Webscraped data for replacing and validating survey questions

Alexander Kowarik

Johannes Gussenbauer

Lukas Mikesa

Marlene Weinauer

Jakob Peterbauer

Wolfgang Rannetbauer

Webscraping can reduce burden and costs in producing business statistics in comparison to the classical survey. Within the ESSNet Big Data II WPC national statistical institutes evaluate the detection of online based enterprise characteristics, as the presence of a website, a link to social media site or an online shop. In this paper the methodology developed and applied in Statistics Austria, as well as first comparison to the Austrian results of the European survey on ICT usage and e-commerce in enterprises are presented.

Introduction

Annually, the European surveys on ICT statistics and e-commerce in enterprises collect business characteristics concerning information and communication technology (Eurostat 2020). A long tradition is the collection of enterprises's web presence: Since 2002, enterprises are asked about having a website, since 2008 the question is extended about specific website's characteristics. The collected characteristics have developed within the last years, the existence of an online shop and a link to social media maintain to be key indicators.

With the vastly developing field of ICT statistics, new indicators, as artificial intelligence, robotics or 3d printing are incorporated in the model questionnaire. On the same hand, traditional items, like web presence, remain to be important. For the reduction of response burden and cost, web characteristics are collected only biannually or new items are not included as trade off.

To overcome these limitations, alternative data sources are explored: For information about enterprises' web presence, the internet itself may be a promising data source. Webscraping of enterprises' characteristics could reduce response burden and cost. Besides webscraping may enable a delivery of figures in closer time intervals and for a wider population: E.g. the restriction of the population to enterprises with at least 10 employees and to specific NACE sectors, as currently defined for ICT surveys, could be extended.

Statistics Austria is member in the ESSNet Big Data II WPC, in which national statistical offices explore the potential of web scraping to produce enterprises' web presence (ESS 2020). Two years of ongoing research, found webscraping to have the potential to mitigate above-mentioned shortcomings, but revealed new problem fields.

In this paper the methodology developed and applied in Statistics Austria, as well as first comparison to the Austrian results of the European survey on ICT usage and e-commerce in enterprises are presented. In the following section main concepts, when using webscrapped data, are described.Then methodology to derive following indicators

- existence of website
- existence of link to enterprise's social media profiles on their website
- existence of online ordering or reservation or booking (e.g. shopping chart) on website

via webscraping is described, respectively. Results of methods are documented with comparisons to survey results of ICT survey 2019.

Main Concepts and Data

In the following section main concepts are introduced, data used is declared.

OBEC

Online based enterprise characteristics (OBEC) is referred to any attribute, which is linked to businesses that has been extracted from webpages. In this study following three OBECs are derived:

- one or many URLs, if existing
- existence of a link to social media link (binary)
- existence of an online shop (binary)

Data Sources

For the production of OBECs the following data sources were used

- Statistical Business Register (SBR)
  - Enterprise ID
  - Enterprise Name
  - Enterprise post address
  - Enterprise VAT and or commercial register number
  - Enterprise grouping ID, to link enterprises belonging to the same enterprise group
- Bing Search
  - up to 10 suggestions for enterprise URL

- Enterprises web sites
    - up to 25 pages from the enterprise website

For the validation of OBECs, the European survey on ICT usage in enterprises 2019 was used.

ICT usage survey.

The European survey on ICT usage in enterprises and e-commerce is carried out annually in all EU countries, according to a European harmonized questionnaire (Eurostat 2020). It covers ICT development in enterprise, ranging from internet access and broadband connections to new fields of digitization as robotics. Questions, compared with derived OBECs within this paper, are formulated in the European-harmonized model questionnaire as follows:

- Does the enterprise have a website?
- Does the website have any of the following: links or references to the enterprise's social media profiles?
- Does the website have any of the following: Online ordering or reservation or booking (e.g. shopping cart)?

A sole description of goods and services or price lists on the website and reservations with e-mails are not included in the concept of an online shop.

Population

OBEC were derived from the 2019 ICT brutto sample of 5200 enterprises. A response rate of 53% yielded in a netto sample of 2745 enterprises. The share of successfully scraped enterprise homepages made out 76.52%. The scraping process of the whole ICT survey sample took approximately 7 days.

In general, webscraping is not restricted to samples: The definition of the population underlies the production process, e.g. depends on the technical feasibility to scrape a very large number of webpages during the observation period. The statistical indicator could also be estimated using the OBEC from a sampled subset of the target population and calibrating sampling weights accordingly.

Unit type

Considered variables of the ICT usage survey refer not only to the website of the business itself but also the website of the parent or holding company. In this paper, OBECs derived for one enterprise are assigned to all enterprises in it's enterprise group. A concrete example: A single URL is valid for all enterprises in the enterprise group. Also an enterprise can own multiple URLs.

Time Periods

OBECs were collected in accordance to ICT survey's field phase (annually February to May). Unequal dates of survey answer and deriving OBECs can still result in contrary results.

In general, periodicity of webscraping can be increased if scraping does not place much burden on the URLs visited. For an extension of the SBR with scraped-URLs, it may be necessary to scrape and retrieve the URLs of a business more frequently in order to keep the inventory up-to date for the retrieval of other OBECs.

ESSnet WPC II

Statistics Austria is member of the ESSnet Big Data, a project with the European statistical system (ESS) to explore the usage of Big Data soucers for Official Statistics (ESS 2020).

ESSnet Big data I ran from February 2016 until May 18 and consisted of 10 work packages, with Work Package 2 exploring ''Webscraping enterprise characteristics''. In the currently conducted follow-up-project ESSnet Big Data II 12 work packages are explored since November 2018 until December 2020. Work Package C''Enterprise characteristics'' builds upon results from results from ESSnet Big Data I WP2.

Software

Calculations were preformed in R (R Core Team 2013). Following R packages were used:

Package	Citation
`data.table`	(Dowle and Srinivasan 2020)
`jsonlite`	(Ooms 2014)
`urltools`	(Keyes et al. 2019)
`wdman`	(Harrison 2020b)
`RSelenium`	(Harrison 2020a)
`processx`	(Csárdi and Chang 2020)
`robotstxt`	(Meissner and Ren 2020)
`rvest`	(Wickham 2019)
`ranger`	(Wright and Ziegler 2017)
`rpart`	(Therneau and Atkinson 2019)
`e1071`	(Meyer et al. 2019)

Methodology

The methodology to produce figures for the existence of a website - equal to finding an enterprise's URL - is substantially different to the technology to answer questions about

existence of a link to social media or an online shop. For the latter, the existence of an enterprise's URL is a precondition.

The following three subsections explain the methodology applied to derive results for the three questions respectively. Methodology is in accordance with general concepts developed in ESSNet Big Data II WPC.

Existence of website

To produce shares of enterprises having a website, the webscraping approach is as follows:

1. Finding potential URLs to be the enterprises' website via search engines.
2. Validate potential URLs with SBR information, e.g. by VAT or commercial register number

If an URL found in step 1, is validated in step 2, this URL is defined to be the enterprise's URL. It may occur that more URLs than one are validated for an enterprise. If, for an enterprise at least one URL is validated, this enterprise counts to have website.

During ESSNet Big Data I WP2 multiple countries explored a variety of approaches for the implementation of step 1, the collection of URLs for enterprises. Results and documentation for those can be found under https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_URL_retrieval1 and https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPC_Documentation. In addition (Ten Bosch, Delden, and Windmeijer 2020) presents an in depth analysis of the search for enterprise URLs in the Netherlands.

Step 2, identifying an URL as the enterprise's website, depends on the Austrian Media Act §25 which instructs enterprises to identify themselves on their own website under certain conditions. This always includes stating the enterprises name and full address and in many cases extends to listing their VAT or commercial register number as well. Using these unique identifiers we can match the enterprise directly with the URL.

If these identifiers are not found or do not match up we might still be able to match the URL to an enterprise if we can find the full name and address. This approach does however have the downside that enterprise names and even addresses might differ slightly between the SBR and what is stated on the webpage. This can include using a space instead dash or specific abbreviations in the name or address. To mitigate this issue we search for each part of the enterprise name as well as each part of an address, e.g. postal code, street, house number, ect…, separately. For each URL and enterprise we thus collect a binary vector holding for each entry: name, postal code, street, house number, ect… a 1 if we found this specific feature on the website. Using this we build a random forest model to classify if a URL can be linked to the enterprise.

A stepwise description of the methodology is as follows

1. Get enterprises from the ICT survey sample with the following data: IDs, Names, post address.

2. Create search string and run Bing Search
    - For the BING Search the software under https://github.com/EnterpriseCharacteristicsESSnetBigData/UrlSearcher was used.
3. Read in URLs from url search and discard blacklisted URLs like yellow pages, etc… .
4. Scrape each URL, specifically site pages containing contact information like the imprint.
5. Load scraped data and extract unique identifier, like VAT or commercial register number or search for company names and address.
6. Match found VAT or commercial register number with SBR.
7. Build model using found indicators for found names and/or addresses and apply on set of URLs which have not already been identified by step 6.

Link to Social Media

Deriving OBECs for link to social media or an online shop, depends on finding a URL for an enterprise. The idea behind retrieval is straight-forward: If a website has a link to social media, this must appear in the enterprise's source code. Therefore, two steps were proceed:

1. Load scraped data
2. Retrieve social media links (Facebook, linkedin, twitter, plus.google, youtube, instagram) found in scraped data. Social Media links not indicating a presence on social media (e.g. sharing links) were excluded.
3. If at least one link was found, a binary variable indicating the existence of a link to social media, was set TRUE.

Online Shop

In a similar fashion scraped data was examined for hints to an online shop:

1. Load scraped data
2. Retrieve core words concerning online-shopping found in scraped data. Specific links (e.g. policies, services, data regulations, etc.) were excluded.

Detecting an online shop with web scraping is less straight forward than the detection of a link to social media: For online shops, no definite word must appear in the website's source code. Therefore, a list of core words was defined in a systematic manual validation process of enterprises' websites. Core words were classified as ''strong indicators'' or ''medium indicators''.

In an expert-based approach online-shops were derived as follows: If at least one strong indicator or a minimum number of medium indicators was found, a binary variable indicating the existence of an online shop, was set TRUE. In a modeling approach, the existence of an online-shop was predicted with classification tree, naive Bayes and random Forest. Two thirds of data were used as trainings data.

In the search of core words, similar problems as for addresses occurred: Core words exist in a variant forms of spelling, e.g. online shop was found as ''online shop'',''onlineshop'' and ''online-shop''.

Results

In the following the derived OBECs are compared with survey results of the ICT enterprise survey. The results do not include the use of survey weights to present a more direct comparison between the survey and webscraped results on a unit level. Including the survey weights for the following results does slightly change the magnitude of the numbers but the differences between the survey and webscraped results are roughly the same.

Existence of website

For 2019, the survey data contained 93.66% of enterprises stating to have a website wheres webscraping detected 82.44% under all survey respondents (see Table 1). The coherence matrix, shown in Table 2, indicates an accuracy rate of 85.06 %.

Table 1: Rate of enterprises having a webpage

	URL rate
ICT Survey	93.66
Webscraping	82.44

Table 2: Crosstable enterprise listed having a URL in ICT survey vs. URL was found using webscraping.

		Webscraping		
		0	1	
ICT	0	123	51	174
	1	359	2212	2571
		482	2263	2745

Enterprises unequally classified were validated manually, reasons for mismatch were systematized: Errors for finding no URL with webscraping, although indicated by survey, may occur on two levels:

1. The correct URL was not found within the Bing search.
2. The correct URL was found, but not validated as corresponding to the enterprise.

Both levels contributed in balance (about 50% of mismatch each). While improvement in first level is limited, the second level validation process was continuously evaluated and improved. The main error source – a lack of Enterprises's ID on website- could be mainly overcome with the above mentioned integration of random forest modelling. Remaining error sources include

- outdated enterprise information on the website
- incorrect enterprise information on the website
- disallowment of scraping via robots.txt

For errors on finding an URL with webscraping, although no URL indicated by survey, webscraping entrenches itself to may be more precise in capturing the statistical unit of the European ICT surveys than the survey itself: Enterprises of the same group are included in the websearch. This is not always the case in survey answers. Limits occur with foreign enterprises in the same group.

Link to Social Media

For 2019, within all survey respondents, 51 % indicated to have social media on their website, for webscraping it were 50 %. With restriction to enterprises with URLs found, the share resulted from Webscraping was higher than the share produced by survey (survey: 56 %, webscraping: 61%). The accuracy rate is 74.05% .

Again mismatches were manually validated: The main reason for false-negatives due to webscraping, may be a conceptional one: Due to the rapid change of social media platforms, no complete list of social media platforms is available for the European ICT surveys. Hence, a different view on social media may lead to incoherence between both methods. Another reason is that social media references in too deeply nested links, are not captured by webscraping.

False-positives mostly occur from social media sites from incorrect enterprises. A way to overcome this issue would be a similar validation process as in the set up of the URL repository. However, the therefore required extensive scraping on social media platforms may be affiliated to costs or is prohibited.

Table 3: Rate of enterprises using social media on their URL. Top: all survey respondents from ICT survey; Bottom: Subset of survey where URL was successfully found through webscraping

Type	ICT Survey	Webscraping
all survey respondents	0.51	0.50
Webscraping found URL	0.56	0.61

Table 4: Crosstable enterprise listed using Social Media on URL in ICT survey vs. Social Media Link was found using webscraping. Subset of survey where URL was successfully found through webscraping.

		Webscraping		
		0	1	
ICT	0	636	348	984
	1	226	1002	1228
		862	1350	2212

Online Shop

For 2019, within all survey respondents, 28 % indicated to have a webshop on their website, webscraping 23 %. The accuracy rate is 75.78 %. Specificity is 80.97 %, sensitivity is 62.46%. Tables 5 and 6 refer to the expert-based approach.

Similar accuracy rates occurred in modelling data with a classification tree model (75%), naive Bayes (76%) or random Forest (79%) - with random forests performing slightly better. Specificity outperforms the expert-based approach in all three model variants (classification tree: 92%, naive Bayes: 87%, random Forest: 92%); on the contrary sensitivity is poorer in all models, with rates smaller than 50% for the classification tree and the Random Forest model (classification tree: 45% , naive Bayes: 52%, random Forest: 48%). No notable difference were observed in precision rates in all 4 methods (close to 80% each).

Considering these quality measures, neither the expert-based approach, nor a modelling variant succeeds the others. Sensitivity, the rate of detecting webshops as webshops with webscraping, is most convincing in the expert-based approach. This conclusion is in accordance with results from ISTAT from 2016, see (Barcaroli 2015): They summarized, that ''as for the proportion of web sales functionality, in general data mining learners fail in reproducing the correct aggregates''. In the referred paper naive Bayes approach was regarded as best method with accuracy rate of 50%.

Systematic validation of mismatches revealed a dependency of online shop wording from NACE categories: E.g. in tourism an online-shop is seldom termed as such or similar, while words as booking are a strong indicator for the presence of an online shop. By contrast, ''booking'' in other NACE categories is a not strong enough indicator. An integration of NACE categories is planned.

Therefore further development of expert-based classification as well as for modelling approaches is needed. By now, sensitivity rates of modelling approaches are too low to be considered as acceptable.

Table 5: Rate of enterprises with online shop on their URL. Top: all survey respondents from ICT survey; Bottom: Subset of survey where URL was successfully found through webscraping

Type	ICT Survey	Webscraping
all survey respondents	0.28	0.23
Webscraping found URL	0.31	0.28

Table 6: Crosstable enterprise listed with online shop on URL in ICT survey vs. online shop was found using webscraping. Subset of survey where URL was successfully found through webscraping.

		Webscraping		
		0	1	
ICT	0	1268	229	1497
	1	298	381	679
		1566	610	2176

Conclusion

In this paper the potential of webscraping to construct online based enterprise characteristics was shown for Austrian enterprises. Overall webscraping exceeded in good accuracy rates, when compared with ICT survey results.

However, rates are not promising enough to replace survey results with online based enterprise characteristics yet: The methods presented in this paper suggests discussed OBECs to be used for supplementary information during micro plaus, but not yet for replacement of survey results. A prolongation of periodicity of discussed items as compulsory items in ICT model questionnaire could be reached trough modelling with OBECs and their population could be extended.

Accuracy rates in comparison to survey results also do not reflect the complete truth. True values of web characteristics can be derived with manual validation - unlike most other survey values. Coherence matrices of OBECs versus true values as well survey results versus true values are needed.

In addition the scraping process should also be extended to include the whole ICT population. For the ICT 2020 survey this is done using the Google Search API instead of BING Search. Final results are however not yet available since the survey results for 2020 are not completely validated yet.

Further development of webscraping could also enable the collection of OBECs, usually not systematically collected with survey or administrative data, as. e.g. whether a business is innovative or not, whether it concerns a family business or not or whether it develops AI technology or not.

Acknowledgments

The content of this paper is heavily based on deliverables produce for ESSNet Big Data II WPC. The whole content of WPC and all other work package can be found on our Wiki platform
(https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page).

Literature

Barcaroli, Giulio et al. 2015. "Internet as Data Source in the Istat Survey on Ict in Enterprises." *Austrian Journal of Statistics* 44: 31–43.

Csárdi, Gábor, and Winston Chang. 2020. *Processx: Execute and Control System Processes*. https://CRAN.R-project.org/package=processx.

Dowle, Matt, and Arun Srinivasan. 2020. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

ESS. 2020. *ESSnet Big Data*. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page.

Eurostat. 2020. *ICT Usage in Enterprises*. https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm.

Harrison, John. 2020a. *RSelenium: R Bindings for 'Selenium Webdriver'*. https://CRAN.R-project.org/package=RSelenium.

———. 2020b. *Wdman: 'Webdriver'/'Selenium' Binary Manager*. https://CRAN.R-project.org/package=wdman.

Keyes, Os, Jay Jacobs, Drew Schmidt, Mark Greenaway, Bob Rudis, Alex Pinto, Maryam Khezrzadeh, et al. 2019. *Urltools: Vectorised Tools for Url Handling and Parsing*. https://CRAN.R-project.org/package=urltools.

Meissner, Peter, and Kun Ren. 2020. *Robotstxt: A 'Robots.txt' Parser and 'Webbot'/'Spider'/'Crawler' Permissions Checker*. https://CRAN.R-project.org/package=robotstxt.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2019. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), Tu Wien*. https://CRAN.R-project.org/package=e1071.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between Json Data and R Objects." *arXiv:1403.2805 [stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Ten Bosch, Olav, Arnout Delden, and Dick. Windmeijer. 2020. "Searching for Business Websites." Discussion paper. Centraal Bureau voor de Statistiek, Den Haag, NL.

Therneau, Terry, and Beth Atkinson. 2019. *Rpart: Recursive Partitioning and Regression Trees*. https://CRAN.R-project.org/package=rpart.

Wickham, Hadley. 2019. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wright, Marvin N., and Andreas Ziegler. 2017. “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.