New Avenues with Web Intelligence – Essnet - WP3

WEBINAR: 2023-11-23

Petrus Munter, Data Department – Registry Unit - SCB
Remy Kamali, Data Department – Development Unit - SCB

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

BACKGROUND

- Work package 3 of the ESSnet Trusted Smart Statistics – Web Intelligence Network project (ESSnet TSS-WIN)

- Use Case 3 is about acquiring additional knowledge based on the content of online sources. Extracting data from enterprise websites to get characteristics is at the center of the project.

What is the name of the institution you are representing, and in which country?

OUTLINE

PART 1. 50 min:

1. Web Scraping
2. Scanner data

3. Break 10 min

PART 2. 40 min:

1. Project goals
2. Where are we at the moment
3. Combined sources
4. Last Project Year

5. Questions/Discussion 15 min

Funded by the European Union

Web Scraping

- Automated data collection through the world wide web
- Implemented in the production for several years
- CPI

Funded by
the European Union

Web Scraping - *Scraping data*

- Imacros (HTML)
- IT ← → Statisticians

Moving forward:

- Python
  - Pandas
  - Beautifulsoup
  - Requests(API)

What code-language and modules/packages are your preferred web scraping tools?

Web Scraping - Methodology CPI

- Uniform samples
- Equal weights
- x most popular plus y more random from the most

Web Scraping - Challenges

- Technical knowledge required
- Person dependency
- Higher risk when problems arise

Web Scraping - Benifits

- Rapid implementation
- Handle large amounts of data
- Regression modeling possible
- More time for analysis
- Cost-effective

Funded by the European Union

Web Scraping - Internal Prerequisites

- Legal aspects
  - Terms of use
  - Internet in general
  - Obligation to provide information
- It – support
- Competency
- Implementing new solutions/languages

In which areas/departments, other than the CPI, do you consider web scraping to be a useful tool for collecting data?

Scanner Data

- Data directly from suppliers own cash register system
- Implemented in production since several years
- Main users: Consumer price index, turnover and food statistics

Funded by the European Union

Scanner data –
Data gathering

- Sftp
- API
- E-mail

Scanner data - methodology CPI

- Similar as before
- A lot more to come
- Larger samples
- Multilateral methods

Scanner data - challenges

- Loads of data
- Long period of negotiations
- Coordinating internal usage
  - CPI
  - Turnover in the service sector
  - Food statistics
- Vulnerable without official agreements

Scanner data - benifits

- Solid reliable data flow
- Loads of data
- Less delivery burden once set up

Scanner data - internal prerequisites

- Legal aspects
  - Several Surveys involved
- It – support
- Competency
- Tools to distribute data internally and externally
- Coordinate between different agencies

In which areas/departments do you consider scanner data to be a useful data source?

Use Case 3

Online prices of household appliances and audio-visual, photographic and information processing equipment

Web Intelligence Network

Funded by the European Union

GOALS – UC 3

- Collection of data regarding online prices of household appliances
- Exploratory!
- Different prerequisites for each attending member
- **Investigating the possibility of combining web scraped data with scanner data**

Funded by the European Union

Status

- Scanner data
  - Status/complications
- Web scraping
  - Status/complications

More knowledge → More value

- Combining sources
- Weights issues
- Comparing indexes
- Increase precision

Web Intelligence Network

Funded by the European Union

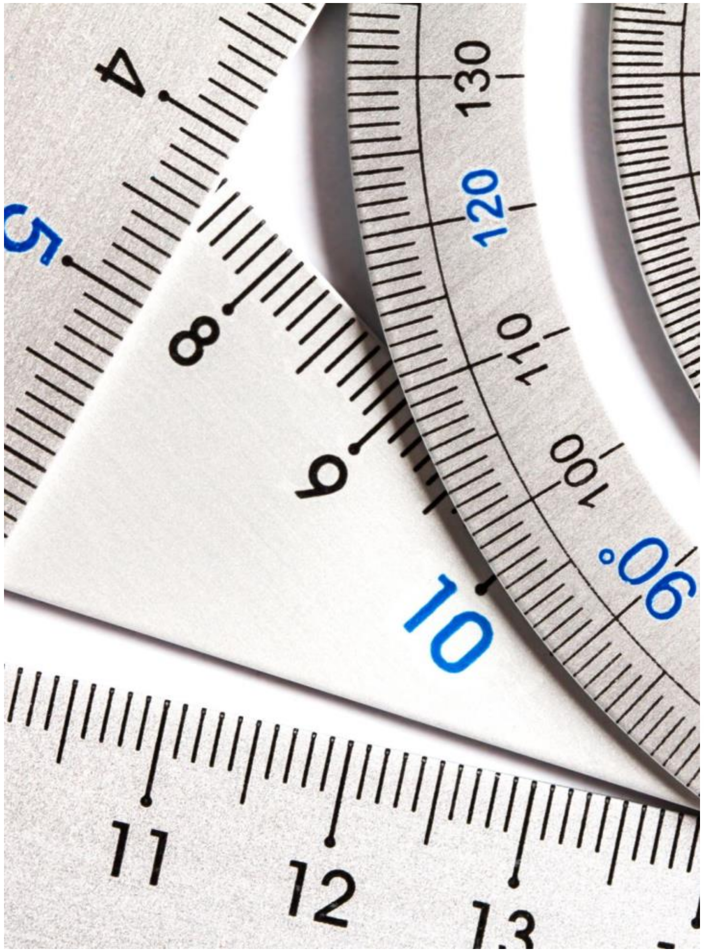
First look

- Elementary index
- Relations
- Maintaining flow of data
- Irregularities, trend but affected by other ways of sorting the data online

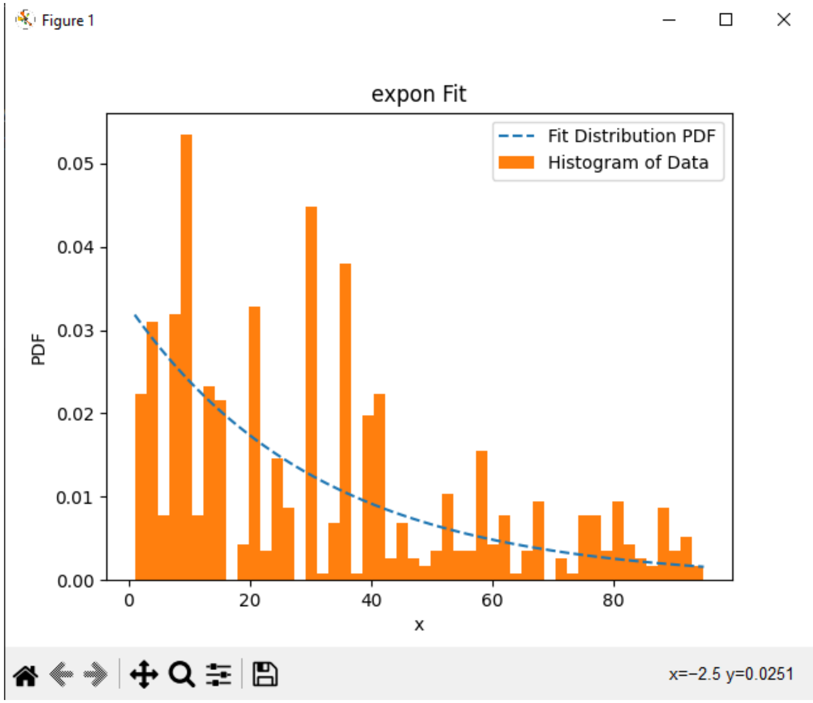
Funded by
the European Union

Last Project Year

- Estimating distributions of weights
- Applying different methodologies and compare to Scanner data results



”Breaks” in data



Web Intelligence Network

Funded by the European Union

FINAL THOUGHTS

Web Intelligence
Network

Funded by
the European Union

Takeaway

- Technical Prerequisites
- Coordinate usage
- Legal Challenges

Funded by the European Union

THANK YOU

Web Intelligence
Network

Funded by
the European Union