Assessing the Quality of Enterprise Characteristics and Online Job Advertisements derived from Web Data

Ville Auno, Statistics Finland
Johannes Gussenbauer, Statistics Austria
WIN Conference, Gdansk 06/02/2025

Trusted Smart Statistics – Web Intelligence Network

Web Intelligence Network

Funded by the European Union

Introduction

- Assessing the quality and usability of web scraped data for official statistics production was one of the tasks carried out in the Web Intelligence Network (WIN) project
- Focus on two different data:
  - Open Job Advertisements (OJA)
  - Online-Based Enterprise Characteristics (OBEC)
- Findings provide insights into the challenges and strengths of web scraped data

Funded by the European Union

Quality Assessment of OJA Data

- Quality of OJA data was assessed with two different ways:
  - Use of pre-defined quality indicators for source evaluation
  - Manual annotation exercises for evaluating classification accuracy
- Quality indicators:
  - Number of relevant (>500 OJAs) and very relevant (>5000 OJAs) sources over time
  - Ranking of the relevant sources over time
  - Time series plots for number of OJAs for all very relevant sources
  - Stability of data over different versions of data

Funded by the European Union

OJA: Quality indicators

- Relevant sources
  - Fairly stable
  - Some fluctuation in Portugal for example

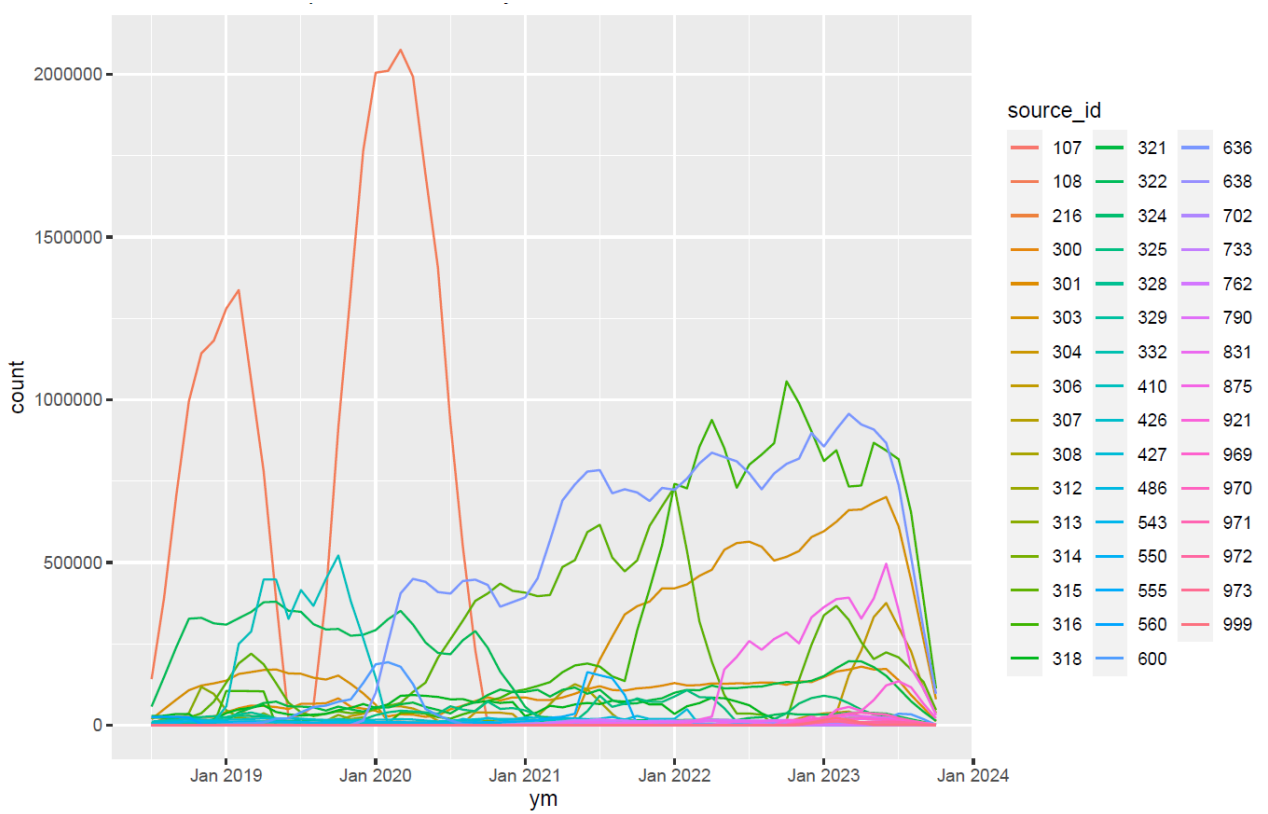
Year	AT	BG	DE	FI	FR	IT	NL	PL	PT	RO
2018	19	7	39	6	38	27	23	12	5	14
2019	18	16	44	10	44	32	29	14	10	24
2020	21	16	40	5	41	31	26	12	25	16
2021	27	21	54	9	47	34	34	14	44	21
2022	20	21	61	7	56	31	36	15	51	19
2023	15	16	43	5	42	28	26	15	44	14

- Very relevant sources
  - Similar with relevant sources
  - Larger fluctuations in relative terms in smaller countries

Year	AT	BG	DE	FI	FR	IT	NL	PL	PT	RO
2018	7	1	27	2	24	16	9	7	2	2
2019	13	5	31	6	25	23	15	11	4	11
2020	9	4	28	4	22	17	12	8	10	7
2021	8	4	28	4	25	21	15	8	14	8
2022	7	5	26	3	31	19	14	11	16	9
2023	4	4	21	3	26	17	13	11	12	6

OJA: Quality indicators

- Stability of the relevant and very relevant sources were analyzed further:
  - Very relevant sources do not remain the same over the years
  - Relative significance of the sources vary from year to year – sources also disappear
- These findings may raise concerns about the stability of the data sources

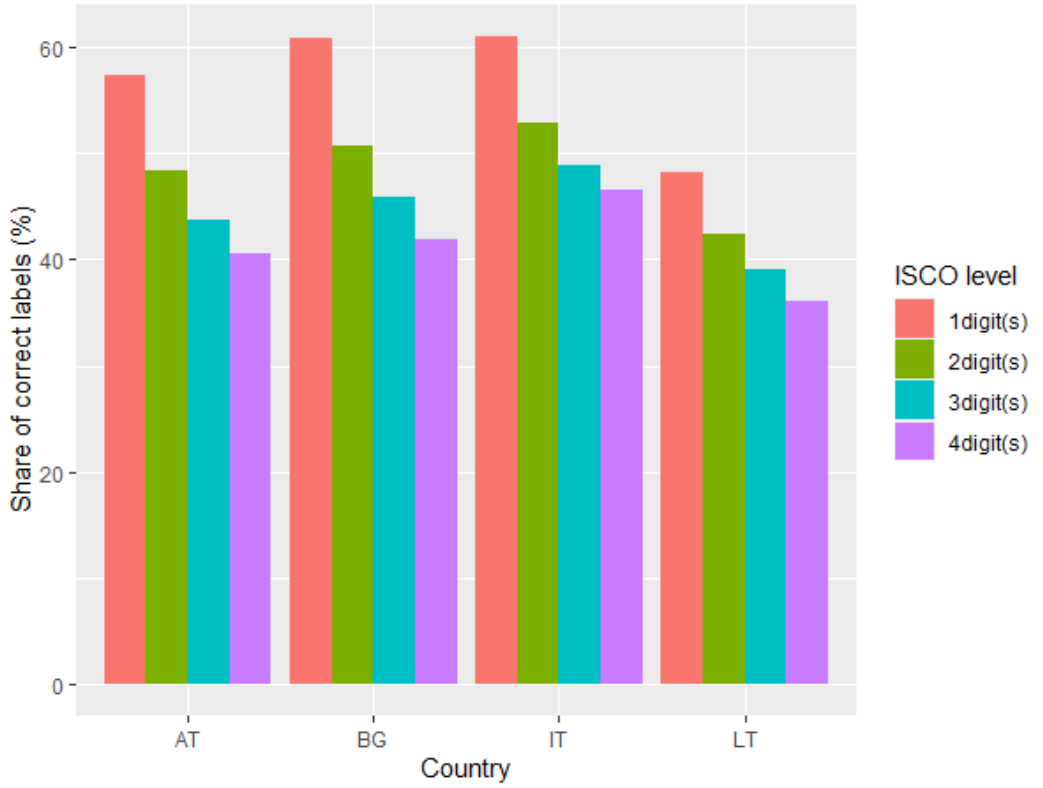


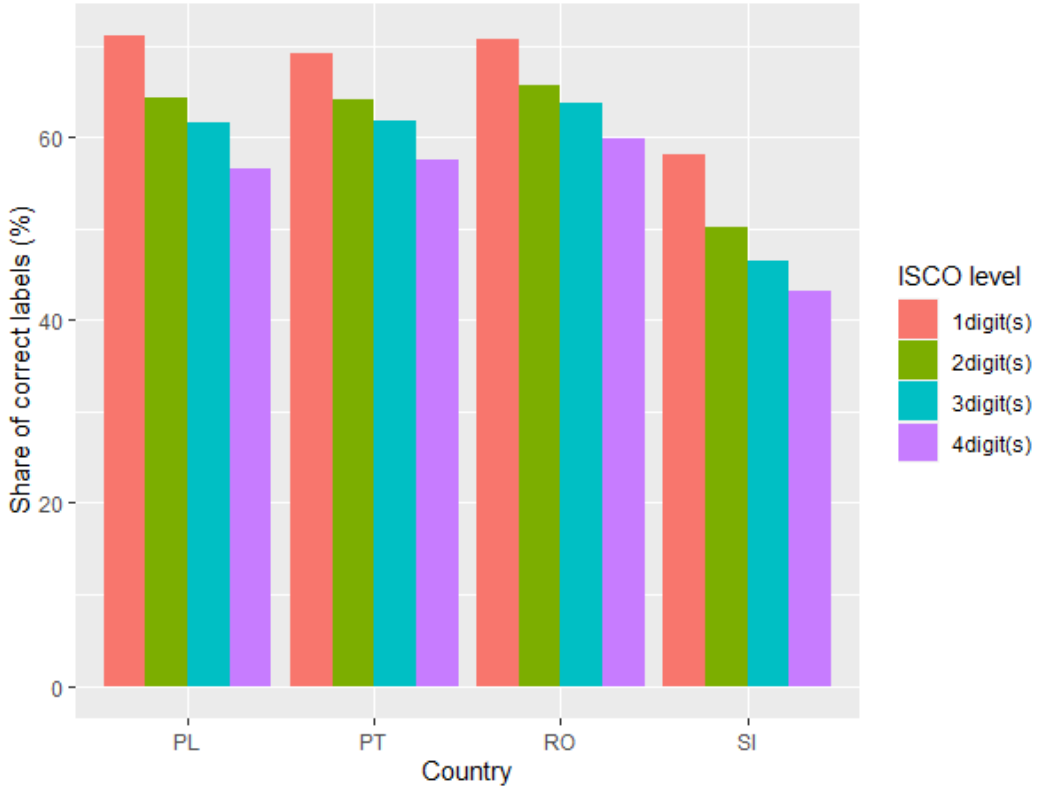
OJA: Manual annotation

- To evaluate the accuracy of various classifications in the data
- Two separate annotation rounds:
  - First one focuces on the classification of occupations (ISCO)
  - Second one included occupation, education, working time, location and economic activity
- First annotation sample was 350-400 job advertisements per country (Austria, Bulgaria, Italy, Lithuania, Poland, Portugal, Romania and Slovenia)
- Second sample was 300 job advertisements per country (Austria, Bulgaria, Finland, France, Germany, Italy, Poland and Slovenia)

Funded by the European Union

OJA: Manual annotation – 1st round





OJA: Manual annotation – 1st round

- Accuracy patterns are similar across all countries
- Highest 1-digit accuracy: Poland 71.07%, lowest: Lithuania 48.28%
- Accuracy drops significantly when moving from 1-digit level to 2-digit level
- Accuracy overall quite poor – probably not good enough for official statistics as stand-alone data source without methodological corrections

Classification	1digit(s)	2digit(s)	3digit(s)	4digit(s)
Correct	62.04	54.83	51.40	47.81

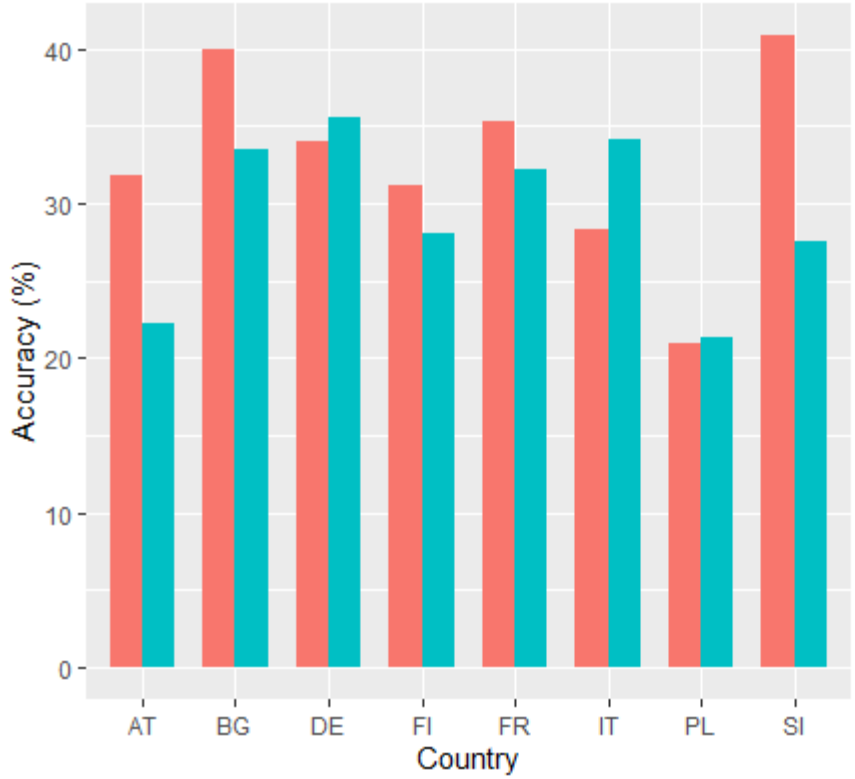
Funded by the European Union

OJA: Manual annotation – 2nd round

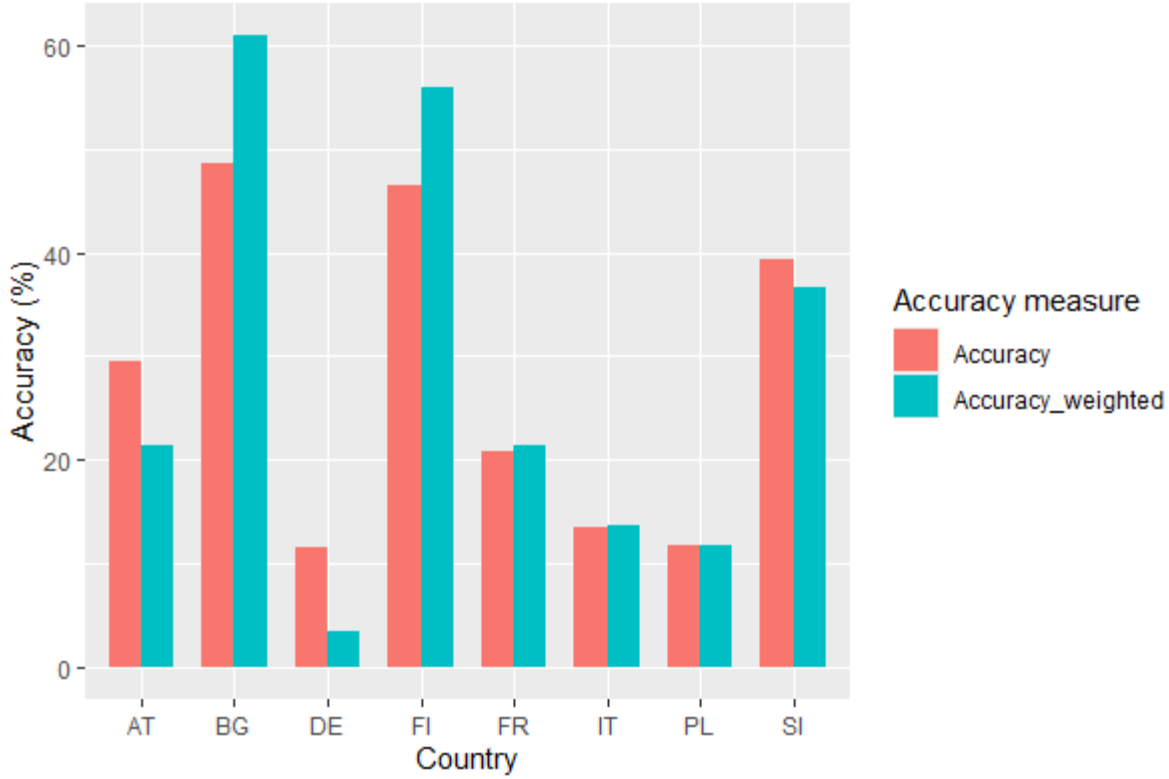
- Accuracy of other classifications within the data leaves hopes for improvement
- Accuracy of occupation even dropped from the 1st round

Classification	Share of correct labels (%)	Weighted share of correct labels (%)
Economic activity	30.99	30.68
Education	25.17	15.54
Occupation	56.23	56.29
Location	64.25	60.74
Working time	67.88	66.99

OJA: Manual annotation – 2nd round



Economic activity



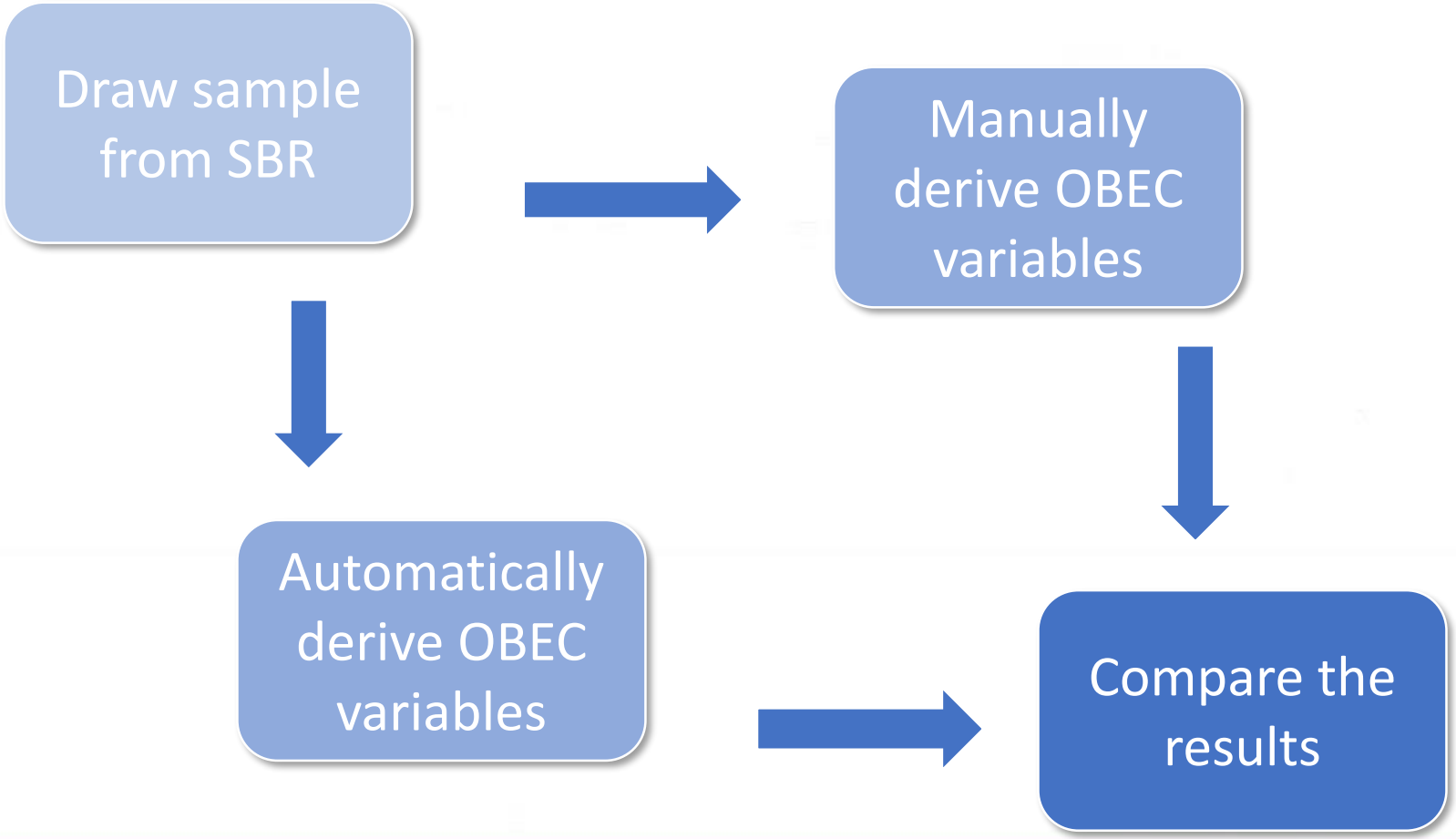
Education

Quality Assessment of OBEC Data

- OBEC variables:
  - Enterprise has a website (URL-Linking)
  - Social Media Link on website (SMP)
  - Onlineshop on website (E-Commerce)
- Multiple implementations among WIN members to do this automatically/programmatically:
  - Statistics Hesse, AT, BG, PL, IT, …
  - R, Python, Java, …
- Compare quality of different implementation ➔ OBEC annotation round

Funded by the European Union

OBEC: Outline of annotation round



Funded by the European Union

OBEC: On the details

- Agreed to sample 500 legal units from SBR
  - Small sample size due to time consuming manual work
  - Sample stratified by NACE-Sections and enterprise size
- Followed annotation guidelines from Statistics Hesse
  - Various rules on when an enterprise has a website and which is the correct one
  - Agreed definition on when an online-shop is present
  - Agreed definition on how to evaluate social media presence
  - Structural recording of manual findings for standardised output comparison
- Participating countries: Hesse(DE), DE, AT, BG, IT, PL, LT

Funded by the European Union

OBEC: Results overall

URL-linking was correct if

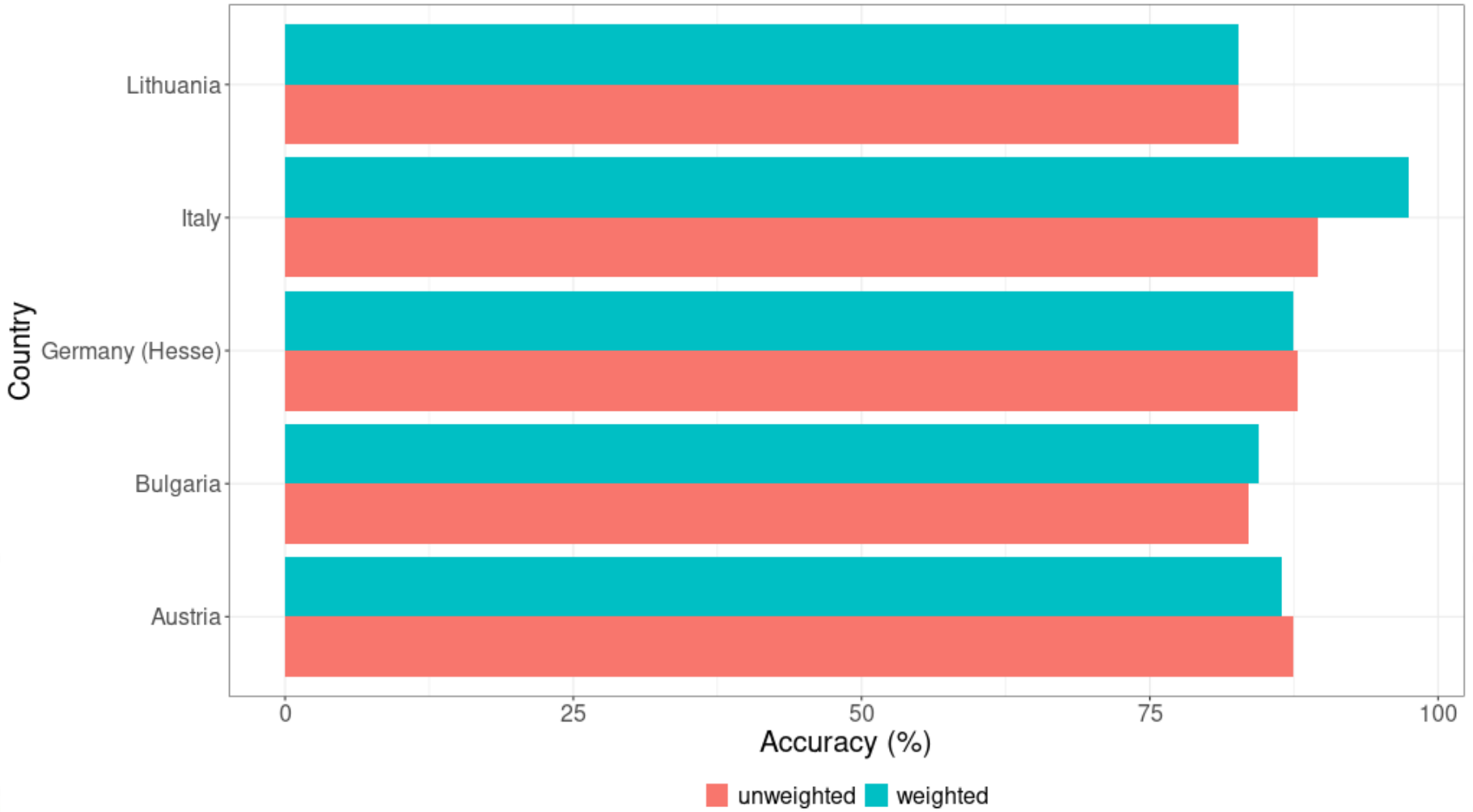
- no website was found in the annotation and none through the automatic procedure
- at least 1 website found in the annotation and through the automatic procedure is the same

SMP and Ecom accuracy only on subset were URL-linking agreed

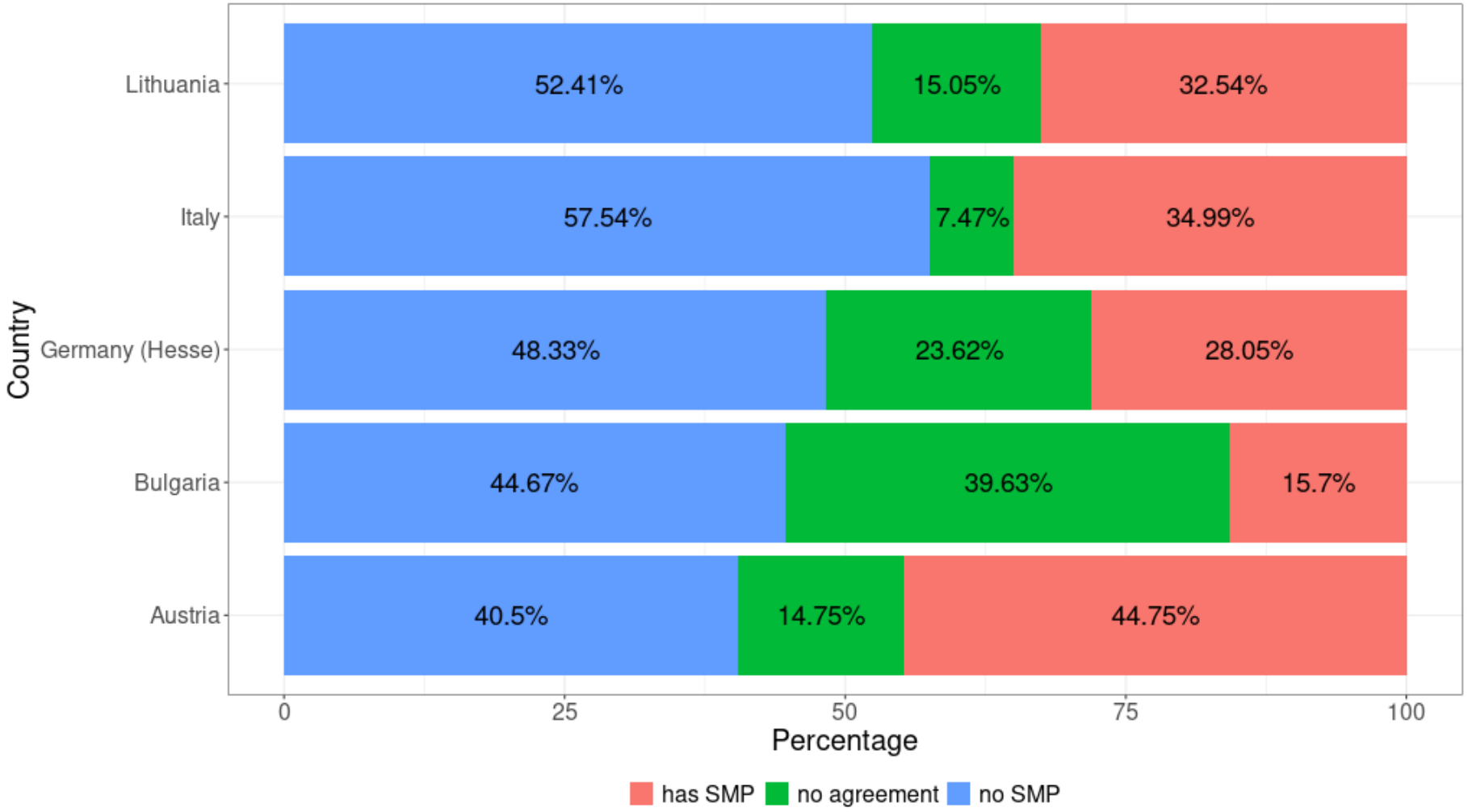
Country	URL-Linking (weighted Accuracy; %)	SMP (weighted Accuracy; %)	E-Commerce (weighted Accuracy; %)
Austria	86.5	79.7	65.1
Bulgaria	84.4	57.0	82.3
Germany (Hesse)	87.5	70.1	79.2
Italy	97.4	90.0	97.4
Lithuania	82.7	77.4	76.8

Funded by the European Union

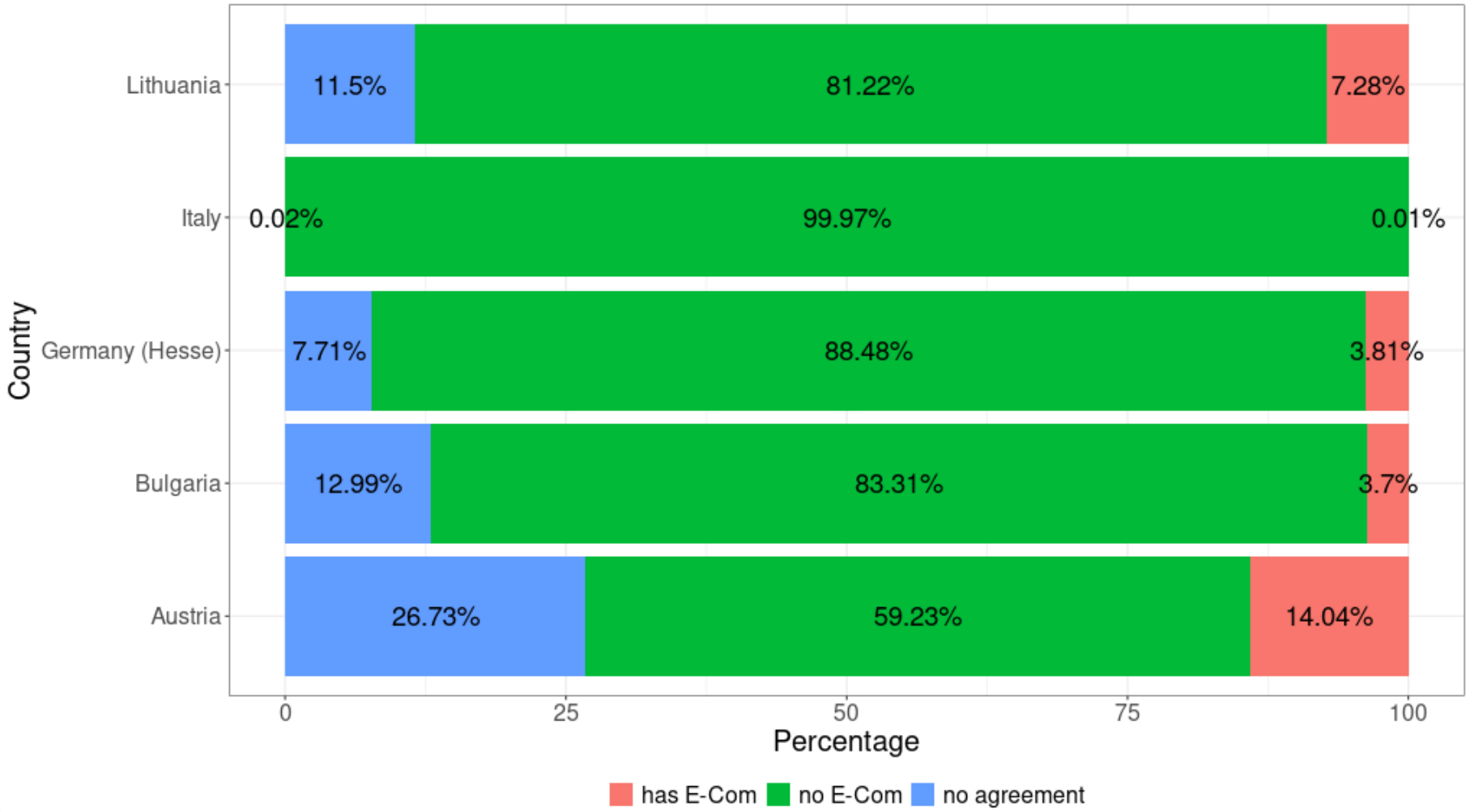
OBEC: Results URL-Linking Accuracy



OBEC: Results SMP



OBEC: E-Commerce



Conclusions

- OJA data's quality with regards of source stability raises concerns
- Accuracy of classifications in OJA data are far from perfect
- One should be very cautious if planning to use such data for official statistics
- In OBEC, URL-linking, SMP and E-Commerce can achieve relatively high accuracy
- Quality of country specific implementations can vary substantially
- Potential to build single sofware package with best practices from each country

Funded by the European Union

Thank you!

Funded by
the European Union