ESSnet Trusted Smart Statistics – Web Intelligence Network
Grant Agreement Number: 101035829 — 2020-PL-SmartStat

WP4: Methodology and Quality

Deliverable 4.8: Quality Assessment for the Statistical Use of Web Scraped Data

FINAL VERSION

Final version, 2024-11-20

Prepared by:
Ville Auno, Statistics Finland
Magdalena Six, Statistics Austria
Johannes Gussenbauer, Statistics Austria
Alexander Kowarik, Statistics Austria

Web Intelligence Network

Funded by the European Union

This document was funded by the European Union.

The content of this deliverable represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

Contents

Web Intelligence Network

Funded by the European Union

1. Introduction

This document showcases the work done in Trusted Smart Statistics – Web Intelligence Network (WIN) project in assessing the quality of web scraped data. Quality assessment is one of the pivotal tasks within the Work package 4 and it is done in close co-operation with Work package 2.

The purpose of quality assessment in general is to provide insight in the quality and usability of web scraped data in official statistics production. The quality assessment in this document focuses on two distinct cases, namely the quality of web scraped data as collected in the WIN project on the topics of Open Job Advertisements (OJA) and Online Based Enterprise Characteristics (OBEC).

The OJA data in the WIN project was assessed in a couple of ways. Firstly, the quality and stability of the sources was assessed with the help of quality indicators that were formulated for the task. Secondly, the accuracy of classification of the OJA data was assessed through annotation exercises. The results for these are presented in this report.

The OBEC data was assessed by drawing a sample of enterprises and manually searching their websites, and identifying if these websites contain links to social media accounts of the enterprise or and whether the enterprise practices e-commerce. The manually obtained results were then compared to the results achieved by automated processes of finding websites of enterprises and scraping them.

This report is structured as follows. The first chapter provides a brief introduction, the second chapter is focused on the quality indicators regarding the OJA data, and the third chapter presents the results from the OJA annotation exercises. The chapter 4 will focus on the quality assessment of OBEC data and chapter 5 concludes.

2. Quality Indicators for OJA Data

Quality indicators discussed in this report are focusing on the quality of the sources (job portals) in OJA data as collected on the WIH within the ESSnet WIN project. The usefulness of the web scraped OJA data in official statistics production is highly dependent on the relevance of the sources and their stability. Therefore, several quality indicators were formulated to assess the quality of the sources. These indicators are:

1. Are the most important portals of your country included?
2. Number of relevant sources over time
3. Presence of very relevant sources over time
4. Ranking of the relevant sources over time
5. Time series plot for number of OJAs for all very relevant sources
6. Stability of data over different versions of data.

These indicators were computed centrally in a standardized way with the help of an RMarkdown script executed on the WIH for the following 10 countries: Austria, Finland, Germany, France, Italy, Poland, Netherlands, Bulgaria, Portugal and Romania.

All country reports as output of the country-specific RMarkdown scripts are available in html format on the ESSnet-WIN internal website:

https://webgate.ec.europa.eu/fpfis/wikis/display/WIN/Quality+assessment+of+OJA+sources

The R code is available in a Gitlab repository with constrained access rights[1]. All WIN members can request access to this repository. Thereby, the output can easily be reproduced also for countries not included so far.

2.1 The most important sources

In order to compute the first quality indicator, domain knowledge is required. The idea is to identify the five most important job portals in a particular country. Subsequently, the presence of those portals in OJA data should be checked:

- Are they included in the list of relevant sources?
- Are they included in the data across all years?

The first quality indicator should therefore be computed by experts from each country with the help of domain experts if needed. As mentioned above, the quality indicators were calculated centrally, based only on the available data. So far, no country experts were involved. Thus, this qualitative indicator could not be completed so far. In addition, the sources are only identified with the help of numerical source_ids in the generally available data sets; no source names are given. The lacking availability of source names makes it barely possible for country experts to compare the relevant sources in their countries with the

[1] The link to the Github Repository: https://git.fpfis.tech.ec.europa.eu/estat/wihp/analysis/oja_sample_annotation/-/tree/develop/quality%20indicators%20OJA?ref_type=heads

sources in the data sets. The source names are available in data sets with special access rights. It is work in progress to gain access also to the source names.

Due to these reasons, results for this indicator cannot be presented in this report.

2.2 Number of relevant sources over time

Relevant sources in the OJA data are those that have a significant number of job advertisements available. With the current setup, it is very difficult to say anything about the statistical relevance of an OJA. Therefore, the relevance in this report is defined by volume only. A source was defined to be *relevant* if it had more than 500 OJAs during a year. A source was defined to be *very relevant* if it had more than 5000 OJAs during a year.

Since the countries differ vastly in size and in the total number of OJAs, 'a relevant source' could be defined differently in different countries. This could be achieved, for instance, by some kind of cut-off analysis on the sources. For now, this indicator was computed using the same thresholds for each country, and therefore the indicator do not give a lot of insight for comparing different countries. However, the indicator provides a good picture on the stability of the number of relevant sources over time within each country.

Table 1 shows the number of relevant sources in different countries. As it can be seen, the numbers remain fairly stable over the years, but some countries have larger fluctuations than others. For example, the number of relevant sources for Portugal varies from 5 in 2018 and 51 in 2022, whereas Poland has rather stable number of 12 to 15 sources over the years.

Table 1: The number of relevant sources (more than 500 OJAs) in different countries across the years.

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

When examining the number of very relevant sources (more than 5000 OJAs), the situation is quite similar. The number of very relevant sources is presented in Table 2. The variability is again the most pronounced in Portugal, ranging between 2 and 16 sources. In countries like Finland and Bulgaria, the number fluctuates less in absolute terms, between 1 and 5 for Bulgaria and 2 and 6 for Finland. However, these fluctuations are significant in relative terms.

Table 2: The number of very relevant sources (more than 5000 OJAs) in different countries across the years.

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

2.3 Presence of very relevant sources over time

For the quality indicator 3, the presence of very relevant sources was studied across different years. The goal was to verify whether the very relevant sources in one year were present in the data in different years. The indicator was computed so that the presence of very relevant sources from 2018 was checked for years 2021 and 2023. Similarly, the presence of very relevant sources from 2021 was checked for the year 2023. Finally, the presence of very relevant sources from 2023 was checked for the year 2018.

The results indicate that the presence of very relevant sources varies significantly over the years. For instance, in Germany, out of 27 very relevant sources identified in 2018, only 17 were present in the data in 2021. This number was further decreased to 10 in 2023. From 28 very relevant sources in 2021, 20 were present in the data in 2023. Conversely, out of 21 very relevant sources in 2023, only 14 were found in the data in 2021.

2.4 Ranking of the relevant sources over time

The fourth quality indicator offers insights into the changes in the ranking of relevant sources. Rather than focusing on the quantity of relevant sources, this indicator tracks how the importance of specific sources evolves relative to other relevant sources.

Table 3 shows the ranking of relevant sources for Austria between 2018 and 2023. It can be seen that the top 5 sources remain quite the same over the years although their respective order changes from year to year. However, it can also be seen that, for example, the number one source (source_id 633) in 2021-2023 did not exist as a relevant source in 2018, and the most important source of 2020 (source_id 108) vanishes for the years 2021-2023. In addition, the third source in 2018 disappears from the group of relevant sources after 2019.

Table 3: Ranking of the sources (source_ids) in Austria between 2018 and 2023.

Rank	2018	2019	2020	2021	2022	2023
1	110	110	108	633	633	633
2	108	108	216	216	216	216
3	564	410	633	113	110	110
4	216	216	110	110	113	113
5	113	564	113	116	855	350
6	410	113	116	734	350	956
7	543	543	187	194	194	716
8	115	116	195	195	716	855
9	427	195	543	427	843	197
10	189	187	212	350	856	843
11	187	427	427	716	197	194
12	332	332	332	753	825	635
13	116	633	500	550	783	106
14	212	192	716	856	753	825
15	195	212	635	843	635	866
16	192	500	194	783	638	-
17	545	191	191	197	106	-
18	500	206	107	855	427	-

19	560	-	169	635	555	-
20	-	-	200	200	190	-
21	-	-	193	825	-	-
22	-	-	-	638	-	-
23	-	-	-	500	-	-
24	-	-	-	106	-	-
25	-	-	-	168	-	-
26	-	-	-	555	-	-
27	-	-	-	187	-	-

The ranking of the relevant sources provides a good image on the dynamics of the sources. Relevant sources disappear, new ones emerge and the relative ranking of existing relevant sources change from year to year. Such dynamics may raise concerns about the stability of the sources.

2.5 Timeseries plots for very relevant sources

Another way to assess the stability of sources is through graphical analysis. For the fifth quality indicator, the monthly number of OJAs per very relevant source is plotted as time series. The construction of the time series was done in a completely standardized way for each country and is published in the country reports. The country-specific plots allow a quick visual assessment of how the ranking of very relevant sources has changed and how stable these sources are over the years.

Figure 1 shows the time series plot for Finland. It is easy to see that the most relevant source changes almost every year. The number of OJAs per source also seems to fluctuate quite a lot from month to month. In addition, the most relevant source during the year 2019 suddenly drops to zero in the beginning of the year 2020.

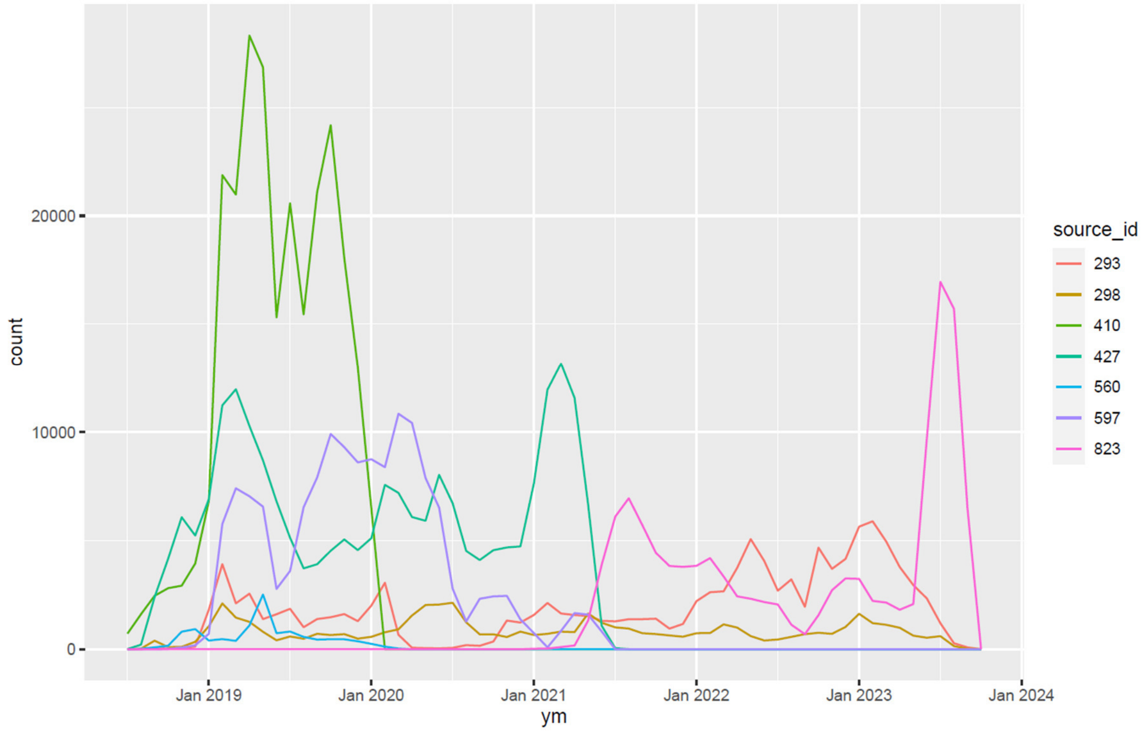

Figure 1: Number of OJAs per month for very relevant sources in Finland.

The situation is similar in other countries as well. Figure 2 shows the time series plot for France. In France, the two large spikes immediately attract attention. Such massive fluctuations suggest instability in the sources. However, in France's case, a couple of sources seem to be rather stable during the years 2021 – 2023.

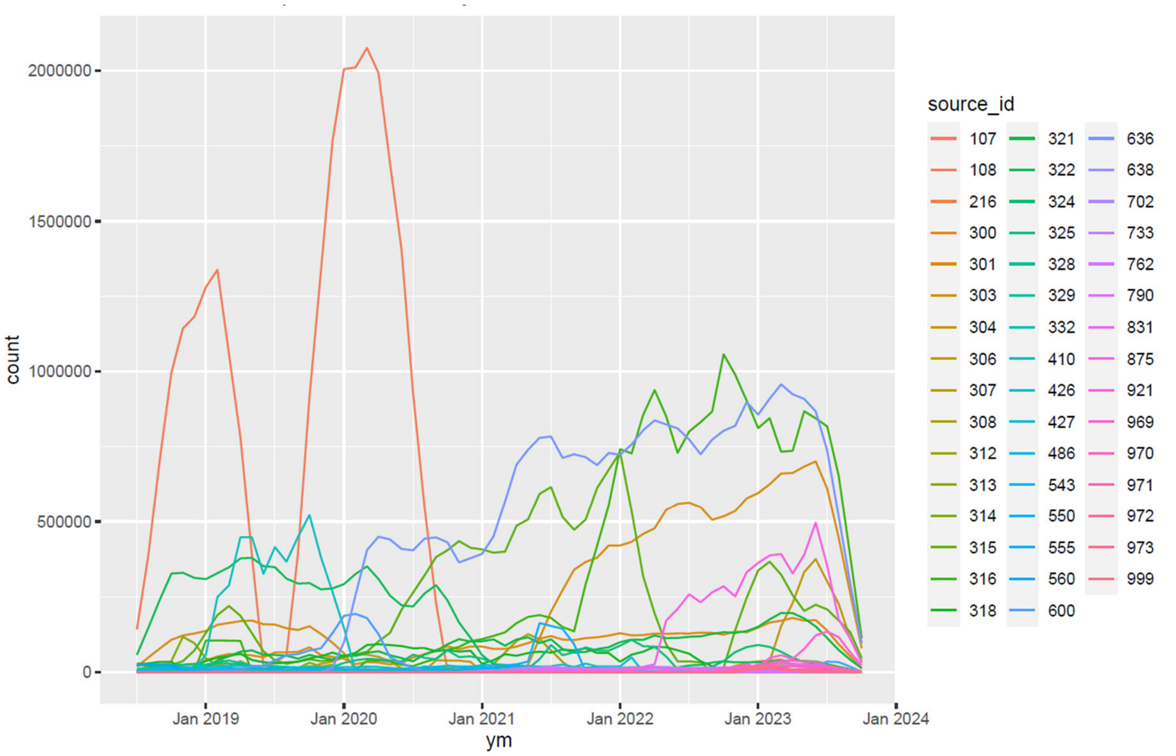

Figure 2: Number of OJAs per month for very relevant sources in France.

Additionally, it is surprising that one cannot detect any decrease or minimum in the time series of OJAs due to the COVID-19 lockdown as it can clearly be seen in European-wide JVS statistics[2]. This observation is not only true for the examples of France and Finland, but holds also for the other countries.

2.6 Stability over different versions of the data

Finally, the sixth quality indicator examines the stability of sources between different versions of OJA data. It is crucial for any analysis that the data does not change for former years when new version of data is published. Of course, there might be occasions when it is necessary to revise the data, but in these cases the revisions should be announced and explained for the users.

For this indicator, two different versions of OJA data were used. The first data set was the one used for all the other quality indicators, where we queried the latest available data on the WIH, with November 2023 as date of execution ("WIHAccessCatalog.wih_oja_latest.wih_oja_blended"). We compared this latest data with data queried from a versioned data set from Q1 2023 ("WIHAccessCatalog.wih_oja_versioned.wih_oja_blended_v2_2023q1_r20230511").

[2] The European-wide job vacancy rate shows a local minimum in the third quartal of 2020, see the Publication about the Job Vacancy Statistics of the European Central Bank: https://shorturl.at/dvBDJ

The number of OJAs per year for each relevant source was counted from both of those versions. Finally, absolute and percentage differences between these two counts were computed. The results raise concerns about the stability of the sources over different versions of the data.
It is important to note that the resulting indicators might vary if we compared different versioned data sets.

Table 4: Averages of absolute and percentage differences of number of OJAs for relevant sources between different versions of the data in Netherlands.

Year	Avg. Absolute difference	Avg. Percentage difference
2018	714	0.61
2019	3271	2.69
2020	1979	2.46
2021	2715	3.35
2022	2115	1.36
2023	34501	159.00

Table 4 shows the average differences between the number of OJAs in different versions of the data for the Netherlands. The absolute and percentage differences are not very big, except for the year 2023, considering that the total number of OJAs for each year is well above 1 million. The huge difference in the year 2023 results from the fact that the considered versioned data set ended already in Q1 2023 However, even if the average differences are low, the numbers can be quite significant for individual sources. For individual sources, the percentage difference between sources was found to be as high as 31 per cent. In absolute terms, the highest difference for an individual source was 45 564 OJAs.

Table 5: Averages of absolute and percentage differences of number of OJAs for relevant sources between different versions of the data in Romania.

Year	Avg. Absolute difference	Avg. Percentage difference
2018	16	1.00
2019	182	2.33
2020	23	1.13
2021	31	0.43
2022	11	0.32
2023	10924	134.27

Table 5 shows the same comparisons for Romania. Absolute differences are small due to significantly lower total number of OJAs when compared to Netherlands, but the percentage differences are of same magnitude or slightly less. In Romania, the highest percentage difference for an individual source was 21 per cent and the highest absolute difference 3 364 OJAs.

It is important to note that for many of the individual sources the number of OJAs remains the same between the data versions. The differences and especially large differences come only from handful of sources.

3. OJA annotation exercises

The web scraped Online Job Advertisement data needs to be classified in order to use it in production of statistics. The key classification is the International Standard Classification of Occupations (ISCO) which organizes the job advertisements into clearly defined groups based on the tasks and duties that are part of the job description.

The classification of the data is done automatically by using machine learning techniques. However, the accuracy of the classification algorithm needs to be assessed carefully. This can be done by annotation exercises, where a sample is drawn from the classified data and the classification is then checked by humans.

3.1 The first annotation exercise

In the first annotation exercise, the correctness of the ISCO classification was evaluated by 8 countries: Austria, Bulgaria, Italy, Lithuania, Poland, Portugal, Romania and Slovenia. The annotation was done for different levels of the ISCO classification: 1-digit, 2-digit, 3-digit and 4-digit levels. The annotators had several values they could assign to each job advertisement: Correct, Incorrect, Not a job ad, Impossible to classify, No reference to occupation, Job description missing and Wrong language. Each country annotated 350-400 job advertisements.

Table 6: Average percentage shares of each class assigned by the annotators in all countries.

Classification	1digit(s)	2digit(s)	3digit(s)	4digit(s)
Correct	62.04	54.83	51.40	47.81
Incorrect	30.77	37.59	40.52	43.53
Incorrect - Missing label	1.16	1.26	1.47	1.73
Impossible to classify	0.74	0.74	0.89	1.61
No reference to occupation/Job description missing	4.40	4.69	4.83	4.80
Not a job ad	1.28	1.28	1.28	1.28
Wrong language	1.92	1.64	1.75	1.75

The average percentage shares of each class assigned by the annotators are presented in Table 6. As expected, the accuracy of the classification declines when moving from 1-digit level to 4-digit level. In the annotated job advertisements, accurate classifications were achieved in 62.04% of cases when using the 1-digit ISCO classification. However, this accuracy dropped to 47.81% for the more detailed 4-digit ISCO level.

Country specific shares of correct labels for different ISCO levels can be seen from Figures 3 and 4.

Funded by the European Union

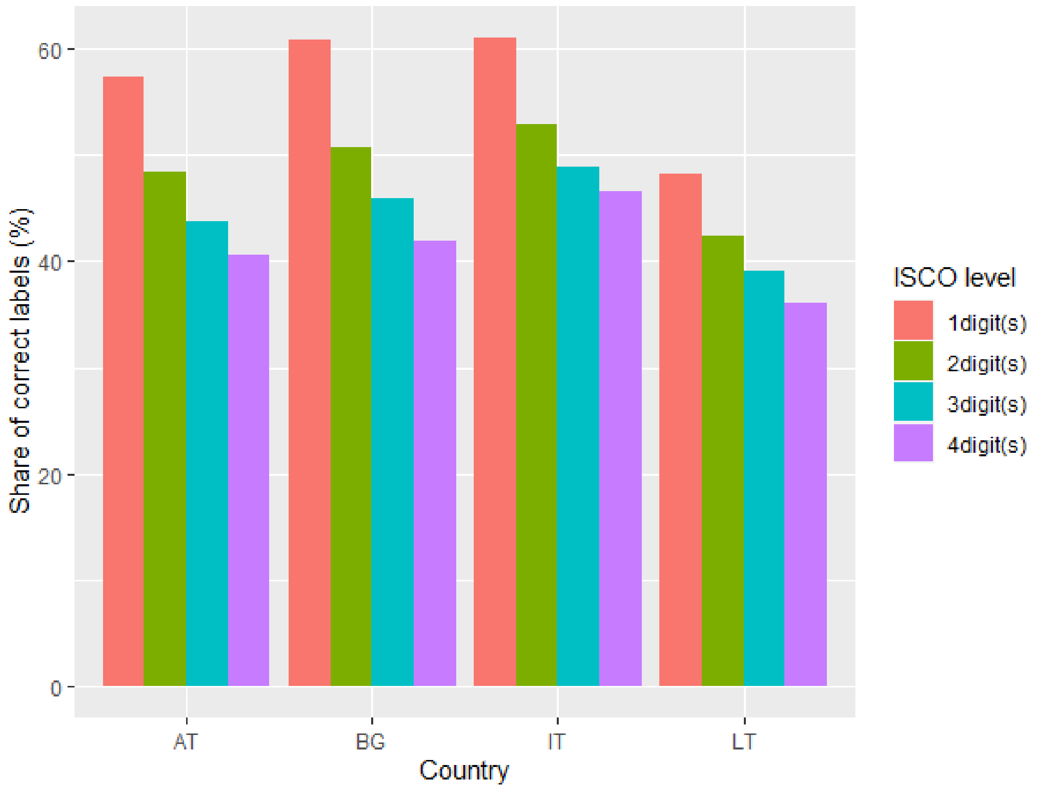

Figure 3: Percentage shares of correct labels for different ISCO levels for Austria, Bulgaria, Italy and Lithuania.

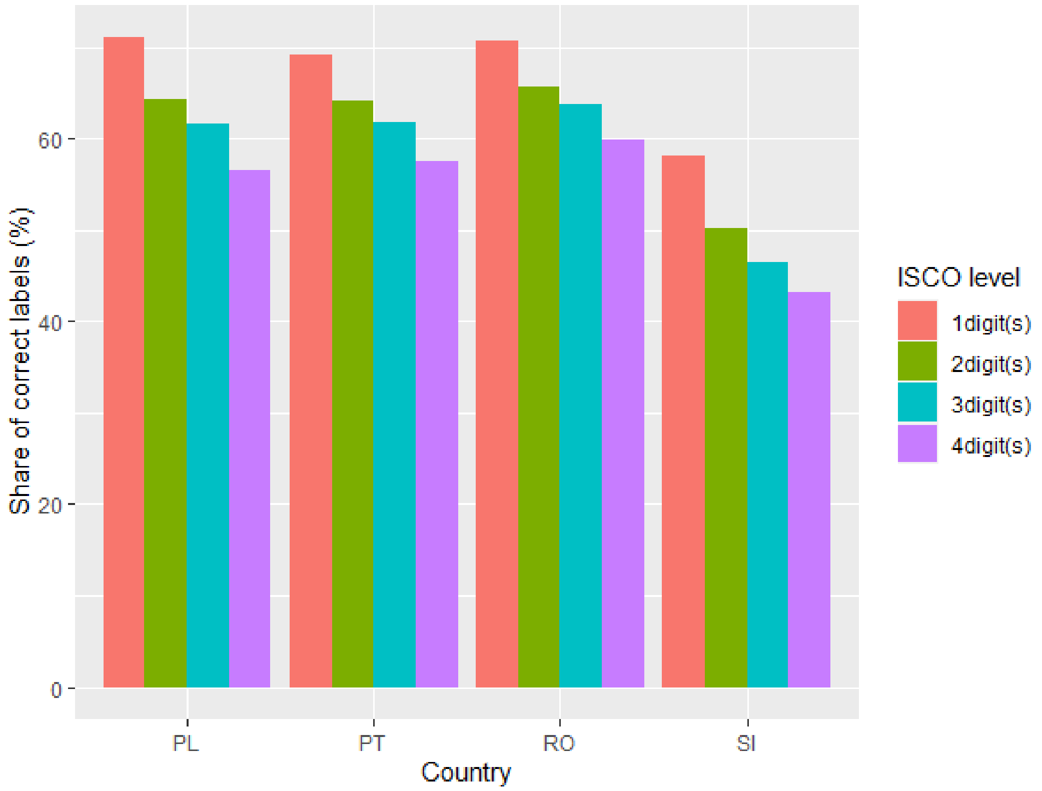

Figure 4: Percentage shares of correct labels for different ISCO levels for Poland, Portugal, Romania and Slovenia.

The share of correct labels are following a similar pattern in all eight countries, as can be seen from the Figures 3 and 4 above. The highest accuracy for 1-digit ISCO level was achieved in Poland with 71.07 % share of correct labels. The accuracy was lowest in Lithuania with only 48.28 % share of correct labels at 1-digit ISCO level. The fall in the accuracy is the greatest when moving from 1-digit level to 2-digit level. For example, in Bulgaria, the accuracy fell more than 10 percentage points.

When looking at the more specific 4-digit level, the highest accuracy was reached in Romania with 59.79 % share of correct labels. Again, the accuracy was lowest in Lithuania with 36.15 % share of correct labels.

This level of accuracy is insufficient for OJA data to be used in statistics production. It is clear that improvements need to be made to the classification algorithm. In order to evaluate whether any improvement has happened in the classification, a second annotation exercise was held later.

3.2 The second annotation exercise

The second annotation exercise expanded the evaluation of classifications in the OJA data. This time, the annotation encompassed all classifications: occupation, economic activity, education, location, and working time. The exercise was conducted in eight countries: Austria, Bulgaria, Finland, France, Germany, Italy, Poland, and Slovenia. The annotated sample size was approximately 300 OJAs per country, with a total of 2,742 OJAs annotated (the Polish sample was annotated twice).

Table 7: Mean shares and weighted shares of correct labels in each classification across all countries.

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

Table 7 shows the mean shares of correct labels for all annotated classifications. The sample estimates were also weighted to match the population. Economic activity and occupation were annotated at the 1-digit level of the classification (NACE and ISCO, respectively). Location was evaluated at the NUTS1 or NUTS2 level, depending on the country. Classification accuracy is very poor for economic activity and education (25.17% and 15.54%, respectively), while occupation, location, and working time showed slightly better accuracy. However, even the highest accuracies among the different classifications are far from perfect.

Country specific accuracies, both unweighted and weighted are presented in the following figures. Economic activity on Figure 5, education on Figure 6, occupation on Figure 7, location on Figure 8 and finally working time on Figure 9.

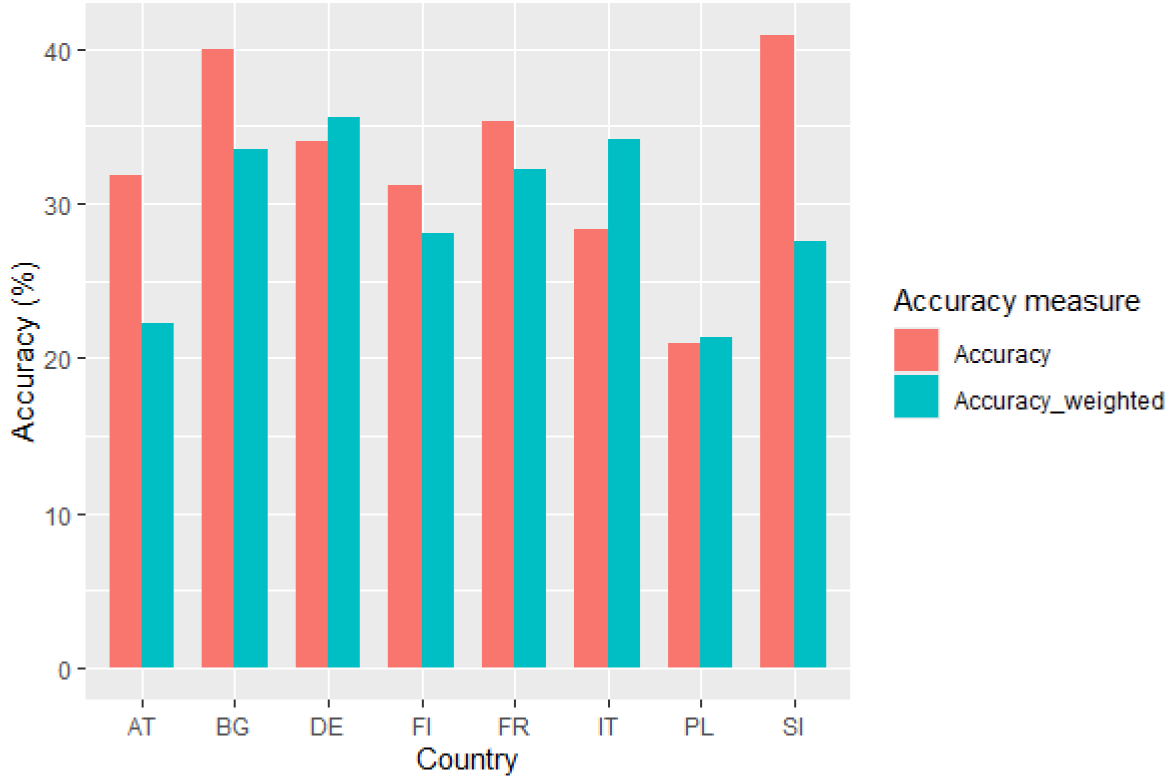

Figure 5: Classification accuracy and weighted accuracy for economic activity in each country.

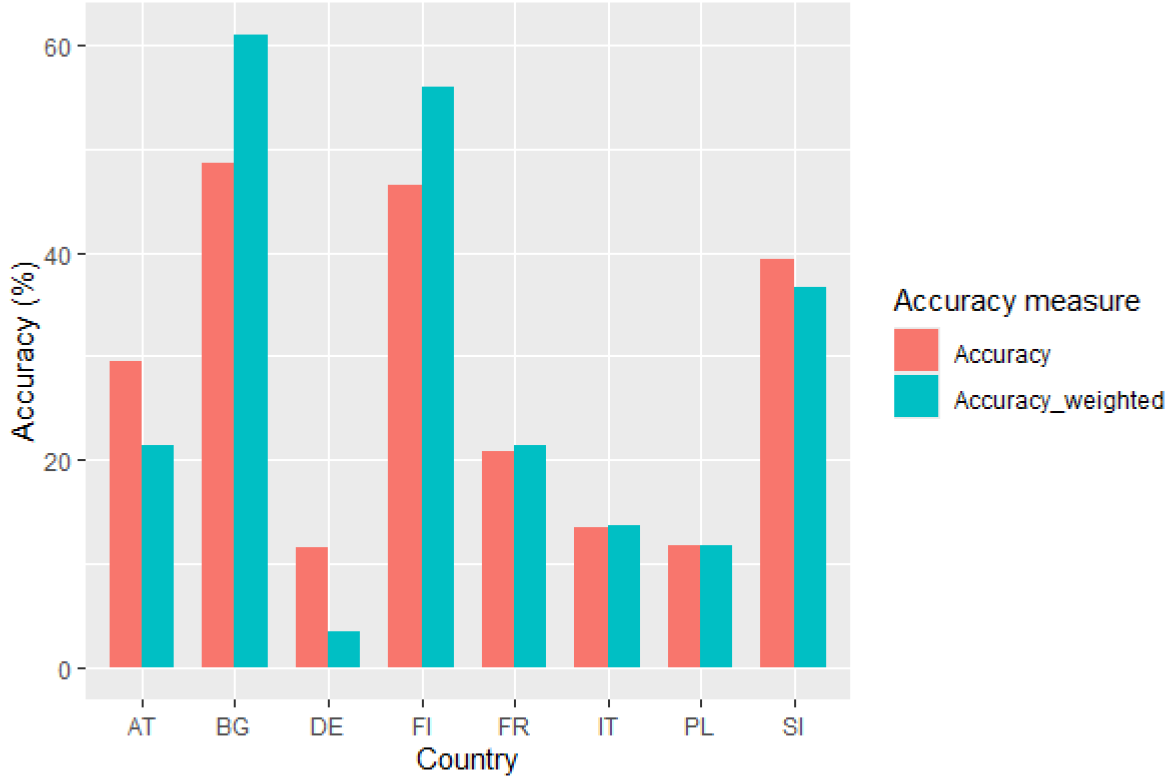

Figure 6: Classification accuracy and weighted accuracy for education in each country.

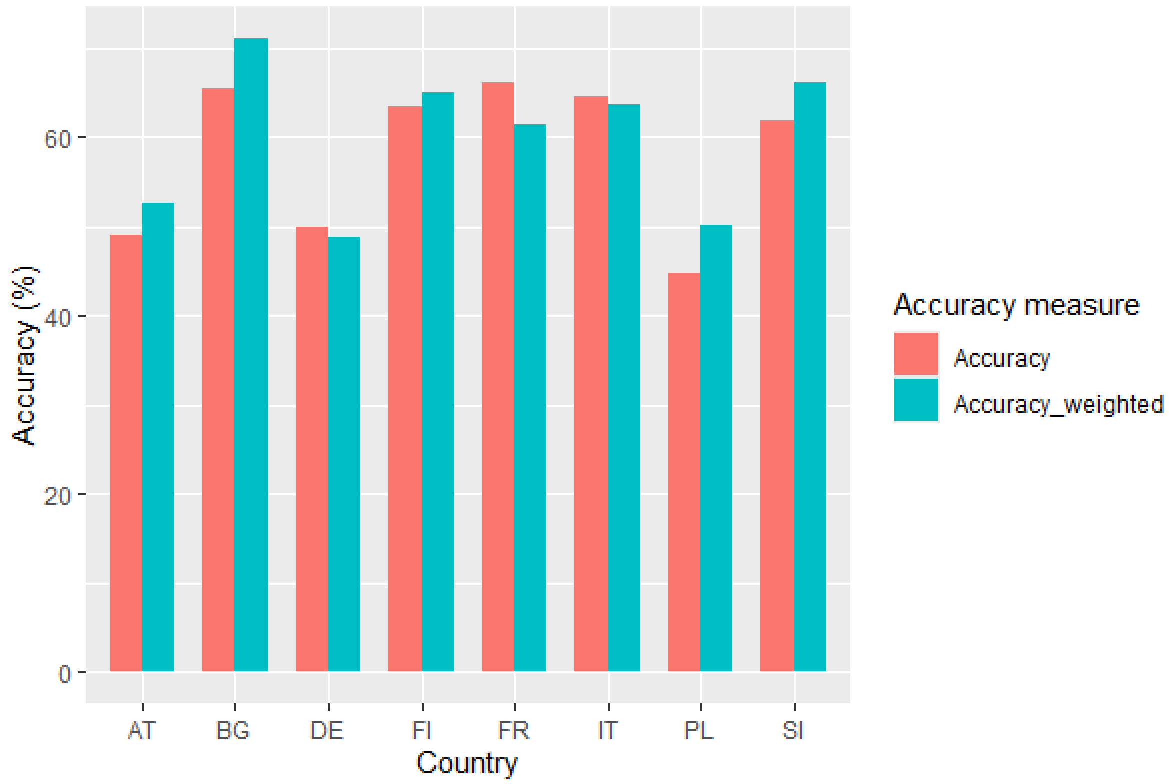

Figure 7: Classification accuracy and weighted accuracy for occupation in each country.

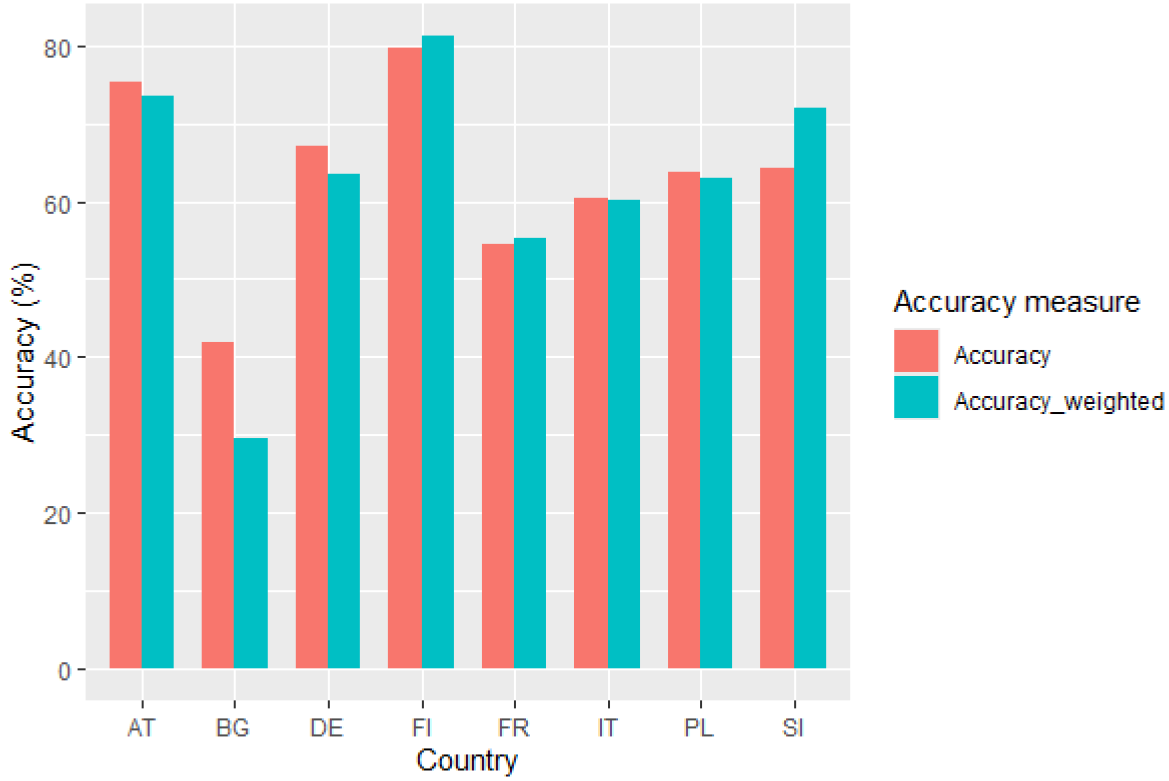

Figure 8: Classification accuracy and weighted accuracy for location in each country.

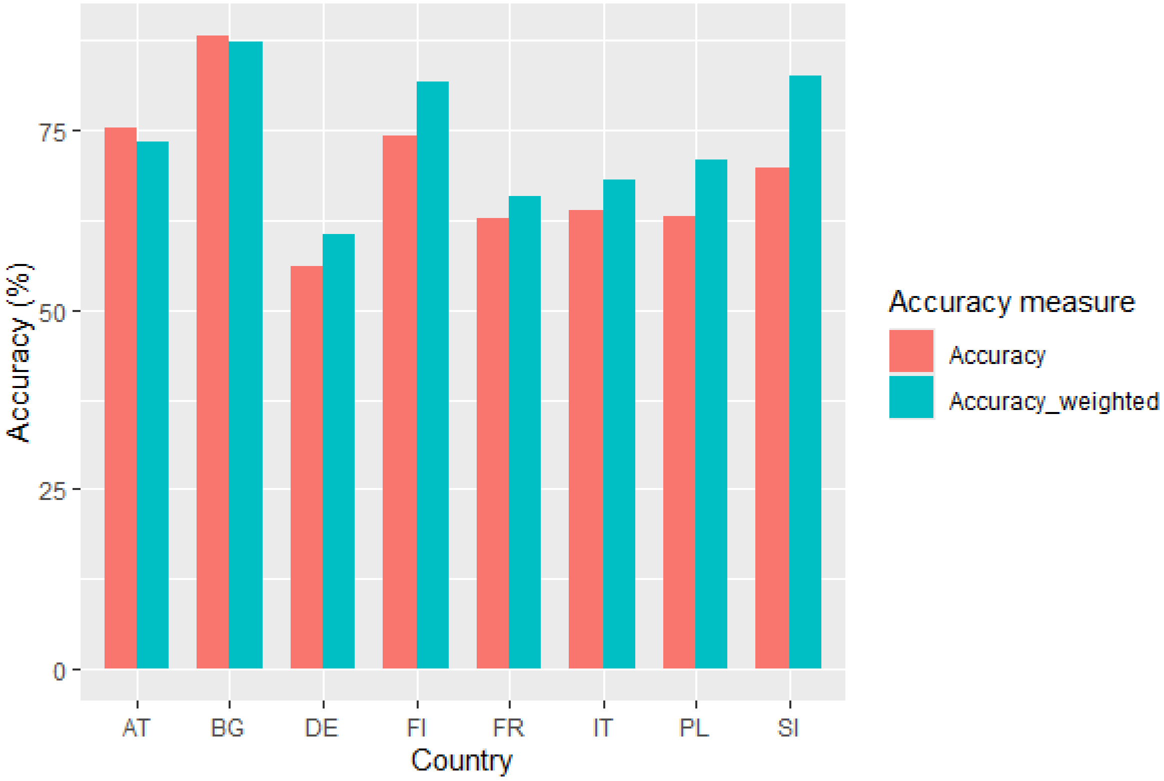

Figure 9: Classification accuracy and weighted accuracy for working time in each country.

One concerning finding from the second annotation exercise was that the accuracy of occupation on 1-digit ISCO classification is actually worse than in the first annotation exercise. In the first annotation exercise, the 1-digit ISCO code was correct in 62.04 % of the cases whereas in the second annotation, the corresponding share was only 56.23 %. Due to the small size of the annotated sample, both for the first and second round, and different sampling design in each round some differences are to be expected.

4. Online Based Enterprise Characteristics

The purpose of the quality assessment of online based enterprise characteristics data was to evaluate how well URL finding and fetching information such as social media presence or e-commerce could be done automatically. In addition, many countries have their own processes regarding URL linking, social media or e-commerce extraction and this exercise aims to ultimately compare the quality of these processes. The assessment was performed by six countries: Austria, Germany, Poland, Lithuania, Italy and Bulgaria.

The quality assessment task began with annotating a sample of 500 legal units, drawn from the statistical business register given a sampling design stratified by NACE section and size class. The sampling population was defined close to the ICT survey population[3]. Annotators manually searched URLs for each legal unit in the sample and checked the websites for social media presence and e-commerce.

The second phase involved using national URL finders or other sources that countries currently use to obtain URLs. These automatically obtained URLs were then compared with the “ground truth” in the annotated sample.

The third phase was to use national web scrapers or the scraper on the Web Intelligence Hub. The scraping results for social media presence and e-commerce were again compared with the manually annotated sample. The comparison consisted of three different measures: one for URL linkage accuracy, one for social media presence on enterprise websites, and one for e-commerce presence on enterprise websites.

The URL linkage was considered correct if the enterprise was found to have one or multiple websites in the manual annotation, and the URL identified through automated URL finding matched to only one or more the manually found URLs. The linkage was also deemed correct if no website was found in both the manual annotation and the automated URL finding. URL linkage accuracies for different countries are shown in the Table 8.

Table 8: Percentage of correct matches between URLs found by manual annotation versus URLs found by an automated URL finding in each country.

Country	Accuracy (%)	Weighted Accuracy (%)
Austria	87.4	86.5
Bulgaria	83.6	84.4
Germany (Hesse)	87.8	87.5
Italy	89.6	97.4
Lithuania	82.7	82.7

The accuracy seems to be similar and reasonably high in each country. With the highest accuracy being 89.6 % in Italy and the lowest 82.7 % in Lithuania.

Social media presence on the enterprise website was evaluate by contingency tables between the manual annotation and web scraping. Contingency tables provide information on how many cases the manual annotation and web scraping agree.

[3] In some participating countries the ICT survey population consists of statistical units and in others of legal units. A compromise was reached to select only legal units which are part of the ICT survey population with 10 or more employed persons and part of NACE sections C-D, L-M and S.

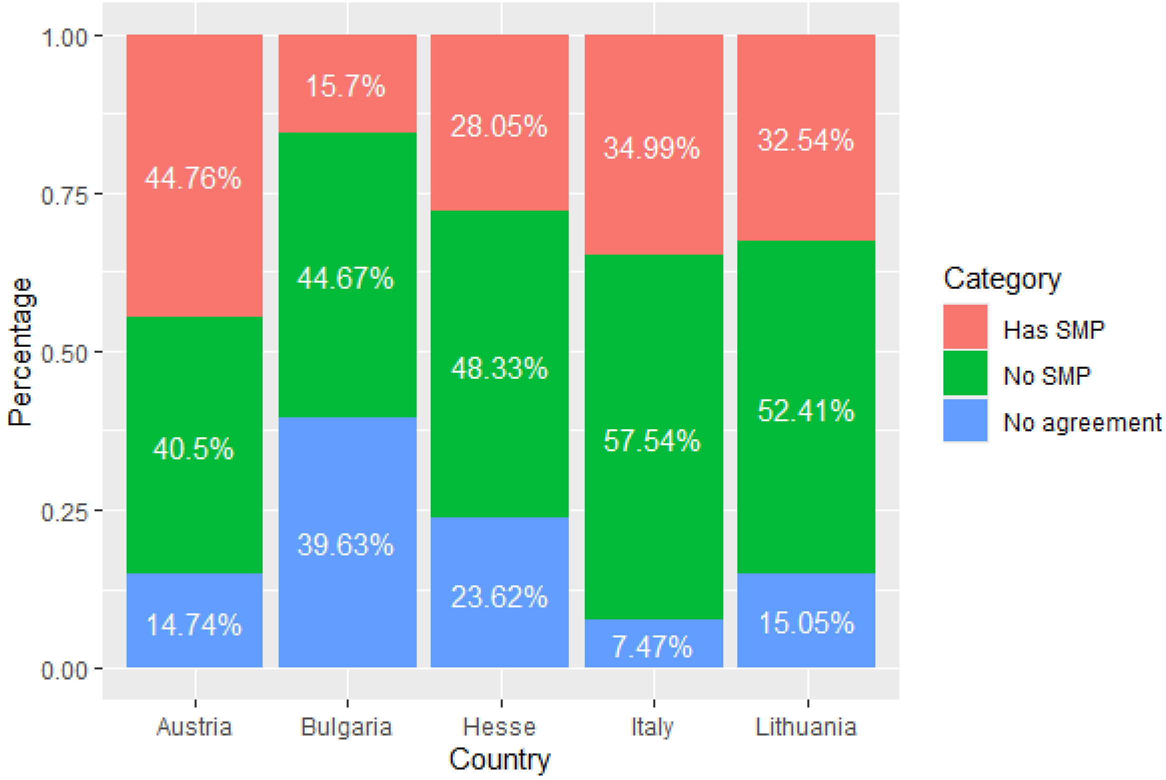

Figure 10: The percentage shares of cases where the manual annotation and web scraping agree on an enterprise having SMP, not having SMP and share of cases with no agreement.

The percentage shares of cases where manual annotation and web scraping agree on an enterprise having or not having social media presence on their website can be seen in Figure 10. It is important to note that these percentage shares only represent the agreement on the SMP. Hence, there are cases where automated process may not agree with manual annotation on the actual URL. The percentage shares of total cases where the automated process and manual annotation agree on both the URL and SMP can be seen in Figure 11.

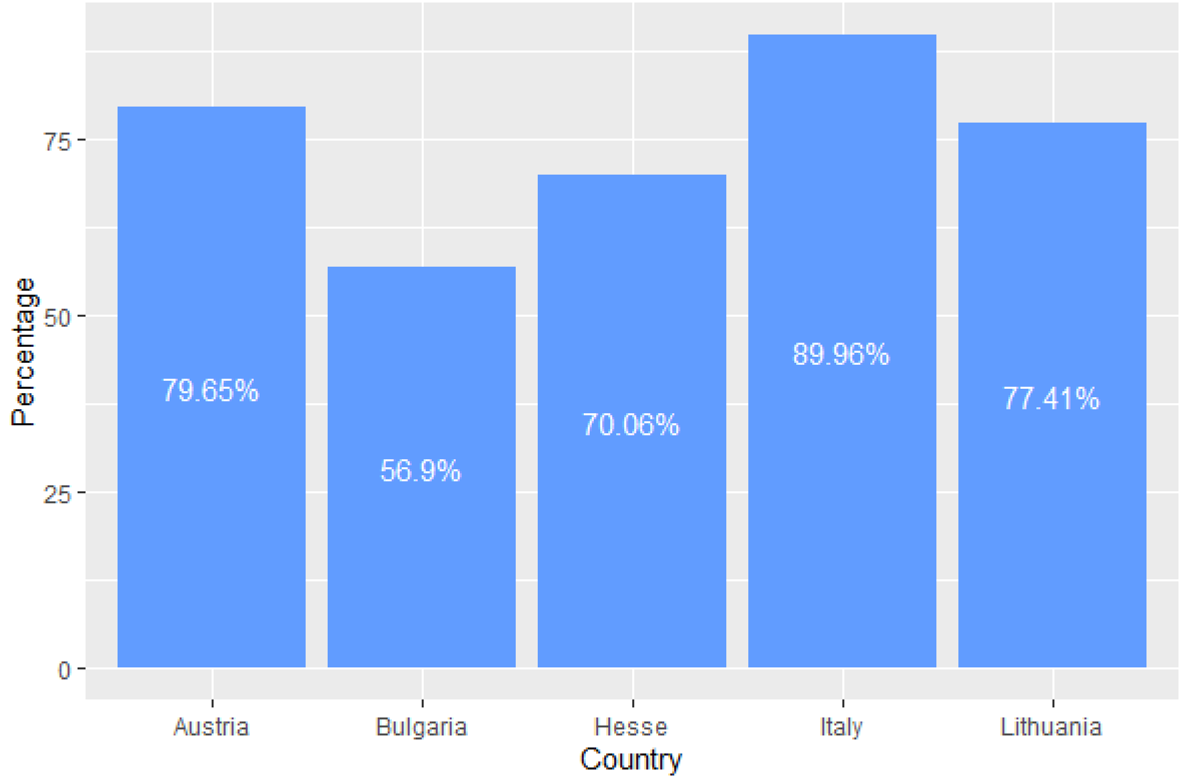

Figure 11: Total percentage shares of agreement on both URL and SMP presence between the manual annotation and web scraping in each country.

Relatively high agreement shares on SMP would indicate that scraping the information could be a viable option to extract the information. However, from the Figure 11, it can be seen that once combined with the URL finding procedure, the share will decrease a few percentage points. Nevertheless, in countries such as Italy, the percentage shares are still high (89.96 %).

The presence of e-commerce on the enterprises' websites was analyzed in same way as the social media presence. The percentage shares of cases where manual annotation and web scraping agree on an enterprise having or not having e-commerce on their website can be seen in Figure 12.

Similarily, when combined with URL finding accuracy, the agreement of the manual annotation and the automated process slightly decrease as is sown in the Figure 13. All in all the presence of e-commerce on an enterprise's website was found more accurately automatically compared to the social media presence.

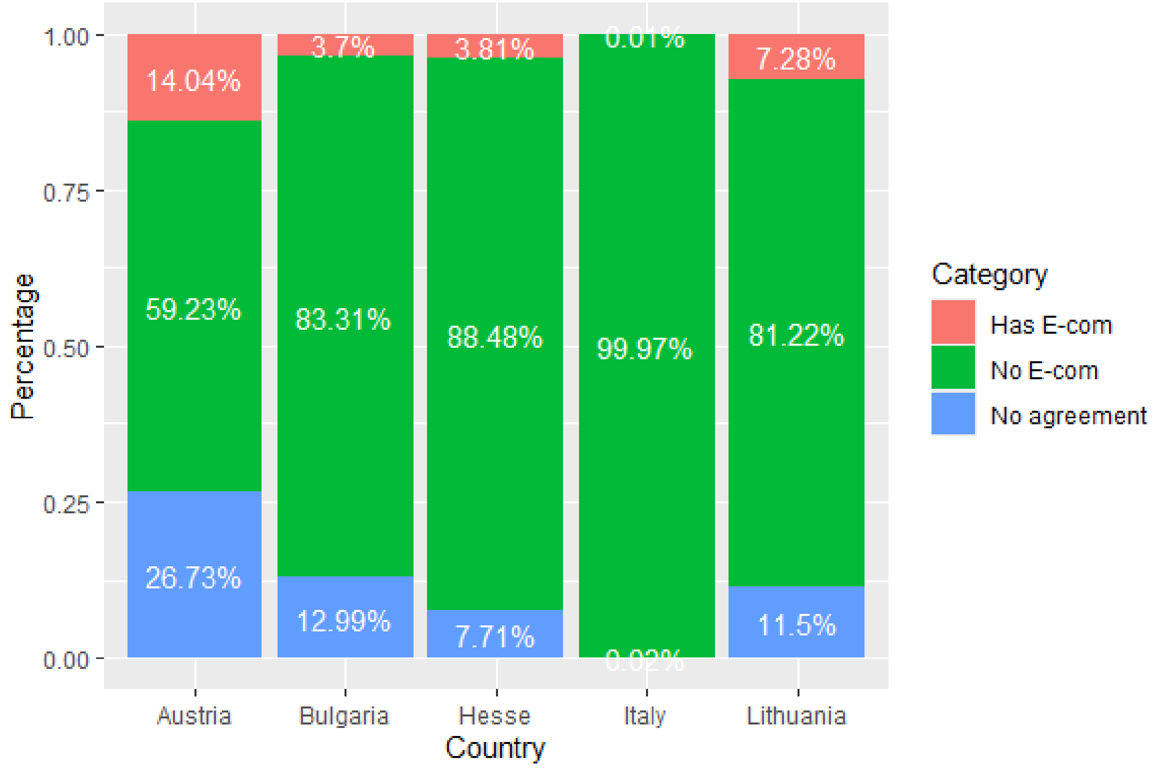

Figure 12: The percentage shares of cases where the manual annotation and web scraping agree on an enterprise having e-com, not having e-com and share of cases with no agreement.

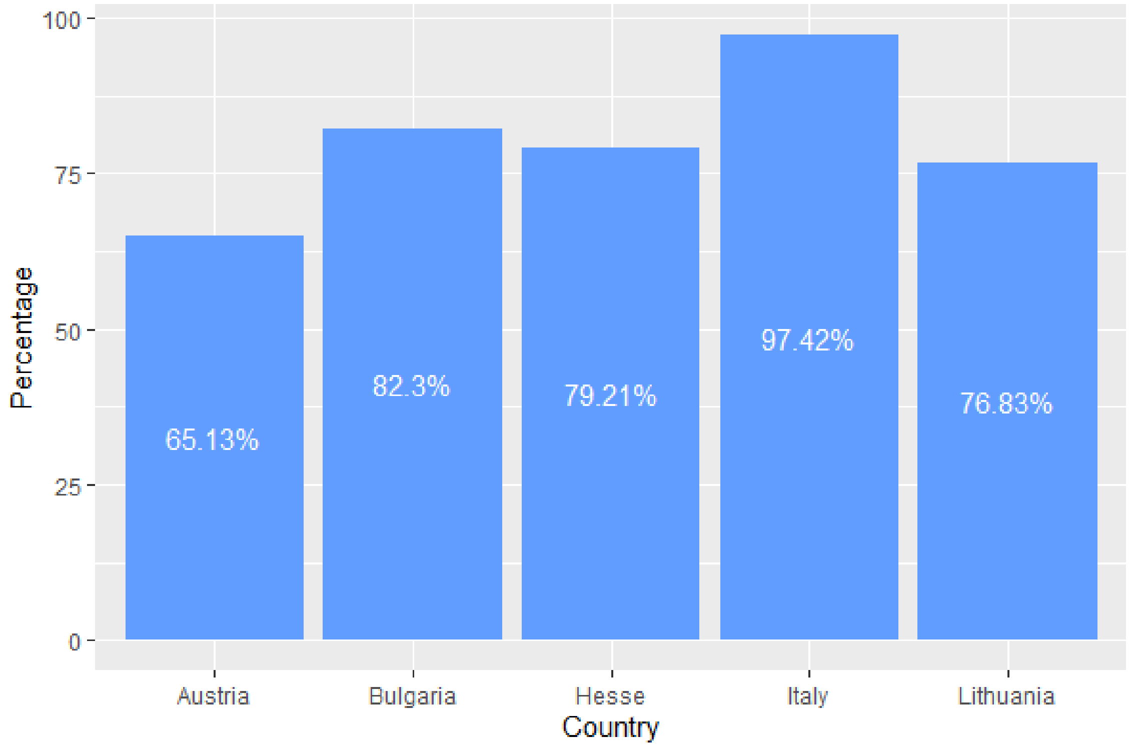

Figure 13: Total percentage shares of agreement on both URL and e-com presence between the manual annotation and web scraping in each country.

These results imply that while some countries may have really well functioning URL finders and web scrapers, in other countries the reality is different. Therefore, caution is needed when automatically acquired information on company websites, social media presence or presence of e-commerce is to be used in official registers and statistics. While the findings certainly underscore the potential of automation, it is recommended to be cautious and refine the processes even further to ensure the data realiability. This is true especially with diverse national systems.

5. Conclusions

This report has been dedicated to assessing the quality of web-scraped data within the Trusted Smart Statistics – Web Intelligence Network project. The quality assessment plays a crucial role in understanding and improving the reliability of web-scraped data when considering its use in statistics production.

In this report, the stability of sources in OJA data has been assessed through several quality indicators. In addition, the quality of classification of OJAs has been assessed through annotation exercises. The key findings are that the stability of the sources in OJA data presents a cause of concern, the lack of source names makes it almost impossible for countries to evaluate their relevance and that the accuracy in classifying the data into ISCO classification is insufficient. The accuracy of other classifications in the OJA data were also evaluated in the second annotation exercise. The accuracies of economic activity, education, location and working time were far from perfect.

The quality assessment of OBEC data demonstrated that automated URL finding, social media presence detection, and e-commerce identification can achieve relatively high accuracy, with results comparable to manual annotation in some cases. However, the results suggest high variability in effectiveness of national URL finders and web scrapers. This leads to a conclusion that automated processes alone may not consistently meet the standards required for official statistical purposes, especially in cases where exact data is crucial.

These findings highlight the fact that more work is needed to improve the quality of OJA data if it is to be used in official statistics production. The same can be said about the OBEC data. While automatic processes for finding enterprise URLs and scraping information such as social media presence or e-commerce can be a viable approach, caution is needed to evaluate the accuracy and reliability of such automatic processes. Additionally, this report serves to emphasize the importance of quality assessment in the adoption of new alternative data sources for statistics production.