Selective scraping, sampling and other methods to minimize known causes of biases of web data

Web Intelligence Network Conference

Alexander Kowarik, Piet Daas

05 February 2025

Trusted Smart Statistics – Web Intelligence Network

Web Intelligence Network

Funded by the European Union

Overview

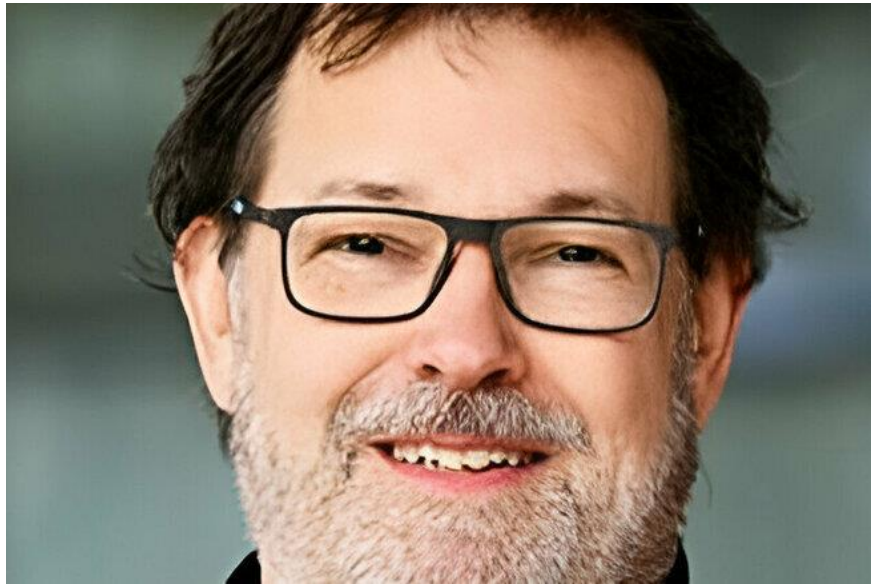
- Sampling in the Context of Webscraped Statistics
- Methods specific to webscraped data and causes of bias

All deliverables of WP4 at https://github.com/WebIntelligenceNetwork/Deliverables

- Co-financed by
  - Web Intelligence Network: 101035829 — 2020-PL-SmartStat
- Contributions to deliverables by several colleagues:
- Olav ten Bosch, Jacek Maslankowski, Magdalena Six, Johannes Gussenbauer, Sonia Quaresma and more

In Memoriam: Prof.dr. Piet Daas

- Methodology lead and
- Main author of “Deliverable 4.6: WP4 Methodology report on using webscraped data” on which this presentation is based.



Sampling – what for

- Sampling for Quality Assessment
- Estimation:
  - Probability and Non-Probability Sampling
    - Methodology for estimation and error estimation very well developed and we do know sampling methodology
  - Selective Scraping
    - Optimized Scraping Strategy

Funded by the European Union

Sampling for Quality Assessment

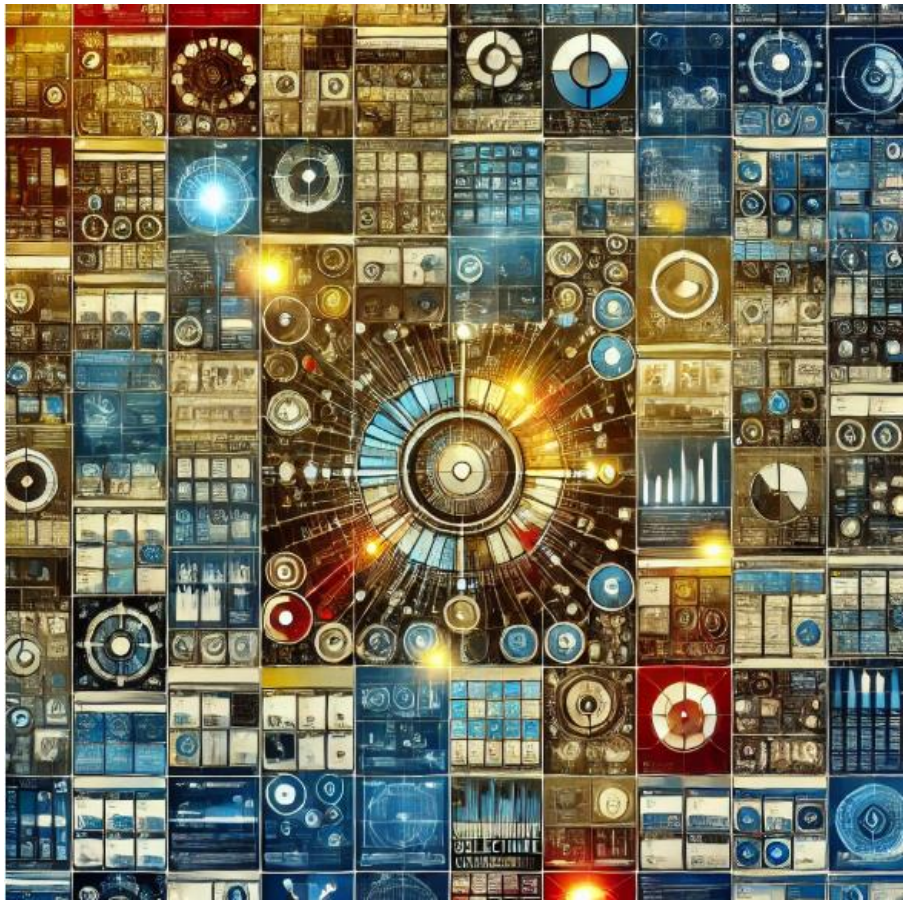
- Why Sampling Matters in Quality Assessment:
  - Labor-intensive nature of manual annotation.
  - Need for high-quality, representative annotated datasets.

- Optimization Strategies
  - Reducing annotation volume with strategic sampling.
  - Ensuring representative marginal distributions.

More on this in the deliverable

Probability Sampling

- Probability sampling if the process of deriving a target variable, is not easily scalable
  - e.g. a statistical classification needs costly manual intervention



- 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.

Funded by the European Union

Non-Probability Sampling

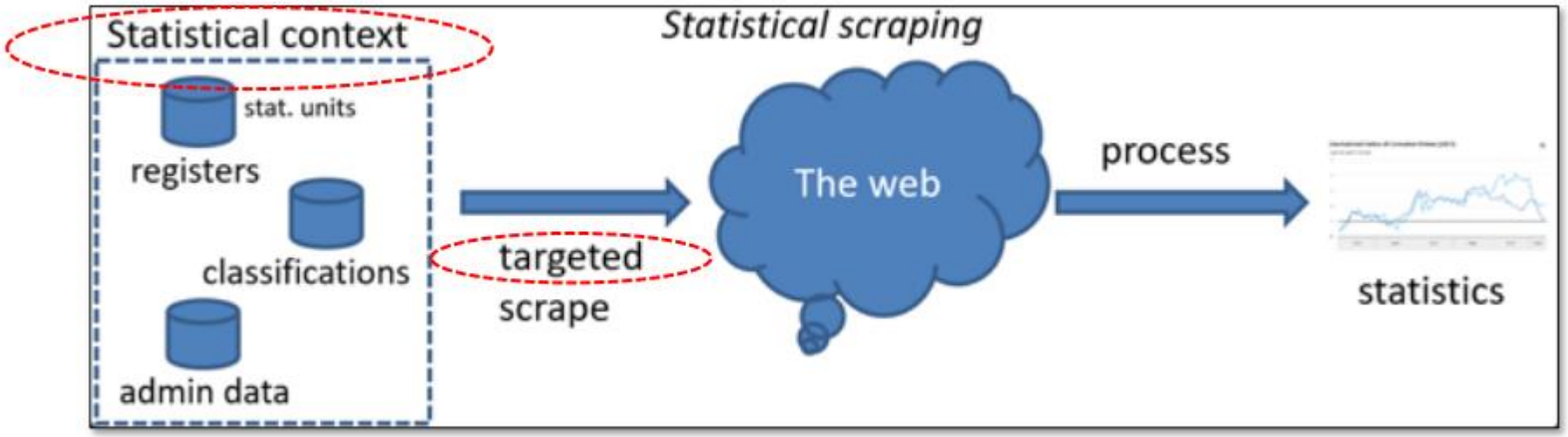
- Web-collected data can also be considered as a non-probability sample if
  - each unit in the target population has an unknown (and unequal) chance of being included in the data set, but it can be explained well enough with know auxiliary information
- This bias introduced by the unequal inclusion probabilities needs to be fixed.

- A wide set of methods exists as extension to probability sampling methodology, e.g.:
  - Calibration / Weighting
  - Statistical Matching (combining a probability sample with a NPS)
  - Propensity Score Matching

Funded by the European Union

Selective (=Statistical) Scraping - Definition

Def 1.1: Statistical scraping is the use of online data starting from a-priori information in the respective statistical domain keeping a clear relation with the statistical context



Olav ten Bosch, Alexander Kowarik, Sonia Quaresma, David Salgado, Arnout van Delden, (2024), *Statistical scraping: informed plough begets finer crops*, European Conference on Quality in Official Statistics, Estoril, Portugal

Selective Scraping …

- Selective scraping = deliberately scraping a subset of the target population
- unlike so-called bulk scraping, which collects large volumes of data on a given subject
- selective scraping uses pre-existing knowledge of the subject to collect specific information on specific units in a more controlled manner
- **Goal: Collecting data of a representative set of units for the topic**

Selective Scraping – Process and Example

Job vacancies in a certain business sector

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).

Source Identification

- Find all units in the business register in the business sector.
- Perform URL finding for them

Source Selection

- Draw a random sample of the collected URLs

Data Extraction and Enhancement

- Scrape the sites and identify and classify job vacancies.

Estimation

- **Apply appropriate estimation methodology, e.g. GREG estimation to compensate for the random selection and the non-probability selection due to the URL finding**

Concept Drift is especially prominent in web data -> 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
- Any decrease in the accuracy (or any other metric) is usually described as concept drift, although model degradation is probably a better description
- **One solution:** Repeatedly scrape and classify identical set of websites and compare over time

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

Correcting for Model-Induced Bias: Binary Classification

- **Misclassification Bias**: Different rates of False Positives (FP) and False Negatives (FN) may lead to distorted estimates.
- **Proportion Bias**: If the ratio of positive to negative cases in the training dataset does not match real-world data, incorrect predictions arise.

- **Example:**
- **Base Rate Estimation**: The proportion of positive cases in the target population must be correctly determined to avoid bias, e.g. by a test set.
- **Calibration of Base Rate**: An adjusted estimate (calibrated base rate) can be derived.

Funded by the European Union

Conclusion

- Big challenges in using web data for proper statistical conclusion
- Proper methodology for design, validation and estimation is needed
  - Quick and dirty is only a solution for experiments

All deliverables at https://github.com/WebIntelligenceNetwork/Deliverables →