Producing contingency table estimates integrating survey data and Big Data

G. Bianchi, G. Barcaroli, P. Righi, M.Rinaldi

Italian National Institute of Statistics (ISTAT)

Outline

- Case study: Istat "Survey on ICT in Enterprises"
- Estimation procedure using data from the websites
- Coherence issues between current and experimental estimates
- Estimator for consistent two-way distributions and composite variables
- Results
- Conclusions

Case study: Istat Survey on ICT in Enterprises

The work deals with coherence issues for integrating estimates based on different sources such as Big Data and surveys

The case study of the Istat Survey on ICT in Enterprises is taken into account

- The survey is the Italian version of the *European Community Survey on ICT usage and e-commerce in enterprises*
- Target population: enterprises with 10 or more employees, in different areas of industry and services (184,000 enterprises in 2017)
- Variables related to: information and communication technology, the internet, e-government, e-business and e-commerce in enterprises

Case study: Istat Survey on ICT in Enterprises

- Sampling strategy:
  - Stratified Simple Random Sampling design combining:
    - economic activity
    - geographical area (NUTS II region)
    - class of number of employed persons

- Sample size ≅ 32,000 in 2017 (sampling rate of 18%).
- Respondents ≅ 21,000 In 2017 (response rate 66%)

- Model assisted calibration estimator:
  - number of enterprises and employed persons by domain defined by the stratification variables

Case study: Istat Survey on ICT in Enterprises

- Questionnaire: many questions focus on ICT usage and particularly on website functionalities

	Use of a Website		
C9.	**Does your enterprise have a Website?** (Filter question)	Yes ☐	No ☐ ->go to C11

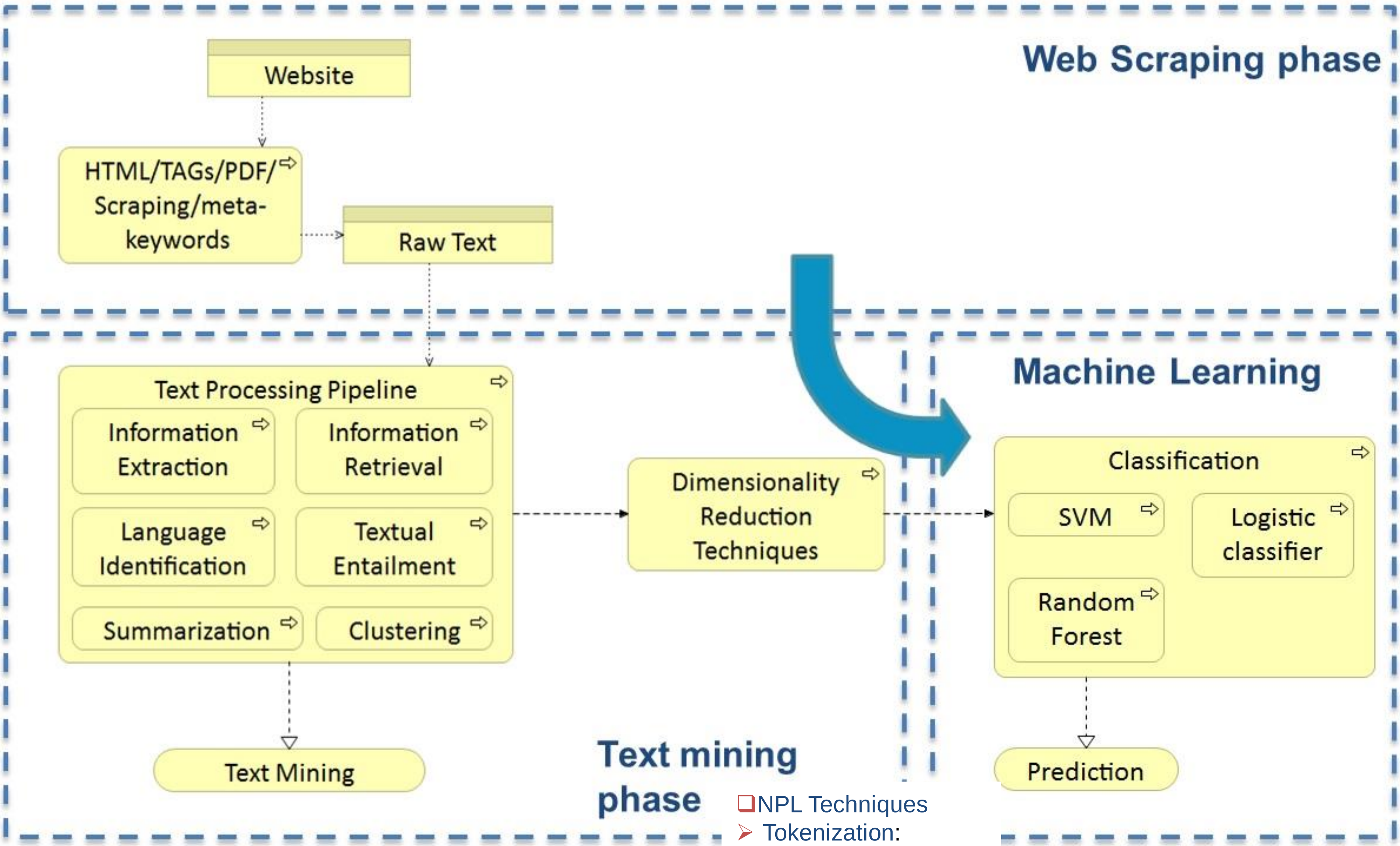
C10.	**Does the Website have any of the following?**	Yes	No
	a) Description of goods or services, price lists	☐	☐
	*[8] b) Online ordering or reservation or booking, e.g. shopping cart	☐	☐
	c) Possibility for visitors to customise or design online goods or services	☐	☐
	d) Tracking or status of orders placed	☐	☐
	e) Personalised content in the website for regular/recurrent visitors	☐	☐
	f) Links or references to the enterprise's social media profiles	☐	☐
	g) Advertisement of open job positions or online job application *- Optional*	☐	☐

Estimation procedure using data from the websites

- In 2016 Istat started to investigate a new procedure to enhance the estimates of website related variables by automatic collection of information from the web

1 – Web address acquisition	URL from the admin sources
	URL from thematic directory sites
	URL from batch queries on search engines (**URL Retrieval** techniques in case of non existing URL)
2 – Enterprise identification	URL validation, check URL's validity (recurring errors and domain extraction)
	Detection of identification variables from the website and **comparison** with the same information available in the SBR register
3 – Data analytics	***Web Scraping*** techniques for web data acquisition
	Text Mining techniques for extracting the requested information
	Machine Learning techniques for the use of algorithms that simulate a learning process for the construction of predictive models
4 – Inference	**From the enterprises with scraped websites to the enterprises of the target population**

Estimation procedure: data analytics phase



- NPL Techniques
  - Tokenization:
  - Lemmatization
  - Part-Of-Speech recognition (POS tagging)

Estimation procedure: Machine learning

- A compared evaluation of learners performance has been carried out for the target variable "web ordering functionalities (yes/no)" (Bianchi *et al.*, 2015, 2018, 2019)
  - A **dataset** of **4,755 enterprise** websites with known class label: the dataset is **imbalanced**, roughly 20% positive and 80% negative
  - **50%** as **training set** and **50%** as **test set**

Learner	Accuracy	Recall	Precision	F1-measure
	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$	$\frac{2TP}{2TP+FP+FN}$
Logistic	0.88	0.64	0.66	0.65
SVM	0.90	0.62	0.76	0.68
Random Forest	0.90	0.72	0.74	0.73

Carried out by Python **scikit-learn** library

Estimation procedure: Inference notation and estimator

- U_1 the target population of size $N_1 \cong 133{,}000$
- U_2 the website scraped population of size $N_2 \cong 90.000$
- Target Parameter: $\bar{Y} \ = \frac{1}{N_1} \sum_{U_1} y_k$
- Assumptions for model unbiased estimator
  - $y_k \ = \tilde{y}_k$ (prediction by ML)
  - $P(k \in U_2 | y_k = 1, z_k = j)$=$P(k \in U_2 | y_k = 0, z_k = j)$

  Akin to **MAR mechanism** (Little and Rubin, 2002)
  - z_k auxiliary variable (vector), $\mathrm{Z} = \sum_{U_1} z_k$ known
- The estimator for one-way distribution

$$\hat{\hat{Y}} = \frac{1}{N_2} \sum_{U_2} \tilde{y}_k \, \gamma_k \, (z_k \, , Z)$$

- γ_k computed by calibration algorithm (Deville and Särndal, 1992) - Pseudo-calibrated estimator $\sum_{U_2} z_k \, \gamma_k \ = \ Z$

Estimation procedure: experimental statistics

Estimates using this process find at Istat website:
https://www.istat.it/en/archivio/216641

EXPERIMENTAL STATISTICS

EXPERIMENTAL STATISTIC

ESTIMATES OF MODALITIES OF USE OF WEBSITES BY ENTERPRISES

The estimates of the modalities of use of websites by enterprises – produced using directly Internet data – are made available by Istat.

In download, in Excel format, the estimates of the rate of enterprises (on the total reference population) that own or use a website in which are available:

1. web ordering functions (e-commerce component);
2. information on job vacancies;
3. links to social media (Facebook, Twitter, Instagram etc.);
4. all the information above, organized by NACE level 2.

The estimates, referring to 2017, concern a reference population of about 184,000 enterprises with at least 10 persons employed operating in different sectors of economic activity.

Data are obtained through a procedure based on web scraping and natural language processing techniques.

Estimation procedure: experimental statistics

ESTIMATES CONCERNING RATE OF ENTERPRISES OFFERING	Design-based estimates	Confidence interval		Estimates with internet data
		Lower bound	**Upper bound**	
WEB ORDERING FUNCTIONALITIES IN THE WEB	**14.97**	**13.81**	**16.13**	**15.51**
JOB ADVERTISEMENTS IN THE WEB	**10.78**	**10.02**	**11.53**	**13.91**
LINKS TO SOCIAL MEDIA IN THEIR WEBSITES	**31.25**	**29.90**	**32.60**	**36.68**

❑ Evidences

- **Higher accuracy** (Barcaroli, Righi, Golini, 2018)
- **Yearly basis statistics** (questionnaire does not collect all variable every year - multi-year basis statistics- while ML prediction is stable over time)
- **Reduction of the Measurement errors** for complex variables (i.e. Web ordering facilities, social media)
- **Reduction of the response burden**
- **Larger differences in small domains**

Coherence issues between current and experimental estimates

- Along with simple one-way distributions **Istat must produce estimates of composite variables or contingency tables where some of the involved variables are collected only by the survey**

- Design based estimates must be consistent with respect to the estimates based on internet data (Internet Data Based – IDB estimates)

Coherence issues: contingency table

- Example: 2017 data – current design based estimates

Did the enterprise sell products or services using the website or apps in 2016? **e_awsell**

	e_webord=0	e_webord=1	Total
e_awsell = 0	83.23	6.88	90.1
e_awsell = 1	1.80	8.10	9.9
Total	85.03	14.97	100.0

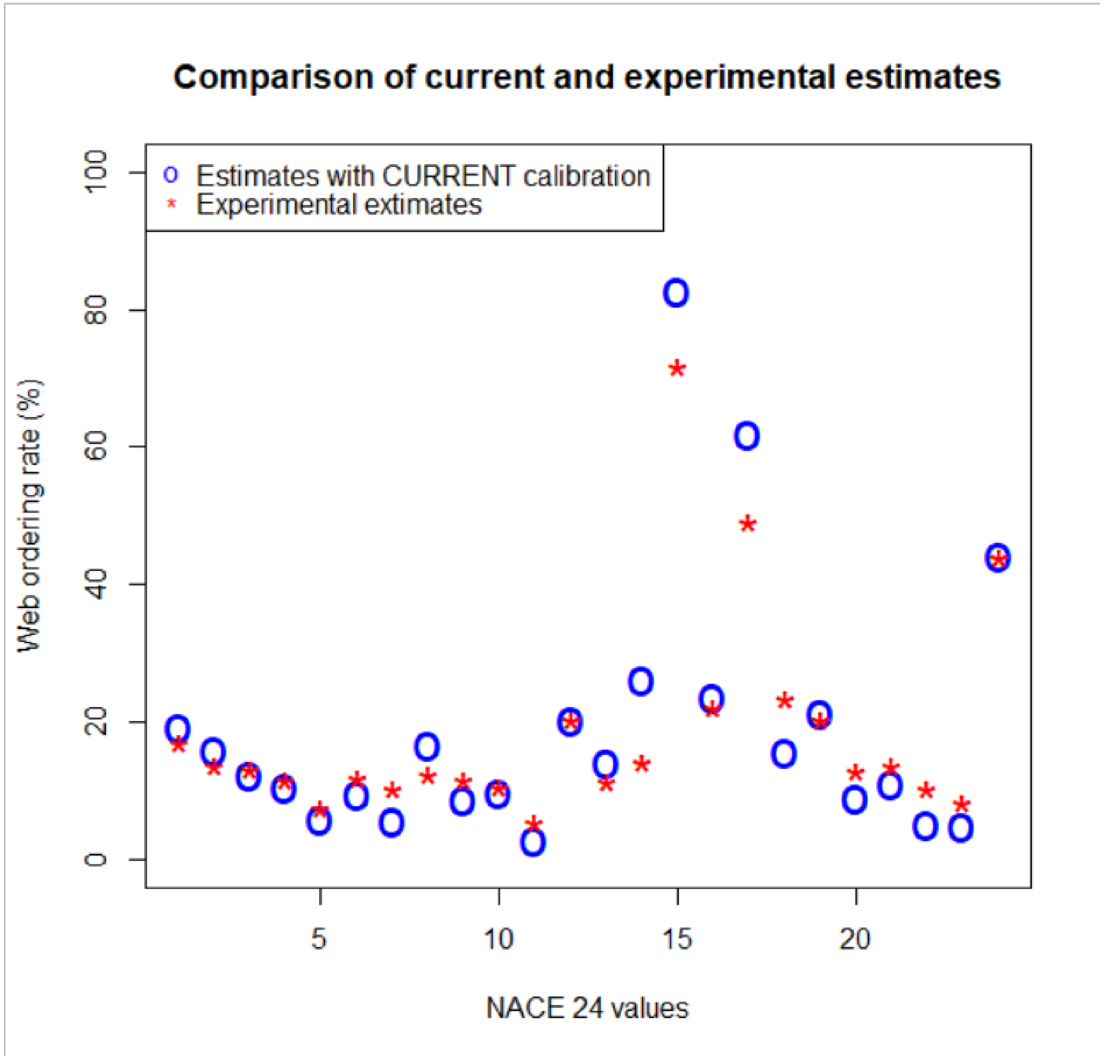
IDB estimator = 15.51

Coherence issues: contingency table

Three-way contingency tables

«e_webord *by* e_awsell *by* Economic Activity (Nace)»

The marginal distribution of e_webord compared with the IDB one-way distribution in the experimental statistics, show some deviations are very high.



Coherence issues: composite variable

- Example: 2017 data
- **e_webf3**= e_webord and[presence of (“Tracking online order” or “Product or price list” or “Functionalities for customizing the website contents” or “Functionalities for customizing the products”)]

- **e_webmaturity** = e_webord and “Tracking online order” and “Product or price list” and “Functionalities for customizing the website contents” and “Functionalities for customizing the products”

	Design-based estimates	Estimates with internet data
e_webord	_	**28,669.31**
e_webf3	**23,276.8**	_
e_webmaturity	**1,346.7**	_

Coherence issues: composite variable

- At domain level the difference could be negative

example e_webord-e_webf3<0

Where e_webord total estimates are IDB and the e_webf3 totals are estimated by the current design-based estimator

	Estimates with internet data (A)	Design-based estimates (B)	A-B
Nace	e_webord	e_webf3	difference
naceistw01	1270,3	1385,9	-115,5
naceistw02	1465,1	1433,2	31,9
naceistw03	682,8	596,4	86,4
naceistw04	996,6	735,6	261,0
naceistw05	1036,3	737,4	299,0
naceistw06	165,0	126,8	38,2
naceistw07	1166,0	608,7	557,3
naceistw08	202,0	268,3	-66,3
naceistw09	964,9	722,7	242,1
naceistw10	322,1	196,9	125,3
naceistw11	1054,3	363,6	690,7
naceistw12	7435,7	6784,6	651,1
naceistw13	1453,1	1022,4	430,7
naceistw14	31,0	57,2	-26,2
naceistw15	4235,9	4398,3	-162,5
naceistw16	3081,5	1919,8	1161,8
naceistw17	219,0	268,9	-49,9
naceistw18	142,0	94,3	47,8
naceistw19	56,0	55,5	0,5
naceistw20	639,0	408,4	230,5
naceistw21	87,0	69,3	17,7
naceistw22	817,0	346,2	470,8
naceistw23	895,7	435,2	460,5
naceistw24	251,0	241,5	9,5

Estimator for consistent two-way distribution

- Two approaches to deal with consistency issues:

  - Mass imputation or projection estimator (model assisted: Kim and Rao, 2012; model based: Valliant et al. 2001). The current estimation procedure changes completely

  - Calibration approach (model assisted) adding the IDB estimates in the current calibration constraints. Minor impact on the current estimation procedure

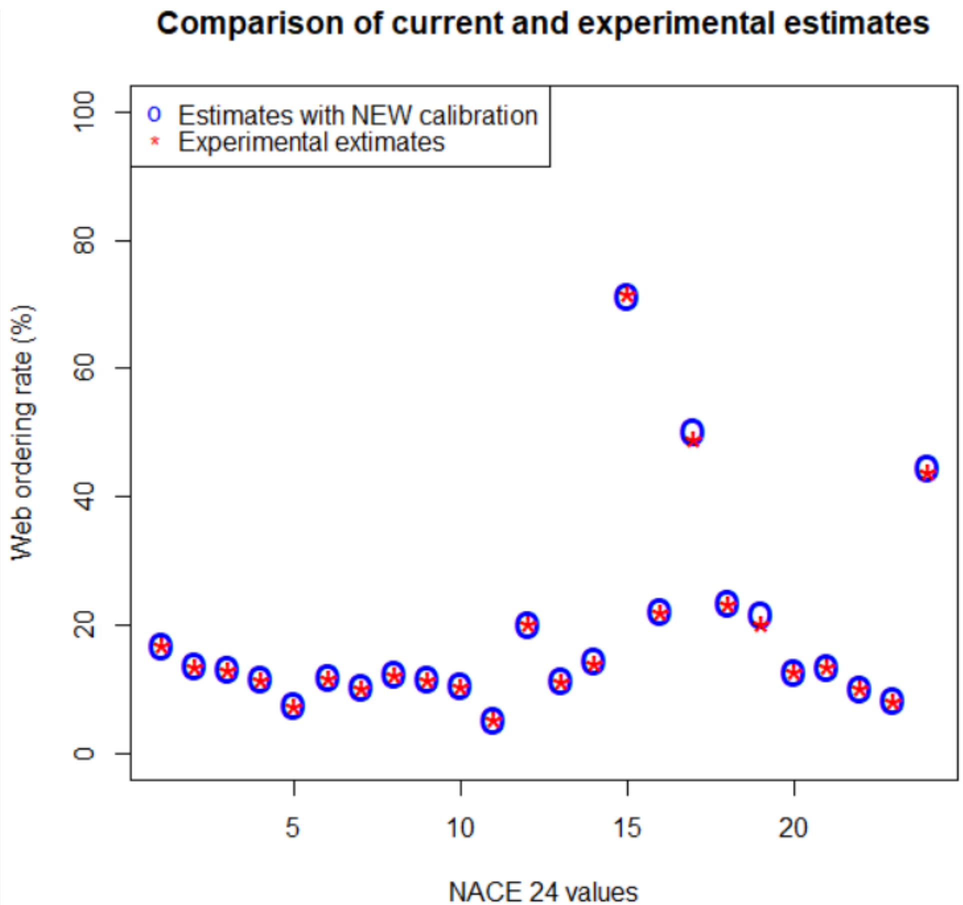
Estimator for consistent two-way distribution

- Preparing the calibration

A. Analysis of the one-way distributions by domain of the IDB estimates and two-way distributions or composite one-way distributions by domain to be published
B. Common domains become calibration domains
C. Replace the observed survey values by the predicted values for the variables of the IDB estimates and use them for the estimates (assuring the consistency)
D. Include the IDB estimates in the set of the calibration constraints

Results: two-way distribution

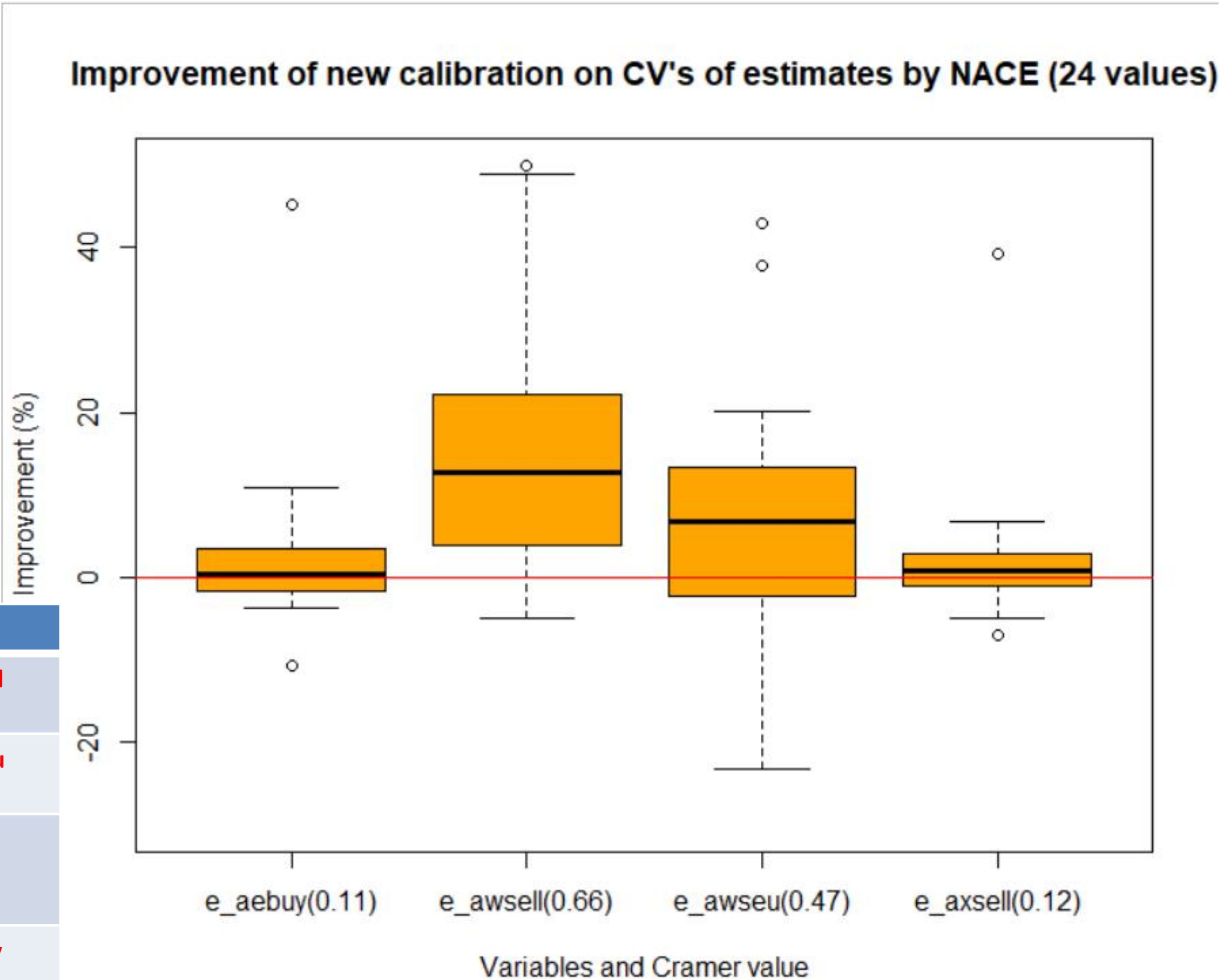
	e_webord=0	e_webord=1	Total
e_awsell = 0	82.46	7.49	89.95
e_awsell = 1	2.03	8.02	10.05
Total	84.49	15.51	100.00



Results: two-way distribution

- Calibration and accuracy of the estimates

Coefficient of variation of the estimates decreases when moving from the old calibration to the new one, but **only if there is dependence with e_webord**.



Question	Id
Did the enterprise sell products or services using the website or apps in 2016?	e_awsell
If yes, the geographic area of customers was in EU countries (Italy excluded)?	e_awseu
Did the enterprise sell products or services using EDI (Electronic Data Interchange) in 2016?	e_axsell
Did the enterprise buy products or services by using websites or apps of other enterprises or by using EDI in 2016?	e_aebuy

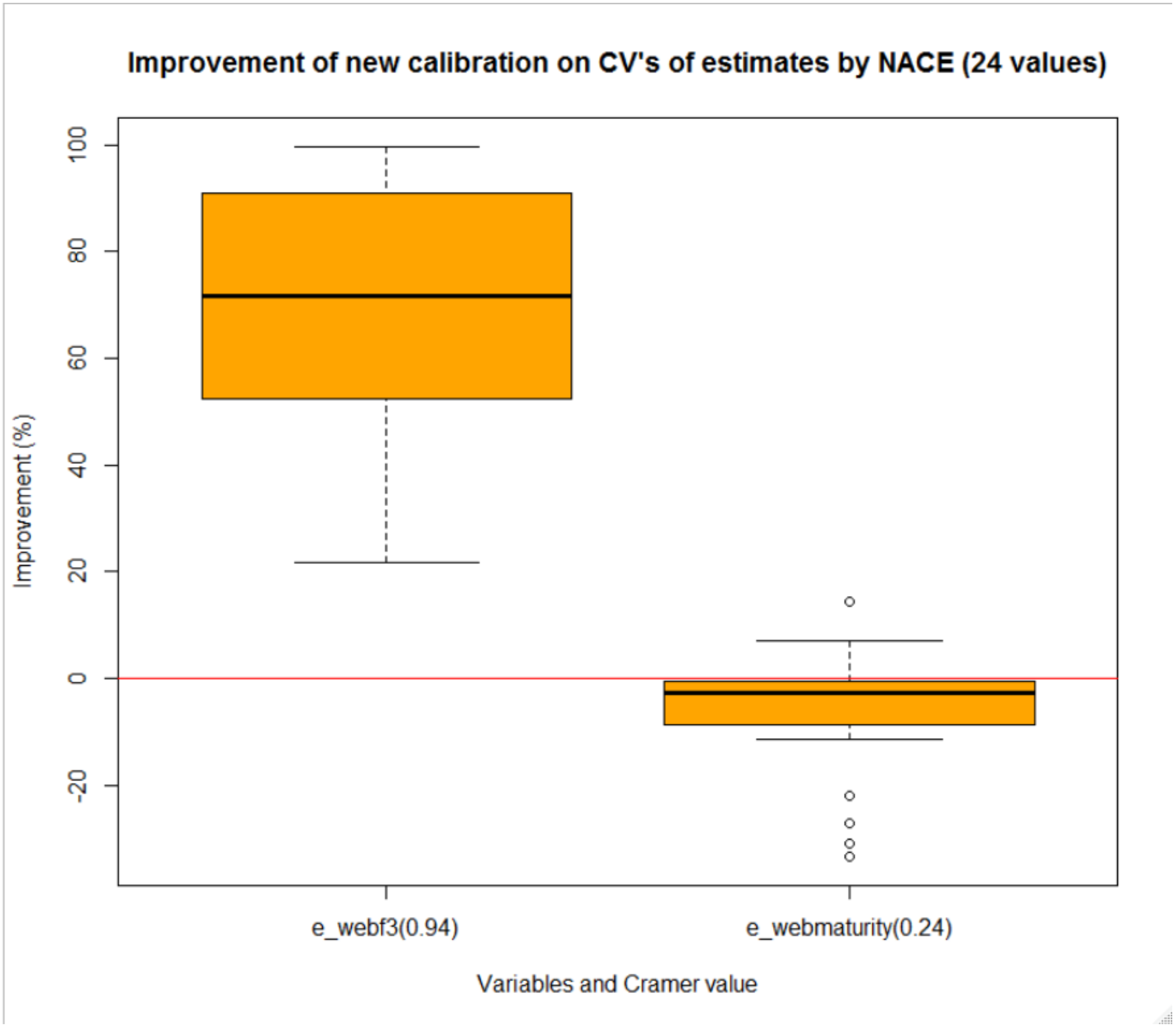
Results: composite variable distribution

Nace	Estimates with internet data (A)	Design-based estimates (B)	A-B	B/A	
	e_webord	e_webf3	difference	New	Old
naceistw01	1270.3	1196.7	73.6	0.94	0.96
naceistw02	1465.1	1240.0	225.1	0.85	0.85
naceistw03	682.8	620.7	62.1	0.91	0.93
naceistw04	996.6	806.2	190.4	0.81	0.83
naceistw05	1036.3	947.5	88.9	0.91	0.92
naceistw06	165.0	152.3	12.6	0.92	0.97
naceistw07	1166.0	1113.7	52.3	0.96	0.97
naceistw08	202.0	193.3	8.7	0.96	1.00
naceistw09	964.9	957.0	7.9	0.99	1.00
naceistw10	322.1	175.7	146.4	0.55	0.67
naceistw11	1054.3	674.6	379.7	0.64	0.65
naceistw12	7435.7	6631.1	804.6	0.89	0.91
naceistw13	1453.1	797.9	655.2	0.55	0.57
naceistw14	31.0	30.4	0.6	0.98	1.00
naceistw15	4235.9	3807.5	428.4	0.90	0.90
naceistw16	3081.5	1934.1	1147.4	0.63	0.59
naceistw17	219.0	204.7	14.3	0.93	0.98
naceistw18	142.0	139.7	2.3	0.98	1.00
naceistw19	56.0	42.5	13.6	0.76	0.94
naceistw20	639.0	568.3	70.7	0.89	0.91
naceistw21	87.0	81.1	5.9	0.93	0.97
naceistw22	817.0	646.0	171.0	0.79	0.83
naceistw23	895.7	749.5	146.2	0.84	0.84
naceistw24	251.0	231.4	19.6	0.92	0.96

Results: composite variable distribution

❑ Calibration and accuracy of the estimates

Coefficients of variation of the e_webmaturity estimates increases in some domains (**low dependence with e_webord**)



Conclusions

- In the era of Big Data, the survey still remains the main source for collecting some variables
- The work deals with the challenge to produce statistics integrating variables from new and traditional data sources
- There can be different approaches each of them affects the data production process differently
- Here it is shown a soft approach that changes as little as possible the current data production process

Conclusions: Challenges

- We assume the Internet Data Based (IDB) estimates as known totals and we are confident the model variability is negligible:
  - Unit level predictions of 2018 real data based on the 2017 training data set show to be stable with respect to the unit-level predictions of the 2017 data
  - --> Small model variance
  - It is true for small domains as well?
  - Should we take into account the variances of the totals when estimating the variance (bootstrap , jackknife)?

Conclusions: Challenges

- In the ICT survey case the new calibration adds only e_webord but in a general more variables can be involved in the calibration
- With too many calibration variables:
  - converge could fail
  - variances could increase

- In these cases other approaches have to be planned. Some ideas:
  - Use calibrated survey estimates as the input for the pseudo-calibration and perform the described procedure
  - Relax calibration constraints in an iterative approach and use of Ridge Calibration estimator (Beaumont and Bocci, 2008);
  - Projection estimators or mass imputation (Specialized Session 7: Inference from informative and non-probability survey samples)

References

- Barcaroli G. , Golini N., Righi P. (2018). Quality evaluation of experimental statistics produced by making use of Big Data, *Proceedings Q2018*, 26-29 June, Krakow (www.q2018.pl).
- Beaumont, J.F., Bocci, C. (2008). Another look at ridge calibration. *METRON - Int. J. Stat.,* LXV I(1), 5–20.
- Bianchi G., Bruni R., Scalfati F. (2018). Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms, *Mathematical Problems in Engineering*, vol. 2018.
- Bianchi G., Bruni R. (2015). Effective Classification using a Small Training Set based on Discretization and Statistical Analysis, *IEEE Transactions on Knowledge and Data Engineering* Vol. 27(9), 2349-2361.
- Bianchi G., Bruni R. (2019). Robustness Analysis of Classifiers for Website Categorization: the Case of E-commerce Detection. *Expert Systems With Applications*, to appear.
- Deville, J. C. Särndal, C. E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 85, 376–382.
- Kim J. K., Rao J.N.K.. (2012) Combining data from two independent surveys: a model-assisted approach, *Biometrika*, 99, 85-100.
- Little R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, Wiley.
- Valliant R., Dorfman A. H., Royall R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, Wiley, New York.