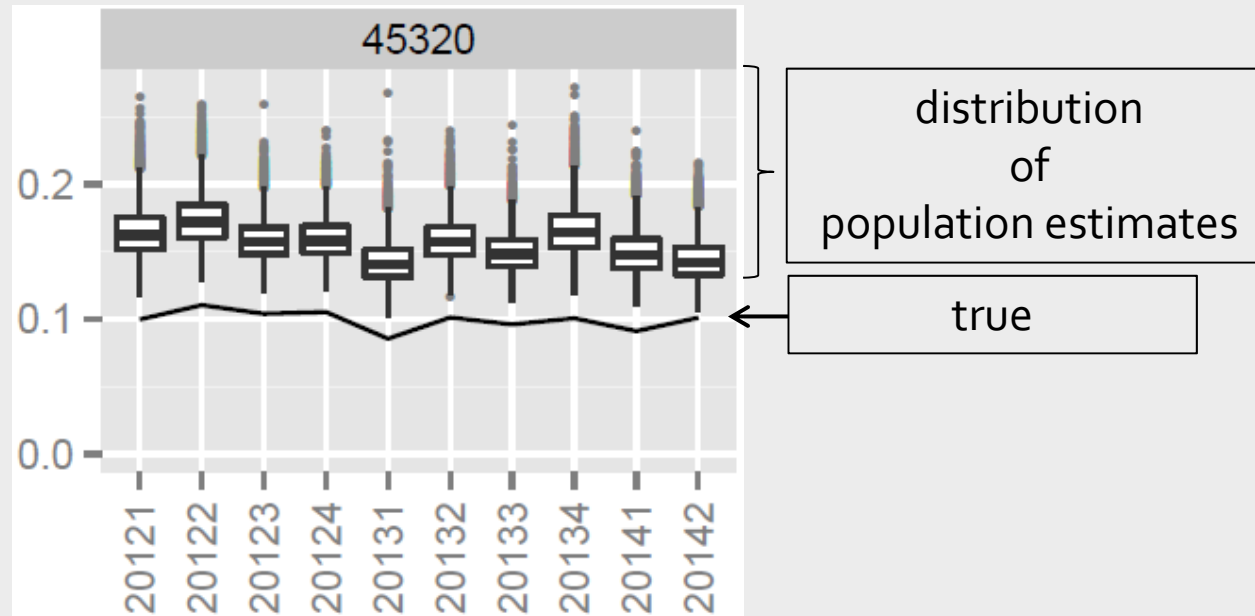
Estimating accuracy

		observed NACE									
		45112	45111	45310	45191X	45200	45401	45402	45320	45194	other
corrected NACE	45112	p_{gg}									
	45111										
	45310										
	45191X										
	45200	p_{gh}									
	45401										
	45402										
	45320										
	45194										
	other										

← Probability distribution →

		bootstrap (resample, R=10,000) NACE									
		45112	45111	45310	45191X	45200	45401	45402	45320	45194	other
observed NACE	45112	p_{gg}									
	45111										
	45310										
	45191X										
	45200	p_{gh}									
	45401										
	45402										
	45320										
	45194										
	other										

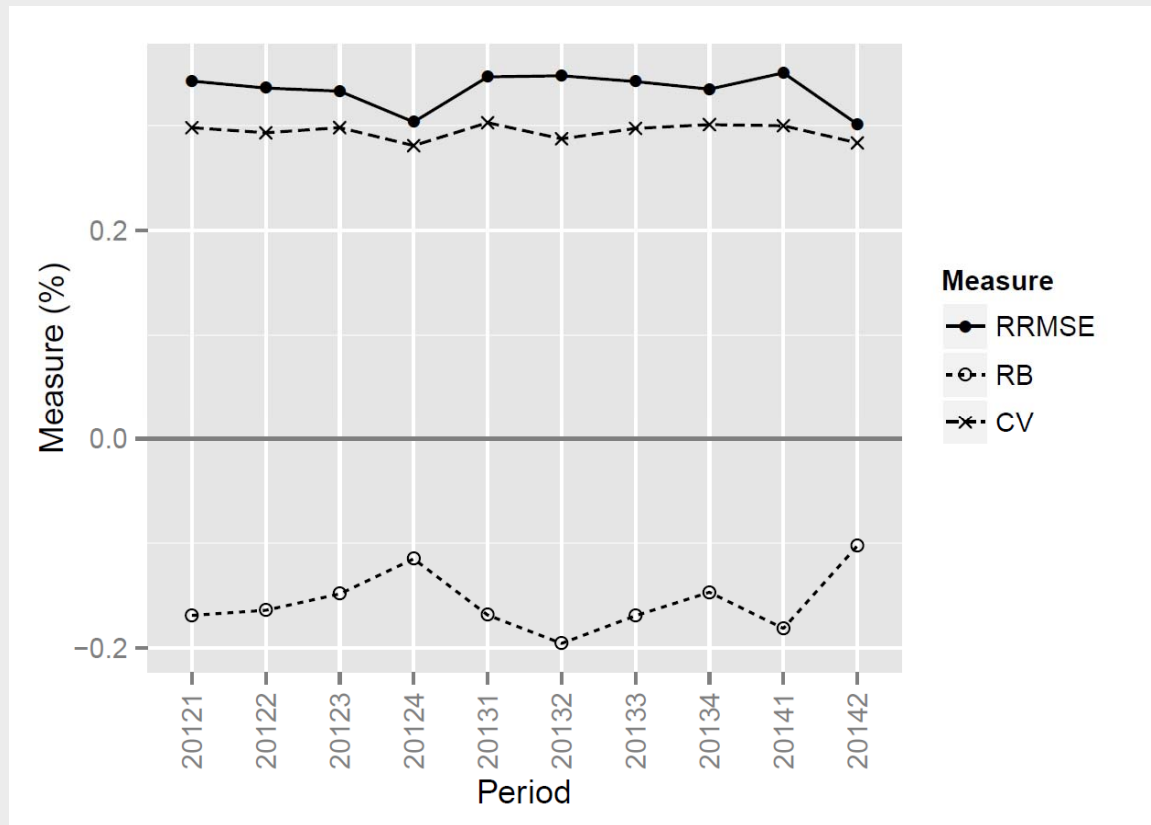
Effect of classification errors



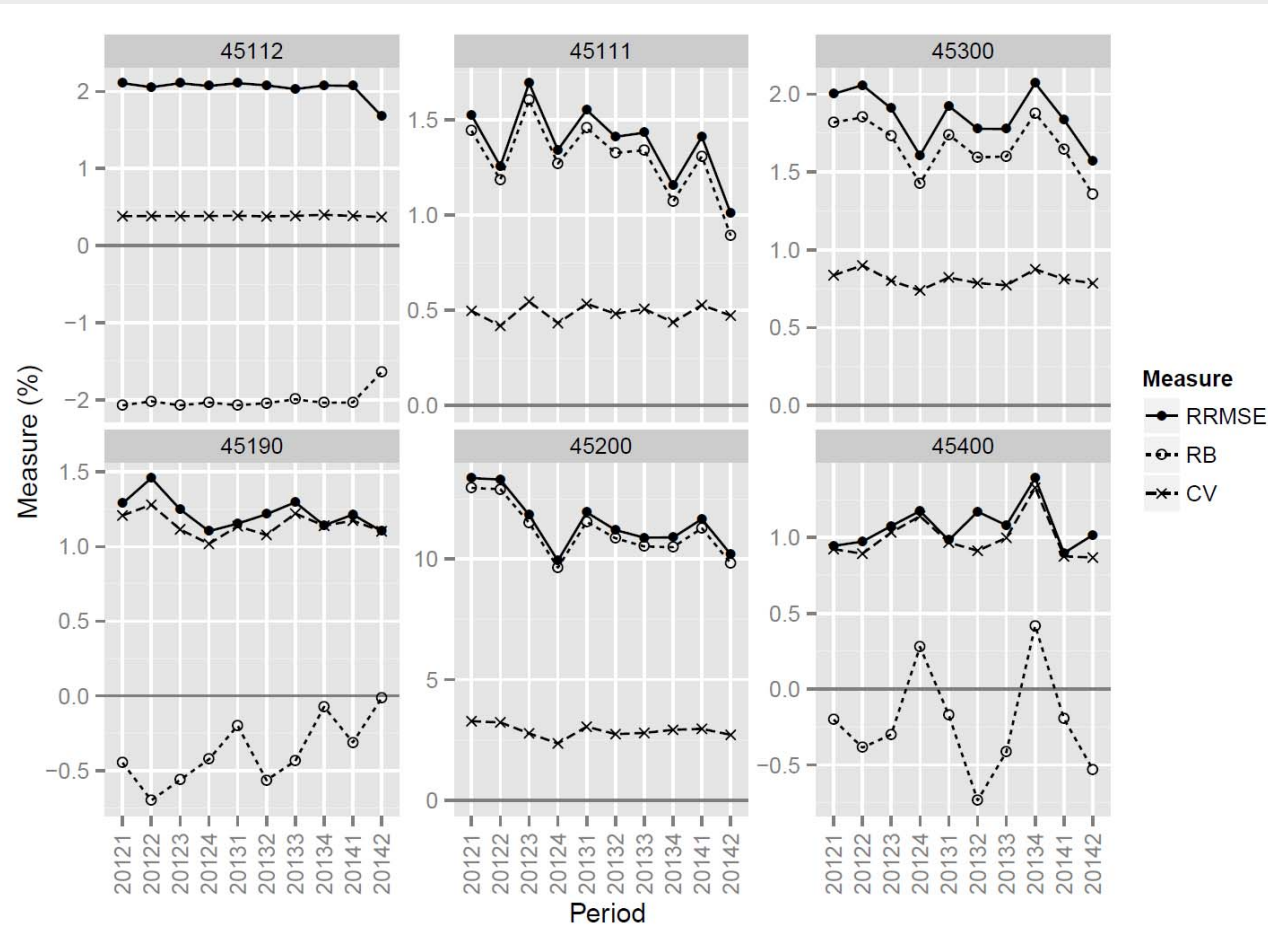
True:
$$\text{RMSE}_y = \sqrt{\text{Var}_y + \text{Bias}_y^2}$$

Simulated:
$$\widehat{\text{RMSE}}_y = \sqrt{\widehat{\text{Var}}_y + \widehat{\text{Bias}}_y^2}$$

Estimated accuracy: car trade



Accuracy six STS-publication cells



Improve accuracy?

We have editing

- A. by a CBS-team for the most complex units across outputs
- B. in daily production for specific output

Vary editing effort B

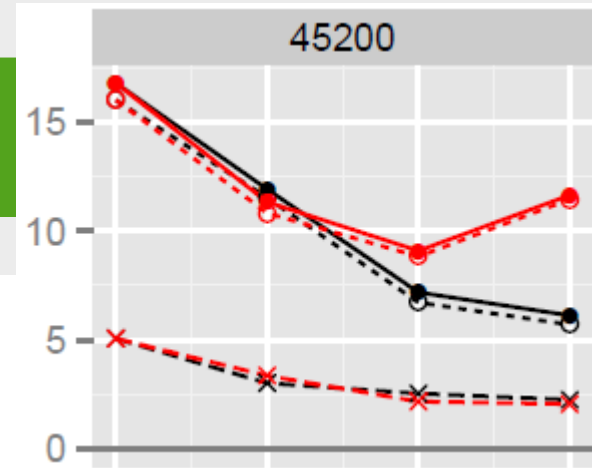
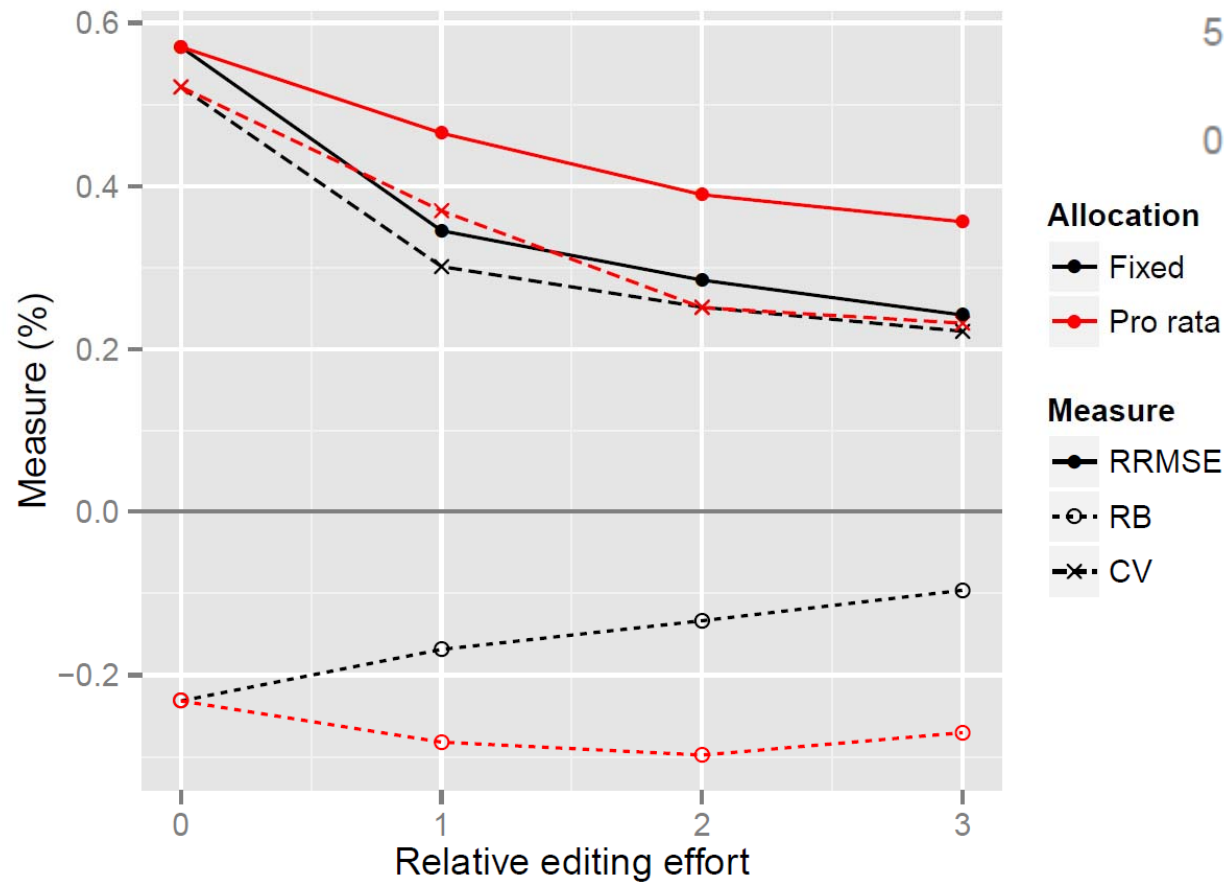
- 9 (base cells) x 25 (avg. # units per base cell) * k
 $k = 0$ (no editing B), 1 (current situation ≈ 60 hrs), 2, 3
- Sort non-edited units from large to small turnover

Scenarios

- Fixed: each base cell same effort
- Pro rata: $n(h) \propto \text{RMSE}(h) * N(h)$

! Outcomes are an approximation (true code not known)

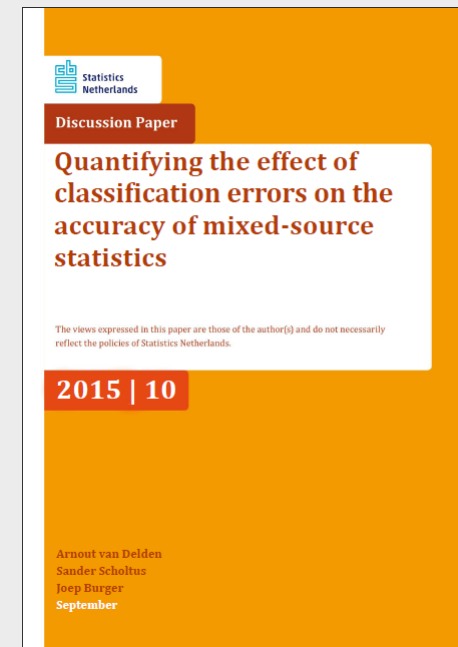
Estimated accuracy



Reference

Technical paper can be found at CBS-internet site:

<http://www.cbs.nl/nl-NL/menu/methoden/onderzoekmethoden/discussionpapers/archief/2015/default.htm>



Conclusions & future work

1. Determine the size of the error type(s)
 - It is possible to collect data on classification errors with a moderate sample size & re-use of available data
 - How to “scale up” to other branches and other error types?
 - How to estimate time-related classification errors?
2. Model error and compute the accuracy
 - We succeeded in obtaining accuracy estimates, but we used a “specific” model with three model “parts”: diagonal, non-diagonal and other elements.
 - Effect of model parameters on ‘accuracy’ of our accuracy estimates unknown
 - How to obtain a more integrated & generic model?



Conclusions & future work

3. Control the accuracy in production

- We found that editing the least-accurate publication cells is not effective
- We seek a method to effectively improved publication cells for classification errors (categorical variable) to meet predefined thresholds

