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

WP4: Methodology and Quality

Deliverable 4.6: WP4 Methodology report on using webscraped data

FINAL VERSION

Final version, 2025-02-17

Prepared by:

1. ***WP leader: Alexander Kowarik (alexander.kowarik@statistik.gv.at) and Magdalena Six (magdalena.six@statistik.gv.at)***
2. *Piet Daas (Statistics Netherlands)*,*
3. *Johannes Gussenbauer,*
4. *Jacek Maślankowski (Statistics Poland/Univ. Gdansk).*

** main author.*

Web Intelligence Network

Funded by the European Union

This deliverable 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.

Content

1. General introduction

This report provides an overview of the methodological work performed in Workpackage 4 of the Trusted Smart Statistics - Web Intelligence Network (WIN) project. An overview of the WIN project and its work packages can be found online at https://cros-legacy.ec.europa.eu/WIN_en.

The WIN project focuses on collecting and using webscraped data for the production of official statistics. A general introduction to the topic of webscraping and using webscraped data -in the context of official statistics- can be found in the paper by ten Bosch et al. (2018) and in the presentation of Greenaway (2017). An overview of the most important methodological topics when using Big Data in the context of official statistics can be found in the ESSnet Big Data (BD) II report of WPK (Del. K9, 2020). Workpackage 4 of the WIN project is divided into four distinct tasks which are: 1. Quality (led by Austria), 2. Methodology (led by The Netherlands), 3. Architecture (led by Italy) and 4. Quality Assessment. The first three tasks are performed by WP4 members. The fourth task focuses on Online Job Advertisement (OJA) data and is jointly performed with WP2 under their leadership.

When we talk about methodology in WP4 it specifically refers to "the way statistics are produced when using webscraped data". Both the order of the steps and the specific methods used in each step are essential to ensure that high-quality statistics are being produced. Considering the work performed in the WIN project, it is obvious that, in principle, specific methodological methods are being developed in work packages 2 and 3, which each focus on producing specific official statistics. The ultimate goal of WP4 is, based on past experiences in the WIN and other research projects, to identify generically applicable methods that can be used to produce multiple web-based statistics of high quality. Based on the current state of the art, this report focuses on three important generic methodological approaches when one plans to use web-data to produce official statistics. These topics are: i) Sampling, ii) the Webscrape process, and iii) Methods specific for dealing with webscraped data.

This report underwent two rounds of review by David Salgado. The first review was conducted approximately one year prior to the final deadline on a partial draft, while the second review focused on the complete version shortly before the submission deadline. We extend our gratitude to the reviewer for his positive feedback and constructive comments, which significantly contributed to enhancing the quality of this document.

1.1 Structure of the report

This report is composed of five chapters. Apart from the introduction, the three major topics discussed are Sampling (Chap. 2), the Webscrape process with special attention to sources of bias (Chap. 3), and Specific methods used when dealing with webscraped data (Chap. 4). The final chapter (Chap. 5) is a discussion on the methodological issues identified. The document ends with a list of references to the papers and presentations cited.

2. Sampling

This chapter gives an overview of using sampling-based methods in the context of webscraped data. First, the use of sampling when annotating webscraped data is discussed. Next, topics on the usefulness of sampling in the general context of webscraping, during (model-based) estimation, and when only a part of the population is scraped are dealt with.

2.1 Sampling for quality assessment

In quality assessment, sampling is crucial for evaluating the accuracy of data. This is especially true if manual annotation should be performed as this labour-intensive task needs to be limited and optimized. An annotated (pre-labelled) data set is of utmost importance when assessing the quality of automatically classified data. Ideally, the annotated data set is large in volume and the annotated variables are of high quality. One obstacle in creating such a data set is the high cost, in terms of time and resources, that comes with manually annotating a large number of records. To mitigate the costs one can reduce the volume by creating a sampling design such that certain margins are well represented in the sampled data. Drawing such a sample can lead to a higher quality given the number of cases to be annotated or to a lower number of cases given the quality of the accuracy estimates.

The annotated data can subsequently be used as test and training data for i) developing a classification model or ii) monitoring the quality of automatically classified records. Depending on the choice made, the sampling design may need to be adjusted. In this chapter, the main focus is on classification.

2.1.1 Sampling design

Given a single classification variables V_i with outcome values $v_1, \dots, v_{n(i)}$ the sampling design can be choosen such that a given quality measure for each dimension can be estimated with a certain degree of accuracy. For instance, Table 2.1.1 shows the degree of uncertainty when estimating the accuracy of the classification algorithm for a single outcome category in dependence of the number of annotated records. When dealing with many classification variables and possibly many outcome values for each variable the number of annotated records needed to derive a high degree of accuracy can quickly reach tens of thousands.

This task has been performed for Austria in the WIN-project. The main purpose of the 2nd annotation round in this project was to measure the accuracy of the following variables

- Location (NUTS - NUTS1 or NUTS2 regions)
- Education level (ISCED)
- Occupation (ISCO 1st digit)
- Economic Activity (NACE - sections)
- Working time (full-time, half-time, not indicated)

The above-mentioned variables have between 3 (Type of Contract) and more than 30 (NUTS2 regions) categories. Choosing a sampling design by taking the joint distribution of these 5 variables and sampling

Table 2.1.1. Approximate range of confidence interval for estimating accuracy given different sample size, assuming that the true accuracy is 50%.

Sample Size	Range Confidence Interval
10	+/-37.7%
15	+/-28.66%
20	+/-24.01%
25	+/-21.06%
30	+/-18.99%
50	+/-14.35%
100	+/-9.97%
150	+/-8.09%
300	+/-5.69%

at least some records in each cell would result in annotating much more than 1000 or even 10000 records. Thus this task in the WIN-project aimed for sampling records such that the marginal distribution of each one of the 5 variables has a given distribution and the overall sample size is equal to n. Tables 2.1.2- 2.1.6 show the expected sample size, given an overall sample size of n=300, for each outcome value in each of the 5 variables using as location variables the NUTS2 regions from Austria. With this approach amount of annotated records was kept very low in return the sample size only allows us to measure the overall accuracy, for each variable independently, to a sufficient degree of precision.

Table 2.1.2. Expected sample size by education level

Dimension	ED1	ED2	ED3	ED4	ED5	ED6	ED7	ED8	Not indicated
Education level	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33

Table 2.1.3. Expected sample size by occupation 1st digits

Dimension	OC1	OC2	OC3	OC4	OC5	OC6	OC7	OC8	OC9
Occupation 1st Digits	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33

Table 2.1.4. Expected sample size by economic activity (sections)

Dimension	A	C	D	E	...	Q	R	S	T
Economic Activity (Sections)	15.79	15.79	15.79	15.79	...	15.79	15.79	15.79	15.79

Table 2.1.5. Expected sample size by working time

Dimension	FT	PT	Not indicated
Working time	100	100	100

Table 2.1.6. Expected sample size by NUTS2 regions (example from AT).

Dimension	AT11	AT12	AT13	AT21	AT22	AT31	AT32	AT33	AT34	Not indicated
NUTS2	30	30	30	30	30	30	30	30	30	30

2.1.2 Selecting the sample

The definition of the sample design is only based on the defined distribution by variable, which we will call margins in this setup $\{\mathbf{t}_1, \dots, \mathbf{t}_M\}$. For our example in the previous sections, these are the tables 2.1.2 to 2.1.6. Furthermore, the inclusion probabilities of each unit is not easily defined by its characteristics as in, e.g., stratified sampling, where the membership to a certain stratum defines the inclusion probability. Inclusion probabilities depend on the joint distribution and can be simulated by repeatedly drawing and optimizing a sample with the below described methodology.

Since a joint distribution is not available for drawing the sample directly, the selection of the sample becomes an optimization problem. Let $f(\mathbf{t}_1, \dots, \mathbf{t}_M, \hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M)$ be an objective function that measures the difference between two sets of margins,

- the pre-defined margins$\{\mathbf{t}_1, \dots, \mathbf{t}_M\}$ and
- the margins of the sample (subset) $\{\hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M\}$,

then the optimisation problem can be defined as follows

$$\min_{\hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M} f(\mathbf{t}_1, \dots, \mathbf{t}_M, \hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M)$$

$$\hat{\mathbf{t}}_m = \begin{pmatrix} \hat{t}_1 \\ \vdots \\ \hat{t}_{R(m)} \end{pmatrix}$$

$$\hat{t}_{r(m)} = \sum_{i=1}^{N} \delta_i \, 1_{[\text{record } i \text{ part of margin cell } r(m)]} \quad m = 1, \dots, M, \; r = 1, \dots, R(m)$$

$$\delta_i \in \{0,1\}, \; i = 1, \dots, N$$

where N is the number of all records which can be annotated and $1_{[.]}$ represents the indicator function which equals 1 if the expression in $[.]$ is true.

Using a heuristic algorithm like simulated annealing we can find a solution to the above problem. The algorithm goes through the following steps

1. Randomly initialize a sample of size n and set starting temperature T
2. Compare the margins resulting from the sample $\{\hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M\}$ to the target margins $\mathbf{t}_1, \dots, \mathbf{t}_M$ and calculate the initial value of the objective function $Obj_0 = f(\mathbf{t}_1, \dots, \mathbf{t}_M, \hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M)$.
3. Randomly add and discard some records from the sample
   - Sample with probability according to over- or under-representation in current target margins $\{\hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M\}$
4. Re-calculate $\{\hat{\mathbf{t}}_1, \dots, \hat{\mathbf{t}}_M\}$. If the difference between sample and target margins is small enough → stop, otherwise go to 4.
5. Check if current solution has become better or worse than previous one $Obj_s < Obj_{s-1}$
   - Accepts worse solution with a probability of $\exp\left(-\frac{Obj_s - Obj_{s-1}}{T}\right)$
6. Cooldown T by fixed factor.
7. Terminate if maximum number of iterations has been reached otherwise go to step 2

This procedure is available for instance in the R-package simPop, see Templ et al. (2017) (function calibPop()) which is on CRAN: https://CRAN.R-project.org/package=simPop and on github: https://github.com/statistikat/simPop

2.2 Sampling in the context of webscraped data

For web-data, it can be beneficial, and especially more efficient time-wise, to study samples (of the websites) of the population under study. Here, various probability-based samples can be used. When datasets with unknown inclusion probabilities are used, non-probability approaches should be considered.

2.2.1 Probability sampling

When dealing with web-data, probability sampling can play a role, especially for estimation. If the process of deriving a target variable, is not easily scalable, e.g. a statistical classification needs costly manual intervention or quality control is needed, drawing a random sample from all available units/websites can be used. The situation is thus similar to a survey where each interview has a high cost and cannot be extended easily to the full population. There is a rich body of methodology developed for inference from random samples from a method for the sampling design and the applied estimation can be selected. The key methodologies are listed in (Särndal et al. 1992).

Simple Random Sampling

Simple random sampling is the most straightforward method, where each unit in the population has an equal chance of being selected. This method is ideal when the entire population is well-defined and accessible and there is no auxiliary information available that would lead to an improved sampling design. It provides a clear, unbiased representation of the population (Särndal et al. 1992, Chap. 3.3).

Stratified Random Sampling

Drawing independent simple random samples from distinct subgroups (strata) of the population. This approach ensures that all relevant subgroups are adequately represented in the sample, e.g., to control the precision within such subgroups. This is particularly useful in heterogeneous populations, where the stratification variables are correlated with the target variables (Särndal et al. 1992, Chap. 3.7).

Systematic Sampling

In systematic sampling, every n-th unit beginning with a random start is sampled from a full list of the population. If the list is randomly sorted, this method is an efficient variant of simple random sampling. If the list is sorted according to some stratification variables and random within, this is a sampling design with so-called implicit stratification. Sorting the list according to a numeric variable, e.g. number of page visits to a website, systematic sampling would ensure a balanced design according to this variable (Särndal et al. 1992, Chap. 3.4).

Cluster Sampling

In cluster sampling, a unit is not sampled directly, but a cluster is selected and then all units within this cluster are part of the sample. In the case of web-data, a cluster can be a subset of adds on a vacation home booking platform, e.g. geographically. This method is useful when a list of all available units is not available or is costly to generate (Särndal et al. 1992, Chap. 4).

2.2.2 Non-probability sampling

Web-collected data can also be considered as a non-probability sample. This means that each unit in the target population has an unknown (and unequal) chance of being included in the data set. This happens, for instance, when non-random criteria like availability, geographical proximity, or expert knowledge are used in the data collection process. The methods applied to deal with the unknown inclusion probabilities of 'non-probability' samples try to limit this effect as much as possible and therefore aim to improve data quality (Baker et al. 2013, Vehovar et al. 2016, Elliott and Valliant 2017, Wu 2022). Reducing the bias is the major concern for official statistics (van den Brakel 2019, p. 12). The most common suggested correction methods considered for these data sources are the following.

Calibration / Weighting

By generating a set of weights for the collected units in the non-probability sample, which match known population margins, one aims to ensure that the sample reflects the population's distribution of key variables (Lee and Valiant 2009).

Statistical Matching

Here, the general idea is to combine the non-probability data set (A), e.g. a subset of scraped websites, with another data set (B), either a population or a probability sample. Statistical matching can help to overcome the limitations of non-probability samples by leveraging information from the second data set (B) with known inclusion properties. The combined data set produced is a synthetic data set in which matched variables common to both data sets are used. In other words, the target variables from data set A are matched to data set B and then either i) traditional estimation methods for probability sampling are used or ii) if data set B includes the full population, it can be used directly for estimation (Baker et al. 2013).

Propensity Score Weighting

Based on a set of auxiliary variables, the probability that a specific unit is part of the non-probability sample is attempted to be estimated. The inverse of this probability can be used in weighted estimation to compute population estimates. To apply this method, some characteristics of the full population need to be known (Lee and Valliant 2009).

2.3 Selective scraping

Selective scraping, also known as statistical scraping (ten Bosch et al. 2024), focuses on deliberately scraping a subset of the target population. It aims to collect limited amounts of web-data tailored to specific statistical needs. Unlike so-called bulk scraping, which collects large volumes of data on a given subject (see, for instance, Daas and van der Doef 2020), selective scraping uses pre-existing knowledge of the subject to collect specific information on specific units in a more controlled manner. Ultimate goal f selective scraping is collecting data of a representative set of units for the topic studied.

Statistical offices have the advantage of having wide sets of data available that can be used as a-priori information for a (selective) scraping process. Examples are statistical registers, classifications, and administrative data sources. For instance, in the enhancement of business registers, selective scraping is using known identifiers, e.g. enterprise names or tax numbers, to search for digital traces online, such as websites, media advertisements, or job postings. This collected data is directly linked to a statistical unit in the business register and can be used to infer information on for example the economic activity of an enterprise. The process of selective scraping can be subdivided into three specific phases:

Source Identification

This phase involves identifying the URLs or more generally the sources associated with statistical units, often using search engines or domain registries. Machine learning models can help select the best matches by scoring entries in the search results. This part is also referred to as ‘URL finding’ (see section 3.2).

Source Selection

The second phase requires one to decide on which units’ information needs to be collected. This could involve scraping all URLs linked to a unit or a subset of the linked URLs based on some selection criteria. For more info on the latter the reader is referred to Deliverable 4.5, section 2.6 (ESSnet WIN WPK, 2024).

Data Extraction and Enhancement

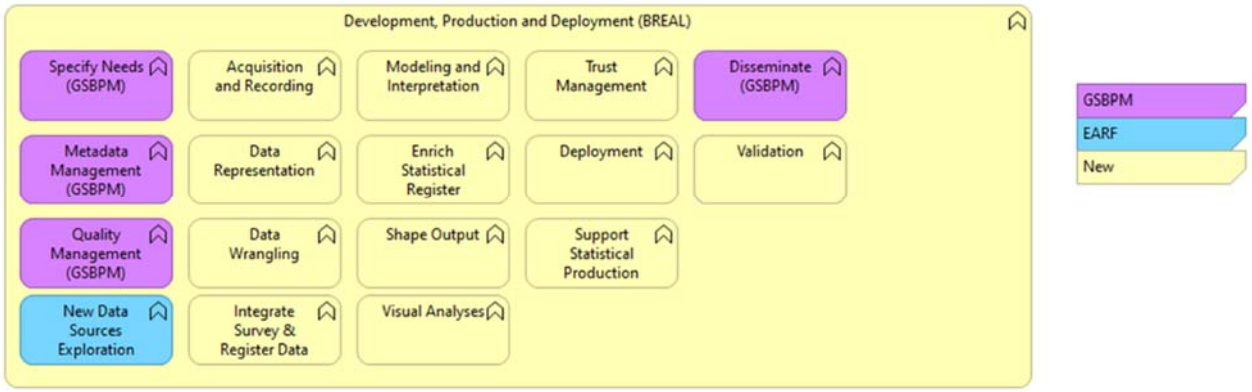
Once relevant sources (URLs) are identified and selected, scraping is performed. Techniques such as natural language processing interpret the raw text and derive the needed target variables (Daas and Maślankowski 2023).

In summary, selective scraping tries to simplify the estimation/inference step by starting from a-priori information. This information, e.g., the link to a unit in the statistical business registers, is then used in the estimation step. In some webscraping projects, this approach allows for a precise and targeted data-collection that directly covers the need for specific information and builds on the knowledge of statistical offices. As such, it can seriously reduce the scraping effort.

3 Webscrape process and causes of bias

This section provides an overview of the web-based statistical process and important methodological issues. Each section in this chapter starts with a BREAL business function (ESSnet BD II, WPF 2019) followed by a description indicating the specific step discussed. Figure 3 gives an overview of the business functions distinguished in BREAL. In each section of this chapter, an overview is given of the most important issues identified and a description of the 'methodological state of the art'. The topics identified are generally discussed in the context of their effect on the estimation process as this has been identified as the most important issue during the P2P WP4 meeting in Vienna (14-15 Feb. 2023).

Figure 3. Overview of BREAL business functions (from ESSnet BD II, WPF 2019).



3.1 Specific needs: Population frame

Depending on the topic studied the most relevant population frame needs to be used. Two different cases are considered here, these are: i) Studying businesses and ii) Studying other phenomena.

Studying businesses

In this case, usually, the Statistical Business Register of the NSI is used, certainly when various types of businesses need to be identified (UNECE 2015). However, even when businesses are studied, this register may not always fully cover the population. For instance, when a topic is studied that includes considerable numbers of internationally active businesses or non-profit organizations. In the first case, one has to realize that internationally active businesses may not be completely included in the Business Register of a country (UNECE 2015, Chap. 3). Here, additional data may be available in, for instance, the European Business Register or other worldwide registers (Wikipedia 2023). When non-profit organizations are the main focus, job-register data or a similar source could be used to identify these types of organizations. Perhaps, lists of such organizations are available online. Another starting point for identifying websites of businesses is an as complete as possible list of domain names[1] in a country (URL finding 2022) to which

[1] A uniform resource locator (URL) contains the domain name of a site as well as other information. In, for example, the URL 'https://ec.europa.eu/eurostat', 'europa.eu' is the domain name,'eu' is the top-level domain name (often a country-code), 'ec' is the subdomain, 'https' is the protocol, and '/eurostat/' is the path to a specific page on the website.

units could be added. The overall goal, in this case, is the creation of a frame with as high as possible coverage for the units of interest.

Studying other phenomena

When studying other types of units, such as job vacancies (ESSnet BD II, WPB 2018), the business register or other administrative registers are not always the best starting point. This is particularly the case when objects are studied for which no register is available. A literature study revealed various phenomena for which webscraped data was found particularly relevant. Examples of such phenomena are hospitality services (Han and Anderson 2021), textile fiber data (Muehlethaler and Albert 2021), ecstasy use (Maybir and Chapman 2021), and local policy variation (Anglin 2019). In each of these studies, webscraping was used to collect (as much as possible) data on a particular phenomenon. Without the web, it would be extremely difficult to study the phenomena in sufficient detail. Depending on the phenomena studied, preliminary searches were performed to find out if websites with relevant information were available. Another approach could be to use search engines to find relevant URLs. Depending on the topic of interest, this could provide interesting information but may not provide a representative overview of the units involved (Daas et al. 2022). More information on such approaches can be found in the Landscaping paper of the ESSnet WIN (2023).

3.2 Enrich statistical register: Adding URLs

After a population frame has been selected, there is a need to obtain the link to the website (the URL) of each unit; if these have not already been included. Several approaches can be used to add URLs here (Barcaroli et al. 2016).

First, the URLs can be searched for by using a search engine[2] and, usually, the name and contact information are used as input. This approach is commonly referred to as the 'URL-search' approach and more details can be found in the ESSnet WIN report on URL-finding methodology (2022). The downside of such an approach is that multiple URLs may be found for various units, which also depends on the type of unit (e.g. legal unit, local unit, etc.) for which the website is being searched. As a next step, this usually requires selecting the most appropriate URL from a number of options. In addition, depending on the topic studied, not all 'businesses' may have a website. For instance, IT and other types of technological businesses are more likely to have a website compared to farmers and hairdressers. In the latter cases, it is extremely likely that URL-search approaches will provide considerable numbers of non-relevant URLs for the businesses at hand. This can, for instance, result in URLs found for businesses that actually do not have a website. The latter needs to be checked (see below). For some businesses, especially small ones, a Facebook link may actually be the most relevant link.

Another way to obtain URLs is by making use of lists of URLs provided by other organizations. Examples of such lists are those produced by i) commercial organizations that actively search the web, such as DataProvider (Oostrom et al. 2016), ii) information provided by businesses when they register at the Chamber of Commerce, such as the domain-name part of their mail and/or web-address, and iii) lists maintained by country-specific or international organizations, such as the organization responsible for the internet domain registration in a country. Such lists may provide a very interesting starting point from

[2] A search engine is a software system that provides hyperlinks to webpages and other relevant information on the Web in response to a user's query. Examples of search engines are google.com, bing.com and duckduckgo.com.

which the website specific for the unit of interest can be obtained. The latter requires checking any information available on the website with those available in, for instance, the Business register. Usually, Chamber of Commerce numbers and/or address information are used for that purpose (Daas and van der Doef 2020). Unfortunately, not every website contains that information. In German-speaking countries, however, business websites need to provide essential business information on a so-called Impressum page. An Impressum page, which is sometimes called an "Imprint," contains legally required information about the business and owner of a commercial website. Because the structure of this page is highly standardized, it is a very interesting way to collect business-relevant data and check it with any other information available for a business.

3.3 Acquisition and recording: Scraping the web

After creating a list of (target) population units with their accompanying URL(s), the associated websites need to be scraped. The ability to scrape a website depends on several (predominantly) technical issues. First, the 'scrape-ability' of the website is an important concern. For instance, a website may not be active at the specific point in time during which a scrape attempt is initiated. This can be solved by performing multiple attempts on different days and times. However, some websites cannot be scraped because they simply no longer exist. The authors have noted that this occurs more often for websites of smaller businesses compared to larger ones. In addition, the webscrape technique used also affects the success of data collection. For instance, Daas and Maślankowski (2023) report a 10% increase in the collection of webscraped data for businesses in Poland when a 'headless browser'[3] scraping technique was used compared to a more direct, simpler, python-based approach. In addition, some websites might even specifically prevent certain groups of webscrapers to collect data by checking the identifier of a scraper or via their robot.txt file[4] (ten Bosch et al. 2018). The combination of all these technical issues may seriously reduce the percentage of websites from which data can be obtained. For official statistics, it would be interesting to perform a webscrape comparability study on a standardized large set of URLs to better understand the effect of these choices, especially regarding their effect on the representativity of the scraped data obtained. The reader is referred to the paper of Daas and Maślankowski (2023) and the presentation of Maślankowski and Daas (2023) for more details, which include a complete overview of the methodological and technical issues identified.

Next is the ability to scrape data from an existing website. Here, several technical issues may be relevant. Whenever a scraper attempts to obtain data from a website, essential information is provided via the so-called header of the scraping program to the webserver. The response of the webserver to the scraping attempt is provided in the form of a code; the so-called HTTP-status code (IANA 2022). Officially, these status codes range from 100 to 512 but sometimes codes with higher numbers are returned. In Table 3.3 an overview of these codes for a Dutch scraping study, that attempted to collect data from about 900.000 unique URLs, is given. In this table, the code 'None' is used to indicate that the scrape attempt failed (no

[3] A headless browser is a browser that runs without a user interface. Headless browsers are very popular in scraping because they can render JavaScript and (programmatically) behave like a browser used by a human. The latter has the advantage that it is less likely to be blocked and, hence, improves scraping. More on https://en.wikipedia.org/wiki/Headless_browser.

[4] A robot.txt file contains instructions for scrapers regarding the accessibility of particular files on a website. It is usually used to prevent access to particular parts of a website. More on https://en.wikipedia.org/wiki/Robots.txt.

code was returned by the webserver) and the results for the codes returned are aggregated at 100-sized intervals. What is interesting to note here is that, starting from a unique set of URLs, a considerable number of websites re-directed the requests to another URL (codes in the 300s). These codes compose nearly 40% of all requests. This suggests that a considerable number of requests resulted in data collected from other URLs than the ones initially visited. Hence, this suggests that potentially a considerable number of identical (duplicate) URLs could be scraped during the study. In addition, codes in the 400 and 500 ranges generally (except code 418 ;-) refer to various client and server errors. These usually result in the inability to collect data from the website. Overall, 82% of the websites were scraped (43%+39%), and a bit more than 17% did either not respond (10.5%) or produce an error code (6.7%+0.6%) during this process.

Table 3.3. Overview of the status codes observed in a large-scale Dutch scraping study*

Status codes	Amount	Percentage (%)	Code range	Description
None	147,088	10.5	-	No response
100s	0	0	100-103	Informational responses
200s	604,826	43.0	200-226	Successful responses
300s	552,243	39.3	301-308	Redirect responses
400s	93,567	6.7	400-451	Client error responses
500s	7,795	0.6	500-530	Server error responses
Other (>= 600)	245	0.02	600-999	Other (user-defined)
All	1,405,764	100	-	

*A maximum of four scrape attempts, at different dates and times, were used to access a website. When different status codes were returned during multiple visits, the one with the lowest value was chosen.

The issues mentioned above influence the ability to scrape websites and may introduce a bias in the final results obtained. Foederer (2023) discusses (part of) this topic, which he identifies as 'sampling bias', and he discerned three major causes: volatility (website change constantly), personalization (websites may show visitor-dependent information), and unindexed (not all websites are commonly known). Potential solutions are described in his paper. Marconi (2022) discusses context removal bias. This is bias caused by the removal of data before it is scraped. Bulk and high frequent scraping may seriously reduce this issue and is the recommended solution to this form of bias (Daas and Maślankowski 2023).

3.4 Data wrangling: Extracting features

One usually assumes that any website from which data is obtained can be used in subsequent analysis. However, this does not have to be the case. In generic scraping, a scraping approach that downloads complete copies of the HTML-files on a website, sometimes zero-sized HTML files are obtained. Since these contain no information (as they are empty), the only thing one can conclude is that the website must be active (otherwise no file would be obtained) and should -very likely- be scraped with a more advanced webscraping technique. Also, invalid (erroneous) HTML files can be obtained. This issue can be

solved by parsing[5] the file with an appropriate HTML/XML library, such as Beautiful Soup. This will ensure that a valid formatted HTML file is stored. Apart from that, it has also been found that some pages don't contain enough data to be used in the subsequent (modelling) steps. An example of this is a website containing a very limited number of words; for example, less than 20 in the study of Daas and van der Doef (2020). Increasing the number of pages to be scraped from a website (a domain) and combining the data from these pages is a way to deal with that issue (Daas et al. 2024). In the case of specific scraping, an approach that focusses on extracting a very specific piece of information from a website (such as the price of a particular product), it sometimes happens that the specific data searched for can no longer be found on the website. This is usually the result of a change in the layout of the page (ten Bosch and Windmeijer 2014). Adjusting the code to extract the required information for the new layout is the common solution to deal with this situation. An alternative, more general way of dealing with this issue, is to first scrape the complete HTML page (resulting in a local copy of the file) and subsequently extract the specific data needed. Such a two-step approach has the advantage that the original HTML page is available and can be studied to deal with the specific extraction issue.

The kind of features that can be extracted from webpages may seriously affect to ability by which a certain concept can be measured. An example that does not suffer from this issue is the price of a specific product. As long as the price is similar for the on- and off-line sold products and the product can be identified without a doubt (Cavallo and Rigobon 2016), this is a non-issue. However, indirectly measuring the concept of interest (Daas 2023), such as innovation based on website texts (Daas and van der Doef 2021) or vacancies from online job advertisements (Beręsewicz and Pater 2021), is much more challenging. Here, an association with the concept of interest needs to be derived in an as much as possible stable and reproducible way (Daas 2023). The study of online platforms is an example of a study in which the concept being measured by a Machine Learning model was confirmed via the response to questionnaires sent to (part of) the population of businesses studied (Daas et al. 2024). The downside of indirectly measuring a concept, is that drift, generally referred to as concept drift, may occur which requires a regular check of the association over time (see section 4.5). Some concepts are more sensitive to this than others. For instance, in a Dutch study, 'Innovation' started suffering from this after 6 months (Daas and Jansen 2022) while 'Online platform' detection (Daas et al. 2024) has been found stable for (at least) 4 years.

3.5 Modelling and interpretation: Model-based estimation

Usually, the data collected from the web is used to produce a model-derived estimate. Here, both traditional and machine learning models can be used. Especially in the latter case, various important sources of bias have been identified (Puts et al. 2022). Here, the effect on binary classifications is used as an example. In the latter case, the ratio of false positives (type I errors) and false negatives (type II errors) produced by the model (Meertens 2021) and the effect of the ratio of the positive and negative cases used in the training phase of the model (Puts and Daas 2021) both affect the outcome. Both can be corrected by a single method, described in Puts and Daas (2021). In the latter reference, a maximum likelihood estimator for the true proportion of positives in data sets is described and has been shown to produce an unbiased result. The reader is referred to section 4.6 for more details.

[5] In the context of the web, parsing is the process of analyzing the string of symbols in an HTML-file to check if they conform to the rules of the formal HTML syntax.

4 Methods specific for webscraped data

4.1 Enrich statistical register: URL matching

As a first step prior to scraping, to each business, a URL of its website needs to be added. This topic is specifically studied in WP3 of the ESSnet WIN (2022).

4.1.1 Direct URL search

When no other sources of information are available, the best way to obtain URLs for a set of businesses is by performing a URL search approach (ESSnet WIN 2022, Barcaroli et al. 2016). Here, the URL of a business is searched for by using a search engine, often Google or Bing, and a set of identifiable information for the businesses. The latter is usually the name and country but sometimes address information and other variables, such as an email address or phone number, are additionally included. The work of van Delden et al. (2019, Chap. 4) describes implementing such an approach in great detail. Here, the search results returned are a list of URLs, usually ten, from which the best candidate needs to be selected. For this task, a machine learning model was developed. The best result obtained had an F1-score of 84% (van Delden et al. 2019).

4.1.2 Via external data sources

Since June 2019, Statistics Netherlands has received a list of active URLs found for Dutch businesses from DataProvider (DP) on a monthly basis. DP is an external company that searches for URLs worldwide. Statistics Netherlands links these URLs to legal units in the Dutch Business Register. The linkage method works as follows: first, some basic text cleaning is applied to standardize the variables in both data sources. Next, all entries in both data sources are compared. A selection of variables, shown in Table 4.1.2 is used in this comparison. Each pair of identification variables is assigned a weight (original weight, third column of Table 4.1.2), that is added to the total linkage score when the values of the pair match. This total score is subsequently used as input for a function to generate a confidence score between 0 and 1 (linkage probability function). The confidence score is an estimate for the probability that a 'DP-URL – legal unit' pair is a true link or not. The accuracy of the approach was found to be 92% (Del 3.2, WP3).

Table 4.1.2. Variables used for URL-matching (from Del 3.2, WP3)

Dataprovider (DP)	SBR (Legal units)	Original weight	Updated weight
CoC-number*	CoC-number	500	5.02
Hostname	Website	500	1.67
Domain	Website	400	5.20
Email	Email	200	1.86
Secondary Email	Email	100	1.16
Zip code	Zip code	100	0.71
Telephone	Telephone	200	1.00
Secondary Telephone	Telephone	100	-0.48
Telephone	Mobile	200	1.78
Secondary Telephone	Mobile	100	0.42

*CoC-number = Chamber of Commerce number

Results of a detailed check of the quality of random samples of SBR-units and the development of a Logistic Regression model resulted in Updated weights (column four in Table 4.1.2) which made the model easier to maintain with the same linking quality (Del 3.2, WP3). For more details, the reader is referred to Del 3.2 of WP3 and the URL-matching document (ESSnet WIN 2022).

4.2 Acquisition and recording: Webscraping

Webscraping is the methodological process of collecting data from websites using automated tools or scripts. When conducting webscraping, it's essential to follow a structured methodology to ensure that the data is collected correctly, efficiently, and ethically. The following (intermediate) steps always need to be considered when performing any webscrape activities

4.2.1 Preparation:

Define objectives: Clearly define the purpose and objectives of the webscraping project. What data is needed and what is it used for? This step guides the scraping effort. Here, it should be determined if complete HTML pages need to be collected (generic scraping) or if specific information from webpages is needed (specific scraping). See Scannapieco (2018), slide 10.

Choose tools and libraries: Select appropriate tools and libraries to be used for webscraping. Popular choices are Python libraries such as BeautifulSoup, Scrapy, and Selenium. Sometimes large-scale oriented webcrawlers, such as Apache Nuts, or other webscraping platforms, such as Octoparse or ParseHub may be valid options. Using a headless browser is another option to be considered.

Identify some target websites: Create a short list of websites from which you intend to scrape data. This list will be used to test the code. Also, ensure that the webscraping activities comply with the website's terms of service and legal regulations.

4.2.2 Actual scraping:

Set up data collection: Write code to send HTTP requests to the target website's servers. Use web-scraping libraries to parse the HTML content and either store the complete page or extract the relevant data and store it. Test the code on the short list of target websites.

Handle authentication and session management: If the website requires authentication or session management (e.g., cookies), implement the necessary procedures in your webscraping code.

Understand website structure: For specific scraping, analyse the structure of the target websites. Identify the HTML elements containing the data you want to scrape, such as divs, tables, or specific CSS classes.

Handle pagination and navigation: If the target website has multiple pages or a complex navigation structure, develop code to navigate through pages and collect data from all relevant sections.

Error handling: Implement error handling mechanisms to address issues such as connection errors, timeouts, or changes in the website's structure.

Data storage: Decide how and where the scraped data is stored. Common options include (local) databases, CSV or JSON files, or cloud storage services.

Testing and validation: Test your webscraping code thoroughly to ensure it collects accurate and complete data. Make sure to test the code on a large list of URLs. Validate the results against manual checks if possible; use a random sample for that.

4.2.3 After initial scraping

Monitoring and maintenance: Set up monitoring systems to detect changes in the website's structure that may break the scraping code. Regularly update and maintain your code as needed.

Feedback and iteration: Ask for help if needed, be open to feedback, and continuously improve the scraping methodology to adapt to changes in website structures or data requirements.

Documentation: Document the scraping methodology used, including the websites scraped, the data collected, and the code used. This documentation aids in transparency and reproducibility.

4.2.4 Legal and ethical considerations:

Rate limiting and politeness: Implement rate limiting and polite scraping practices to avoid overloading the website's server with requests. Respect robots.txt files when applicable.

Legal compliance: Ensure that your webscraping activities comply with relevant legal regulations, such as copyright laws, data protection laws (e.g., GDPR), and anti-bot policies.

Ethical considerations: Respect the website's terms of service and terms of use. Avoid scraping sensitive or personal data, and be aware of legal and ethical constraints.

Data privacy and security: Protect any data you collect, especially if it contains personally identifiable information (PII). Secure your data storage and access.

4.3 Validation: Dealing with over- and under-coverage

4.3.1 Over-coverage

It is possible that the list of websites to be scraped contains websites for units not belonging to the target population. Examples of these are websites of persons or non-profit organizations while studying businesses. Another example are so-called ‘ghost’ vacancies while studying OJA’s. How one exactly deals with these units depends on the goal of the study.

In OJA’s, there is over-coverage when some advertised vacancies are out of scope for purposes of official statistics (ESSnet BD I WP8, 2018). It is important to identify those advertisements as early as possible in the process. The same holds for non-business websites when, for instance, one is studying enterprise characteristics. Looking for generic features specific to enterprise websites, such as contact information and a description of the business is useful here (Barcaroli et al. 2016, Daas et al. 2022).

4.3.2 Under-coverage

When considering scraping websites, we should be aware of the fact that not all enterprises may be present on the web. Particular types of business (branches) and, especially, small enterprises and self-employed persons may not be represented well online. Therefore, there is the problem of under-coverage of particular enterprises in webscraped data.

A number of OJA-data studies have been performed in the ESSnet BD II to study the under-coverage of advertisements (ESSnet BD II WPB, 2018). Three different approaches have been studied: i) micro-level comparisons, ii) aggregate comparisons, and iii) measuring the use of advertising channels via the Job Vacancy Survey (JVS). The outcomes revealed that, for the first approach, micro-unit comparison is very challenging; linking at the micro-level is messy and difficult. This was mostly caused by the difficulty in deriving the correct business unit for which the advertisement was collected. Comparison at the aggregate level was found to be much easier and those findings looked promising. The biggest concern was deriving the NACE code for the advertisements; private portals were found to usually have their own taxonomies, which are often only approximately comparable with NACE. The estimates were found to be close to the JVS-based estimates at the economic activity section and regional levels. This was independently confirmed by the work of Pedroza et al (2019) and Lovaglio et al. (2020). The third approach involved surveying enterprises and asking specific questions about their advertising channels. Here, it was found that large businesses are more likely to advertise online compared to smaller businesses.

An example of using multiple sources to deal with under-coverage, including webscraping, is described by Young and Jacobsen (2021). In this study, potential under-coverage of the business register was assessed by using two combined frames and employing capture-recapture methods. The units in the business register and those in a constructed webscraped dataset were combined by linking the units included in both. Because record linkage was conducted prior to drawing samples, the sample design incorporated information from records i) only in the business register, ii) only in the webscraped frame, and iii) in both frames. After drawing the sample and conducting the survey, a composite estimator was applied that allowed full use of the overlap design and the sample information to produce survey-based estimates. The findings demonstrated that making use of a combined list seemed a valuable addition but the overall improvement could not be accurately determined in the study.

Obtaining a list of all websites that need to be scraped to (as completely as possible) cover the population studied has been indicated to be one of the most important issues to solve (ESSnet BDII WPB, 2018). Under coverage is much less of an issue when the Business Register is used as the starting point (ESSnet BD II WPC 2019, Daas and van der Doef 2020).

4.4 Validation: Deduplication of units

Certainly, when multiple data sources are used, data from the same websites/units/objects may become included more than once. This could be the result of URL redirection (see section 3.3), multiple businesses with the same website (such as franchises), or the fact that the same text, such as a job advertisement, occurs on multiple websites. It is recommended to deal with these duplications as they may negatively affect the findings obtained (Stateva et al. 2020).

Deduplication is an important and challenging issue for OJA's. As mentioned before, one and the same job vacancy can be published in several different places on the web. So it is necessary that adds for the same vacancy and workplace are identified and that all duplications are removed. An important way to deal with this issue is pre-processing of the texts and combining the cleaned OJAs data with administrative sources, e.g. job vacancies data from the Public Employment Agencies at the national level (if available). Although the OJA-data may still be unstructured at this stage, as a consequence of the pre-processing steps, the data will become more and more structured. This process is not linear and contains many cycles as findings at the end of the chain may initiate the need to improve some, or a number, of the previous

step (ESSnet BD II WPB, 2019). To make it even more complex, sometimes an OJA can contain adds for multiple jobs. Each of them needs to be extracted and checked for duplicates.

For scraped webpages, duplicates can also be a major concern, especially when machine learning-based models are produced. When duplicate pages occur in both the training and test set, the metrics used to indicate the performance of the model will be overrated. Hence it is recommended to check and deduplicate the combined training and test set (for the features) included prior to model development.

4.5 Validation: Concept drift detection

When a model is developed to measure a specific concept it is important to regularly check if the model is "still measuring what it is supposed to be measuring". This is especially relevant when concepts are indirectly measured (Daas 2023). Any decrease in the accuracy (or any other metric) is usually described as concept drift, although model degradation is probably a better description (Daas and Jansen 2020).

Concept drift can be generally described as "a phenomenon in which the statistical properties of a target domain change over time in an arbitrary way" (Lu et al. 2019, p. 2347). Formalized, this implies that $P_t(X, y) \neq P_{t+1}(X, y)$, where $P(X, y)$ is some joint probability distribution of feature vector X and labels y at time point t. Subscript t refers to the time point right before the joint distribution of $P(X, y)$ changes. Subscript $t + 1$ then denotes the time point at which the joint distribution of $P(X, y)$ has changed, and concept drift has arisen. For the reader, it might be interesting to know that in essence three probabilistic forms of concept can be defined (see for more details the paper of Lu et al. 2019), but these are somewhat less important for the remainder of this paragraph.

Let's start by assuming one has developed a model, based on webscraped data collected at time point t, with a particular accuracy; $Accuracy_t$. If all is done well, the accuracy is obtained from an independent test set, e.g. using data not included while training the model, and the accuracy value is the mean (μ) of a 1000 independent test sets; hence $N = 1000$. This indicates that the standard deviation (s) can be determined for this large set of accuracy values. When one wants to check if - at a certain point in time - model degradation has occurred, the simplest approach is to scrape the complete set of websites used for model development (i.e. training set + test set) at the new time point. This results in data for $t + 1$, which needs to be processed in exactly the same way as the data collected at the previous time point. Next, the model, trained on data of time point t, is applied to the newly collected and processed data collected of $t + 1$. Assuming a normal distribution of the accuracy values, concept drift/model degradation has occurred when the metric of choice (accuracy in our case) at $t + 1$ deteriorates. This means that, for the single accuracy measurement of the $t + 1$ dataset, one has to determine the probability that the new value belongs to the distribution of values (of size N) observed at t. Because only lower accuracies are a concern and μ and s are the result of 1000 measurements, this means that when the following statement is true:

$$Accuracy_{t+1} < \mu - 1.96 \cdot s$$

it is highly unlikely that the accuracy value found is from the same distribution; this makes it likely that concept drift has occurred.

A more detailed concept drift analysis can be performed by individually checking the outcomes for the identical set of websites, i.e. those successfully scraped at both t and $t + 1$. For classifications, the results

of the websites included in both datasets can be compared in a confusion matrix. This enables one to derive multiple metrics and obtain additional insights on the difference observed.

4.6 Modelling and interpretation: Correcting for model-induced bias

Depending on the topic studied, the application of machine learning (ML) based models may result in a biased estimate (ESSnet BD II 2020, section 6.6). From the work of Wang et al. (2023) it becomes clear that carefully trained ML models do not have to suffer from bias; provided that sufficient auxiliary variables are included in the data and model. However, this is not always possible. We will start by discussing model-induced bias in the context of a binary classification.

4.6.1 Model-induced bias: Binary classification

Here, bias correction is discussed with the most simple example of a model-based approach in mind, that of a binary classification model. An example of this is the detection of certain types of businesses based on website texts. For this task, a model is going to be developed that, based on the texts on a website, aims to determine if the business is either a positive or a negative case.

When texts are used as input, it is fairly standard nowadays to make use of ML algorithms. These algorithms are trained to obtain a classification model that performs the task as well as possible. Training is commonly done on a dataset for which the classification outcome is known. This is known as supervised learning. From such a dataset, a random sample is drawn - often around 80% - on which the model is trained. The ultimate goal of training is to produce a model that is able to identify the positive and negative examples included as accurately as possible. How well the model actually performs is independently determined by applying it to the remaining, unselected, examples; this is the so-called test set (in our example composed of the remaining 20%). During training, a specific metric is chosen on which the classification by the model is optimized. Let's, for the sake of simplicity, choose Accuracy in this example. The latter means that the number of correctly classified positive and negative cases of the total number of cases classified is used to determine how well the model performs. For a binary classification task, the well-known confusion matrix (Table 4.6.1) is usually used to obtain the values needed. A confusion matrix contains the correctly classified cases, i.e. the so-called True Positives (TP) and True Negatives (TN), and the erroneously classified cases, e.g. the so-called False Positives (FP, Type I errors) and False Negatives (FN, Type II errors).

Table 4.6.1. Confusion matrix for a binary classification task

		Predicted class	
		Positive (1)	Negative (0)
Actual class	Positive (1)	*True Positive*	*False Negative*
	Negative (0)	*False Positive*	*True Negative*

Based on the confusion matrix, Accuracy is calculated as:

$$Accuracy = \frac{(True\ Positve + True\ Negative)}{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)} \quad (2).$$

The ultimate goal is to develop a classification model of the highest Accuracy possible. However, it is to be expected that not all cases will be correctly classified, which means that a number of False Positive and False Negative cases remain. When the number of False Positives is not equal to the number of False Negatives, this will negatively affect the model-based findings. This phenomenon has been described as *misclassification bias* (Meertens 2021, Chap. 1). In addition, if the proportion of positive cases in the training set (and obviously also in the test set) differs from the proportion of these cases in 'real world data' (on which the model will be eventually applied), this will also introduce a bias (Puts and Daas 2021). This bias is described as *proportion bias* in this document. Both biases are discussed in more detail below.

Misclassification bias

Dealing with the misclassification bias essentially focuses on correcting for differences in the numbers of False Positives and False Negatives cases produced by the model. This, for instance, occurred in the innovation detection model described by Daas and van der Doef (2020). In order to correct for this bias, estimates of the algorithm's (mis)classification probabilities are needed. These can be obtained from the confusion matrix results of the test set. Here, it is assumed that the misclassifications are independent across objects and that the (mis)classification probabilities are the same for each object, conditional on their true class label. Also, it is assumed that the test set used is a representative sample of the target population (more on that below).

With this classification-error model in mind, the probability that the algorithm predicts an object of class 1 correctly (TP) is denoted by p_{11} and the other probabilities TN, FP, and FN are defined analogously; i.e. as p_{00}, 1-p_{11}, and 1-p_{00}, respectively. The classification probabilities p_{00} and p_{11} are not known, but can be estimated by using the results for the test set. For a test set of size n, with the notation shown in Table 4.6.2, the classification probabilities are then estimated by: $\hat{p}_{00} = n_{00}/n_{0+}$ and $\hat{p}_{11} = n_{11}/n_{1+}$. A similar notation is used for the target population of size N.

Table 4.6.2. Confusion matrices for the test set (left) and target population (right).

		Predicted class		
		1	0	*Total*
Actual class	1	n_{11}	n_{10}	n_{1+}
	0	n_{01}	n_{00}	n_{0+}
	Total	n_{+1}	n_{+0}	n

		Predicted class		
		1	0	*Total*
Actual class	1	N_{11}	N_{10}	N_{1+}
	0	N_{01}	N_{00}	N_{0+}
	Total	N_{+1}	N_{+0}	N

Furthermore, the base rate for the target population ($\propto$), which is the proportion of data points for which the observed class is equal to 1, is defined formally as: $\propto = N_{1+}/N$. From this, it follows that the base rate can be estimated from the test set as: $\hat{\propto} = n_{1+}/n$. Hence, correcting for the misclassification bias is composed of determining i) the base rate of the model and ii) the bias caused by the number of false positives and false negatives produced. Both can be estimated from the test set results.

From the work of Meertens (2020, Chap. 3) it becomes clear that the misclassification bias can best be corrected for by using the calibrated base rate ($\hat{\alpha}_c$):

$$\hat{\alpha}_c = \hat{\alpha}^* \frac{n_{11}}{n_{+1}} (1 - \hat{\alpha}^*) \frac{n_{10}}{n_{+0}} \quad (3)$$

The required 'classify-and-count' estimator ($\hat{\alpha}^*$) in (3) is obtained by:

$$\hat{\alpha}^* = (n_{1+}/n)\ (n_{00}/n_{0+} + n_{11}/n_{1+} - 2) + (1 - n_{00}/n_{0+}) \quad (4)$$

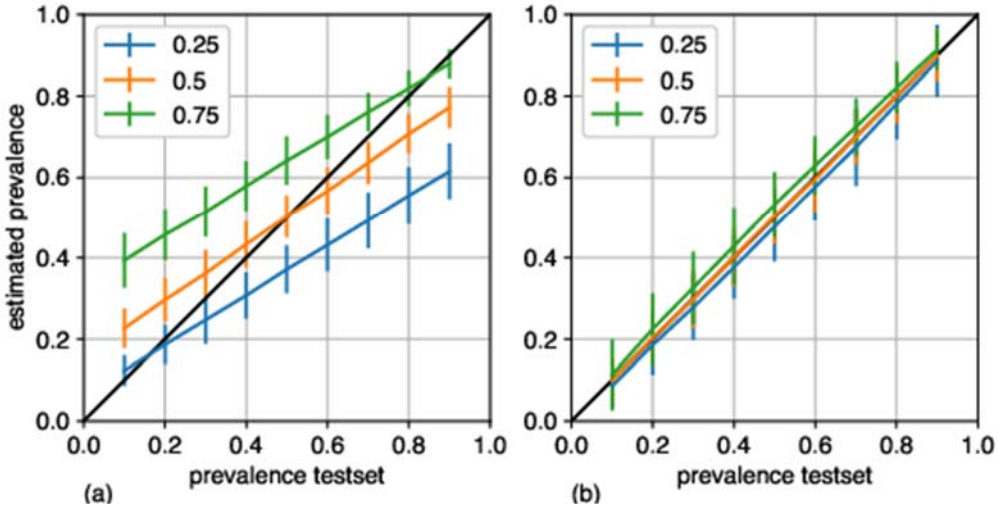
The reader is referred to Meertens (2020) for more details.

Proportion bias

Additional classification model research revealed that binary classifiers trained on a certain proportion of positive items introduced a bias when applied to data sets with different proportions of positive items. Since the latter is usually not known, this topic was initially studied with simulated data. The effect of this difference is illustrated in Figure 4.6. From this figure, it becomes clear that a solution had to be developed. Hence, the challenge was to solve this issue based on the limited information available in the test set.

The bias correction method developed is described in Puts and Daas (2021) and makes use of the *probability distribution* for the positive and negative cases in the data set used for model development. Hence, the correction method can *only* be applied to classification models that are able to produce *probability* estimates as output. Examples of these are Logistic regression, SVM, Randomforest, and Neural Networks (MPLC). Here, the starting point is a data set ($\mathbf{T}$) where X is the domain of all feature vectors, and the outcome of the model is a probability of being a positive case with features $\vec{x}$ ($\vec{x} \in X$).

Figure 4.6. Estimates of the proportion positives in simulated data for (a) models trained on 25% (blue), 50% (orange), and 75% (green) positives against their true proportion and (b) after applying the correction method developed.



Using Bayes theorem, it becomes apparent that:

$$B(\vec{x}) = P(+|\vec{x}, \mathbf{T}) = \frac{P(\vec{x}|+,\mathbf{T})P(+|\mathbf{T})}{P(\vec{x}|\mathbf{T})} \quad (5)$$

Note that all probabilities are conditional on $\mathbf{T}$. When $\mathbf{T}$ is representative for the population, $P(+|\vec{x}, \mathbf{T})$ should also be representative; i.e. the model should have learned the right features. However, it is not known if the proportion of positive items, $P(+|\mathbf{T})$, in the dataset is representative. Hence, the question is 'how to get a good estimate for the proportion of positive items?' Let the unknown proportion positives in the data be denoted by π, then the probability of a score b given the proportion π given $\mathbf{T}$ is:

$$P(b|\pi, \mathbf{T}) = \pi P(b|+, \mathbf{T}) + (1 - \pi)P(b|-, \mathbf{T}). \quad (6)$$

Bayes' law gives the probability of π:

$$P(\pi|b, \mathbf{T}) = \frac{P(b|\pi, \mathbf{T})P(\pi)}{P(b|\mathbf{T})} \quad (7).$$

Assuming $P(\pi)$ is uniform and $P(b|\mathbf{T})$ is a normalization constant, this can be formulated as a likelihood function, which, over the complete data set ($\mathbf{B} \subset \{b(\vec{x})|\vec{x} \in X\}$), is equal to:

$$L(\pi|\mathbf{B}, \mathbf{T}) = \prod_{b\in\mathbf{B}} P(b|\pi, \mathbf{T}) \quad (8)$$

From this, the maximum likelihood estimate for π is:

$$\hat{\pi} = argmax_{\pi}\, L(\pi|\mathbf{B}, \mathbf{T}) \quad (9).$$

The reader is referred to Puts and Daas (2021) for more details. The code is available on Github (Puts 2023). The method described in this section also corrects the misclassification bias mentioned in the previous section.

4.6.2 Other selectivity-based correction approaches

The methodology applied above is specifically developed to correct for model-induced bias in a binary classification context. However, there are other (potential) causes of bias (Groves and Lyberg 2010) that, for a large part, can be solved by the traditional bias (selectivity) correction methods used in official statistics. Certainly when survey data forms an important part of the data used, (standard) survey methodology can be applied. It stands to reason that this methodology need not be discussed in much detail here. Readers are referred to Särndall et al. (1992), Beręsewicz et al. (2018), and two Makswell project reports (2020a, 2020b) for more details on those views. However, the reader also needs to realize, that when Big Data is used information needs to be reliably extracted from the latter source to enable the application of the survey-based methodology. The reader is referred to section 3.4 for a specific overview when web-data is used and section 6.4 of the ESSnet BD II project (Del. WPK, 2020) for a more general overview. It suffices here to discuss a number of interesting examples observed during the literature review. Since the composition of the units for which website data is collected may differ from that of the target population, there is a need to apply correction methods that (aim to) solve these issues. Certainly when the differences are related to the target variable studied.

In the study of Daas and van der Doef (2020), three important correction methods were applied to determine the number of large innovative companies based on the text on the main page of their website. First, the model-based classification results were corrected for the bias resulting from the model developed. Here, it was observed that the number of False Positives (Type I errors) and the number of False Negatives (Type II errors) of the model were unequal, which negatively affected the estimate of the number of innovative companies; see Meertens (2021) for more details. Correcting this imbalance resulted in an increase in the number of large innovative companies. Next, a correction was applied for

websites that contained less than 20 words after processing. Since it was found that model-based classification results were unreliable when less than 20 words remained, these pages could not be classified by the model. A detailed study of those webpages revealed that no particular types of businesses dominated this (sub)group. Hence, it was assumed that the of ratio innovative and non-innovative businesses for this subset was similar to those already classified. This also increased the estimate. Finally, the number of innovative companies without a website was determined; the Community Innovation Survey-based results already indicated that such a group existed. Here, it was found that 0.1% of the innovative businesses included had no website, correcting for it slightly increased the estimate. In the end, the estimated number of large innovative businesses in the Netherlands based on the survey and on website texts were found to be nearly identical; they were 19,916 ±680 (survey) and 19,276 ±190 (web) respectively (Daas and van der Doe, 2020). The latter finding suggests that the most important causes of bias were identified and appropriately dealt with.

OJA work found that correction at the micro-level is challenging. This was mainly caused by the difficulty of linking the webscraped job advertisements to the relevant businesses. This obviously limited the success of any correction methods applied.

An interesting example in the context of selectivity is described in a German study on the location of businesses with a website (Thonipara and Haefner 2023). This study used the fact that a business has a website as a proxy for the degree of digitalization of the respective business. The work revealed that having a website is highly correlated with the location of the business. Businesses located in urban areas were almost twice as likely to have a website compared to those located in rural areas. They therefore suggested including the NACE code of businesses in any future website based studies.

4.7 Dealing with very dynamic population changes

Nowadays, for the Consumer Price Index (CPI), webscraped prices of particular or ranges of products are used by many National Statistical Institutes (Belchev and Lamboray 2021). Meaningful price indices can only be computed when the products used are homogeneous (Chessa 2016). However, especially for data on supermarket products - whether collected via the web or obtained as scanner data - it has been observed that the composition of (the population of) these products is very dynamic. Use of these products and their properties can be hampered by the occurrence of so-called “relaunches”. The latter refers to barcode changes of products repositioned in the market. However, often a relaunch caused barcode change does not actually indicate an ‘actual’ change of a product. Hence, in many cases, the characteristics of a product, such as its composition and use, have not changed. However, the fact that the barcode has changed means that the former code (which disappeared) and new (reintroduced) code must be related to correctly capture any price changes of that particular ‘product’.

Chessa has studied how (the effect of) the relaunches of products affect the CPI calculation for the Netherlands. In this work, it is suggested to focus on forming so-called homogeneous groups of products, e.g. a group of relaunched products (Chessa 2016). The latter can be achieved by using retailers’ product codes (Stock Keeping Units) and/or through item characteristics and may need expert involvement. In his 2021 paper, Chessa (2021) describes a method that groups ‘barcodes’ into strata (‘products’) by balancing two measures: an explained variance (R squared) measure for the ‘homogeneity’ of the ‘barcodes’ within products, while the second expresses the degree to which products can be ‘matched’ over time with respect to a comparing period. The resulting product ‘match adjusted R-squared’ (MARS) combines explained variance in product prices with product match over time, so that different stratification

schemes can be ranked according to the combined measure. The MARS method uses the proportion of explained variance in product prices, relative to the total variance in item prices, as a measure of product homogeneity. The contribution of each product (k) or item (i) is weighted by the quantities sold. This yields the following weighted R-squared measure:

$$R_t^k = \frac{\sum_{k \in K} q_t^k \left(\bar{p}_t^k - \bar{p}_t\right)^2}{\sum_{i \in G_t} q_{i,t} \left(p_{i,t} - \bar{p}_t\right)^2}$$

Where k is the product, K the set of products, t the period (month), G the group of items, q_t^k denotes the number of items sold for product k in month t and $q_{i,t}$ denotes the items i sold in that month. The price of an item i sold in month t is denoted by $p_{i,t}$ and $\bar{p}_t^k$ denotes the unit value for product k in that month. Note that $R_t^k = 0$ when all items are combined into one product and $R_t^k = 1$ when each item is a separate product. The MARS method has been applied to a broad range of product types and demonstrated positive results (Chessa 2021).

When there are many relaunches of products, the price changes of strata should preferably be used for the relevant products. By using quantity weights for strata of products, direct indices can be calculated, based on prices and quantities of the products in each stratum, with respect to the base month. To better deal with these changes, a new index method was introduced; the so-called Geary-Khamis method (Chessa 2016). This method is more adequately suited to deal with the effects of including relaunched products. The Geary-Khamis method helps track price changes while considering product replacements or relaunches. It calculates average prices in a way that ensures fair comparisons over time and regions, even when product availability is heterogeneous. It improves the accuracy of the price index and ensures that shifts in prices closer reflect actual changes rather than technical changes to product listings.

4.8 Filtering methods for Online Job Advertisement data

Online Job Advertisement (OJA) data is available on the web and can be used to obtain information on the job market. The paper of Beręsewicz and Pater (2021) provides an overview of the findings of a large pan-European study on inferring job vacancy statistics from large amounts of OJA's. This was promising and challenging as the OJA-based results significantly differed from the official job vacancies for the countries studied. The study recommends taking a subsample of the online collected data, for instance by filtering out a part of the advertisements, so the remainder will be (more) suitable to predict job vacancies. The findings indicate issues with the representativeness of the OJA data and try to deal with them.

Representativity of OJA data was specifically investigated in a separate study (Napierala et al. 2022). Here, OJA data was compared to two external data sources: the Labour Force Survey (LFS) and the Job Vacancies Survey (JVS), which are available in most EU countries. The conclusion of this study was the following:

> "… this comparison suggest that both the sources of information used as a benchmark for carrying out this evaluation [LFS and JVS] also have their weaknesses. The reliability of information collected in surveys is heavily dependent on response rates, which in turn rely on various factors (e.g. willingness to participate in a survey). Moreover, the methodology of JVS is not homogenous across Member States, which also impedes comparisons. Lastly, the analysis indicates that none of the three sources of information allows precise estimates on the number of vacancies to be derived." (from Napierala et al. 2022, p. 4).

This is an interesting conclusion indeed. The reader should realize that the units compared in the study of Napierala et al. (2022) are the vacancies and online job adds and that these units are not the basis for the sampling procedures applied in the LFS (households/persons) and the JVS (enterprises), respectively.

OJA data can be of particular interest when studying specific topics, such as the skills or prior experience required for specific jobs or the emergence of new skills. For such and other applications, filtering methods have also been suggested (Napierala 2023, Napierala et al. 2022). Here, subsets of OJA data are obtained by either selecting job adds with specific words, such as data scientist, or removing specific adds, such as apprenticeship or training opportunity adds. The latter are, from a statistical and labour law perspective, not defined as a vacancy (Beręsewicz and Pater 2021).

5 Discussion

This document provides an overview of the methodology of using webscraped data for official statistics. This process starts with defining a target population and ends with producing statistics based on web-data. The most important methodological issues identified are related to sampling, causes of bias, and methods specifically developed for dealing with web-data in the statistical process. Each topic is described in a separate chapter.

The sampling chapter (Chap. 2) discusses issues relevant to the need to obtain a subset of URLs and indicates the pros and cons of sampling-based approaches. One of the topics discussed is the difference between probability and non-probability-based approaches. The former has the advantage that the inclusion weights of the units are known while this is certainly not the case for the latter. Methods for dealing with non-probability samples usually tend to 'derive' weights for the units included by obtaining information from other data sources or by using (expert-based) assumptions. It is a topic of current research (Boyd et al. 2023, Golini and Righi 2024). Another topic discussed in this chapter is bulk vs. selective scraping. Selective scraping has the major benefit of reducing the amount of work required but must ensure that data for a representative part of the target population (for the topic studied) is collected.

Bias is a major concern when using webscraped data. Chapter 3 provides an overview of each step in the statistical process that this data goes through. Apart from methodological issues, any technical considerations that affect the latter are additionally discussed. This results in an overview of each step in the webscraping process and all the relevant methodological considerations in relation to their effect on the bias of the estimates obtained. A more detailed overview can be found in the recently published paper of Daas and Maślankowski (2023).

Chapter 4 gives an overview of methods that were specifically developed for dealing with the peculiarities of using (large amounts of) web-based data in the production of official statistics. A literature study was performed that aimed to obtain an as complete as possible overview of such specific methods. Each method is described, in a condensed way, focused on the typical problem solved with references to the relevant literature. It results in a unique overview that cannot be found in any other document to date.

From the above, it becomes clear that webscraping is a very interesting way to produce (business) statistics. Here, webscraping:

I. stimulated the development of web-specific legal and ethical regulations
II. enables the collection of (large amounts of) data that can be used to improve the business register, improve the quality of already existing business statistics, and produce statistics on new topics.
III. does not put any response burden on the businesses involved (apart from the low burden on their website, of course)
IV. has the advantage that the data collection step is - for a large part - under control of the researcher/NSI.
V. stimulates the development of a new methodology
VI. when performed routinely, requires the need for a dedicated webscraping infrastructure.

From the above, it is clear that webscraped data provides some major advantages for NSIs, specifically for Business-related statistics. Webscraping not only reduces the response burden but has the additional advantage that, for many of the business units of interest, more recent data can be collected at regular intervals. This is expected to not only improve the quality of the statistics produced but may also stimulate the development of new statistics (at relatively low additional costs). Webscraping has a great future indeed!

References

Anglin, K.L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. Journal of Research on Educational Effectiveness 12(4), 685-706. DOI: 10.1080/19345747.2019.1654576

Baker, R., Brick, J., Bates, N., Battaglia, M., Couper, M., Dever, J., Gile, K., Tourangeau, R. (2013). Report of the AAPOR Task Force on Nonprobability Sampling. AAPOR report, June 2013. Link: https://aapor.org/wp-content/uploads/2022/11/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Barcaroli, G., Scannapieco, M., Summa, D. (2016). On the use of Internet as a Data Source for Official Statistics: A strategy for Identifying Enterprises on the Web. *Rivista Italiana di Economia Demografia e Statistica* (RIEDS) 70(4), 25-41. Link: https://ideas.repec.org/a/ite/iteeco/160401.html

Belchev, P., Lamboray, C. (2021). How to start with web scraping in the HICP: Evidence from EU member states. Paper for the online meeting of the Group of Experts on Consumer Price Indices, UNECE, June. Link: https://unece.org/sites/default/files/2021-05/Session_2_Eurostat_Paper.pdf

Beręsewicz, M., Lethonen, R., Reis, F., Di Consiglio, L., Karlberg, M. (2018). An overview of methods for treating selectivity in big data sources. Statistical Working Paper, Eurostat. Link: https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf

Beręsewicz, M., Pater, R. (2021). Inferring job vacancies from online job advertisements. Statistical Working Papers, Eurostat, Luxembourg. http://data.europa.eu/doi/10.2785/96387

Boyd, R.J., Powney, G.D., Pescott, O.L. (2023). We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, 38(6), 521-531. DOI: 10.1016/j.tree.2023.01.001

Cavallo, A. Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives* 30(2), 151-178. Link: https://www.nber.org/system/files/working_papers/w22111/w22111.pdf

Chessa, A.G. (2016). Processing scanner data in the Dutch CPI: A new methodology and first experiences. Paper for the 2016 meeting of the Group of Experts on Consumer Price Indices, UNECE. Link: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_1._Netherlands_Processing_scanner_data_in_the_Dutch_CPI.pdf

Chessa, A.G. (2021). A Product Match Adjusted R Squared Method for Defining Products with Transaction Data. *Journal of Official Statistics* 37(2), 411-423. DOI: 10.2478/jos-2021-0018

Daas, P. (2023). Big Data in Official Statistics. Inaugural speech at the Eindhoven University of Technology, The Netherlands, May 26. Link: https://pure.tue.nl/ws/portalfiles/portal/296764797/Rede_Daas_26_5_2023.pdf

Daas, P., Hassink, W., Klijs, B. (2024). On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms. *Journal of Official Statistics* 40(1), 190-211. DOI: 10.1177/0282423X241235265

Daas, P., Jansen, J. (2020). Model degradation in web derived text-based models. Paper for the 3rd International Conference on Advanced Research Methods and Analytics (CARMA), 77-84. Doi: 10.4995/CARMA2020.2020.11560

Daas, P., Maślankowski, J. (2023). Current challenges in the use of web data as a source for official statistics. Case study on data collection. *Wiadomości Statystyczne* (The Polish Statistician) 68(12), 49-64. DOI: 10.59139/ws.2023.12.3

Daas, P., Tennekes, M., De Miguel, B., De Miguel, M., Santamarina, V., Carausu, F. (2022). Web intelligence for measuring emerging economic trends: the drone industry. Statistical Working Papers, June, Eurostat. Link: https://ec.europa.eu/eurostat/documents/3888793/14722798/KS-TC-22-004-EN-N.pdf

Daas, P.J.H., van der Doef, S. (2020). Detecting Innovative Companies via their Website. Statistical Journal of IAOS 36(4), 1239-1251. DOI: 10.3233/SJI-200627

Delden, A. van, Windmeijer, D., ten Bosch, O. (2019). Finding enterprise websites. Discussion paper, Statistics Netherlands, the Netherlands. Link: https://www.cbs.nl/en-gb/background/2020/01/searching-for-business-websites

Elliott, M.R., Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science* 32, 49–264.

ESSnet BD II WPB (2018). Methodological framework for processing online job adverts data for Official Statistics V2 . Workpackage B Deliverable B3, ESSnet Big Data II. Link: https://cros.ec.europa.eu/group/31/files/1558/download

ESSnet BD II WPC (2019). Reference Methodological Framework for processing online based enterprise characteristics (OBEC) data for Official Statistics V.1. Workpackage C Deliverable C2, ESSnet Big Data II. Link: https://cros.ec.europa.eu/group/31/files/1565/download

ESSnet BD II WPF (2019). BREAL: Big Data REferenc e Architecture and Layers, Business Layer. Workpackage F Deliverable F1, ESSnet Big Data II. Link: https://cros.ec.europa.eu/group/31/files/1590/download

ESSnet BD II WPK (2020). Revised version of the methodological report. Workpackage K Deliverable 9, ESSnet Big Data II. Link: https://cros.ec.europa.eu/group/31/files/1635/download

ESSnet WIN (2022). URL finding methodology. Joint Workpackage 2 & 3 deliverable, ESSnet Trusted Smart Statistics. Link: https://cros.ec.europa.eu/system/files/2023-12/20220131_url_finding_methodology.pdf

ESSnet WIN (2023). Landscaping of Websites for Webscraping with Focus on Selection Modes. ESSnet WIN, Draft version of WP4 report, Oct. 2023.

ESSnet WIN (2024). Quality Guidelines for acquiring and using web scraped data. ESSnet WIN, Deliverable 4.5, version, Version, Oct. 2024.

Foederer, J. (2023). Should we trust web scraped data? Paper on Arxiv. Link:10.48550/arXiv.2308.02231

Golini, N., Righi, P. (2024). Integrating probability and big non-probability samples data to produce Official Statistics. *Stat. Methods Appl*. 33, 555–580. DOI:10.1007/s10260-023-00740-y

Greenaway, M. (2017). Better Scraping, Better Statistics?: Using web-scraped data in statistical outputs. Presentation at the Governmental Statistics Service (GSS) conference, 22-23 Nov., Sheffield, UK. Link: https://analysisfunction.civilservice.gov.uk/wp-content/uploads/2018/01/Better-Scraping-Better-Statistics-1.pdf

Groves, R.M., Lyberg, L. (2010). Total Survey Error: Past, Present, and Future, *Public Opinion Quarterly* 74(5), 849-879. DOI: 10.1093/poq/nfq065

Han, S., Anderson, C.K. (2021). Web Scraping for Hospitality Research: Overview, Opportunities, and Implications. Cornell Hospitality Quarterly. 62(1), 89-104. DOI: 10.1177/19389655209735

IANA (2022). Hypertext Transfer Protocol (HTTP) Status Code Registry. Webpage of the Internet Assigned Numbers Authority. Link: https://www.iana.org/assignments/http-status-codes/http-status-codes.xhtml

Lee, S., Valliant, R. (2009), Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research* 37, pp. 319–343.

Lovaglio, P.G., Mezzanzanica, M., Colombo, E. (2020). Comparing time series characteristics of official and web job vacancy data. Quality & Quantity 54(1), 85-98, DOI: 10.1007/s11135-019-00940-3

Lu, J., Liu, A., Dong, F., Gu, F. Gama, J., Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering* 31(12), 2346-2363. Doi: 10.1109/TKDE.2018.2876857

Makswell project (2020a). Methodological aspects of using big-data. Deliverable 2.2. Link: https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.2.pdf.

Makswell project (2020b). Report on identification of future research needs in terms of statistical methodologies and new data. Deliverable 2.3. Link: https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.3.pdf

Maślankowski, J., Daas, P. (2023). Current Challenges and Possible Solutions for the Use of Web Data as a Source for Official Statistics. Presentation at the Methodology of Statistical Research 2023 (MET2023) conference, 3-5 July, Warsaw, Poland. Link: https://met2023.stat.gov.pl/Content/Presentations/Sesja%2016.4%20MET2023.pdf

Maybir, J., Chapman, B. (2021). Web scraping of ecstasy user reports as a novel tool for detecting drug market trends. *Forensic Science International: Digital Investigation* 37, 301172. DOI: 10.1016/j.fsidi.2021.301172

Marconi, G. (2022). Content removal bias in web scraped data: A solution applied to real estate ads. Open Economics 5, 30-42. DOI: 0.1515/openec-2022-0119

Meertens, Q.M. (2021). Misclassification Bias in Statistical Learning. PhD-thesis University of Amsterdam, Amsterdam, the Netherlands. Link: https://pure.uva.nl/ws/files/59712159/Thesis.pdf

Muehlethaler, C., Albert , R. (2021). Collecting data on textiles from the internet using web crawling and web scraping tools. *Forensic Science International* 322(1), 110753. DOI: 10.1016/j.forsciint.2021.110753

Napierala, J. (2023). The Feasibility of using Online Job Advertisements in Analysing Unmet EU Demand. Luxembourg: Publications Office. Cedefop working paper No 18. http://data.europa.eu/doi/10.2801/10233

Napierala, J.; Kvetan, V., Branka, J. (2022). Assessing the representativeness of online job advertisements. Luxembourg: Publications Office. Cedefop working paper No 17. http://data.europa.eu/doi/10.2801/807500

Oostrom L.A.N., Walker A.N., Staats B., Slootbeek-Van Laar M., Ortega-Azurduy S., Rooijakkers, B. (2016). Measuring the internet economy in The Netherlands: a big data analysis. Discussion paper 2016-14, Statistics Netherlands, The Hague/Heerlen, The Netherlands. Link: https://www.cbs.nl/-/media/_pdf/2016/40/measuring-the-internet-economy.pdf.

Pedroza, P., de, Visintin, S., Tijdens, K., Kismihók, G. (2019). Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data. IZA Journal of Labour Economics 8(1). DOI: 10.2478/izajole-2019-0004

Puts, M. (2023). BayesCCal: Bayesian Calibration of classifiers. Code available on Github. Link: https://github.com/mputs/BayesCCal

Puts, M.J.H., Daas, P.J.H.. (2021). Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach. Paper for the 2021 Symposium on Data Science and Statistics, Machine Learning session, online. Link: https://arxiv.org/abs/2102.08659

Särndal, C.-E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. New York: Springer.

Scannapieco, M. (2018). WP2 Web Scraping Enterprise Characteristics. Presentation at the Big Data for European Statistics meeting, 14-5 May, Sofia, Bulgaria. Link: https://cros-legacy.ec.europa.eu/sites/default/files/BDES_2018_WP2_Presentation.pdfspecific

Stateva, G., Saucy, F., Lucarelli, A., Wu, D., Maślankowski, J., Dumesnil de Maricourt, C., Grahonja, Č., Špeh, T. (2020). Methodological framework for processing online job adverts data for Official Statistics. Workpackage B Implementation Online Job Vacancies, Deleverable. ESSnet BD II, 18 March.

Templ, M., Meindl, B., Kowarik, A., Dupriez, O. (2017). Simulation of Synthetic Complex Data: The r Package simPop. *Journal of Statistical Software* 79(10), 1–38. DOI: 10.18637/jss.v079.i10

Ten Bosch, O., Windmeijer, D. (2014). On the Use of Internet Robots for official Statistics, UNECE MSIS conference, Dublin, 2014. Link: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf

Ten Bosch, O., Windmeijer, D., van Delden, A., van den Heuvel, G. (2018). Web scraping meets survey design: combining forces. Paper for the BigSurv18 conference, 25-27 Oct., Barcelona, Spain. Link: https://www.bigsurv.org/bigsurv18/uploads/73/61/20180820_BigSurv_WebscrapingMeetsSurveyDesign.pdf

Ten Bosch, O., Kowarik, A., Quaresma, S., Salgado, D., van Delden, A. (2024). Statistical scraping: informed plough begets finer crops. Paper for the Q2024 conference, Estoril, Portugal. Link: https://drive.google.com/file/d/1Blxlb80Vth4ono1gdgY289t4_FMxB-J9/view , page 61.

Thonipara, A., Haefner, L. (2023). Digital divide, craft firms' websites and urban-rural disparities—empirical evidence from a web-scraping approach. Review of Regional Research 43, 69-99. DOI: 10.1007/s10037-022-00170-5

UNECE (2015). Guidelines on Statistical Business Registers. United Nations Economic Commission for Europe report, New York & Geneva. Link: https://unece.org/fileadmin/DAM/stats/publications/2015/ECE_CES_39_WEB.pdf

Van den Brakel, J. (2019). New data sources and inference methods for official statistics. Statistics Netherlands Discussion paper, July 2019, Heerlen, The Netherlands. Link: https://www.cbs.nl/-/media/_pdf/2019/27/newdataofficialstatisticsdp.pdf

Vehovar, V., Toepoel, V., Steinmetz, S. (2016). Non-probability Sampling. In: Wolf, C., Joye, D., Smith, T.W., Fu, Y-C. (eds), *The SAGE Handbook of Survey Methodology*, Chapter 22. Sage Publications Ltd., Londen, UK. pp. 329-345. Link: https://methods.sagepub.com/book/the-sage-handbook-of-survey-methodology/i2461.xml

Wang, R., Chaudhari, P., Davatzikos, C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. PNAS USA 120(6), e2211613120. DOI: 10.1073/pnas.2211613120

Wikipedia (2023). List of official business registers. Wikipedia page. Link: https://en.wikipedia.org/wiki/List_of_official_business_registers

Wu, C. (2022). Statistical Inference with Non-probability Survey Samples. *Survey Meth.* 48(2), 283-311.

Young, L.J., Jacobsen, M. (2021). Sample Design and Estimation When Using a Web-Scraped List Frame and Capture-Recapture Methods. *Journal of Agricultural, Biological, and Environmental Statistics* 27(2), 261-279. DOI: 10.1007/s13253-021-00476-w