EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Updating of Statistical Register by Web Scrapping

Jaroslav Sixta

Vice President

Czech Statistical Office

Instituto Nacional de Estatística
Statistics Portugal

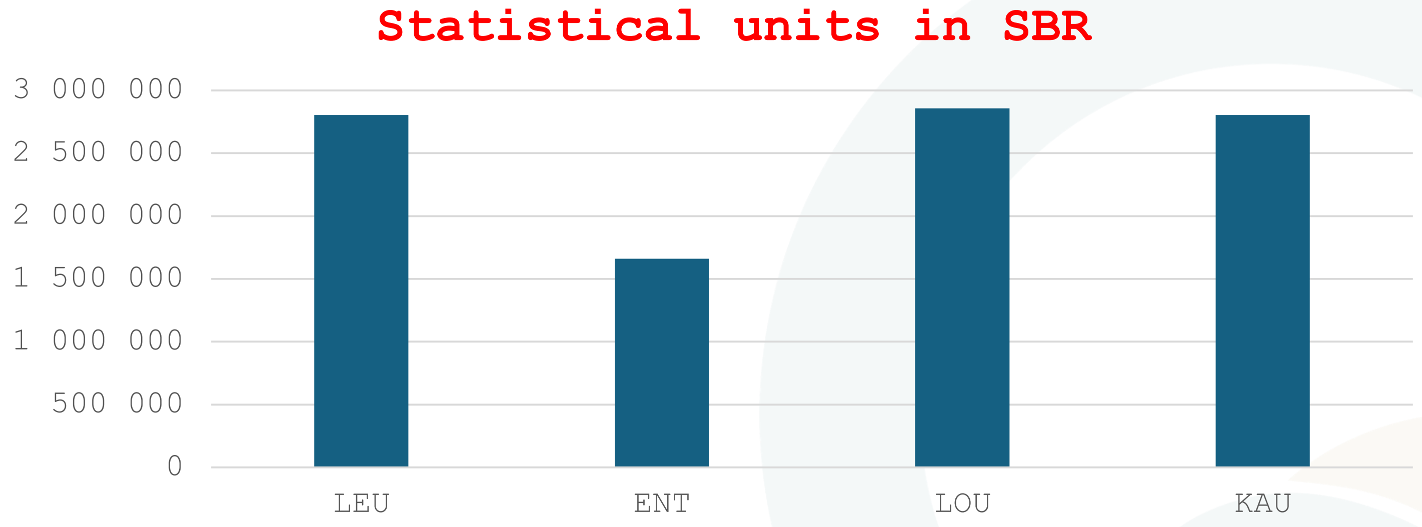
eurostat

The conference is
partly financed by the
European Union

CZECH STATISTICAL OFFICE

Statistical Business Register at the CZSO

- main backbone for statistical surveys
- production of register based statistics



- problems with NACE misclassification – project for updating SBR by web scraping started in 2023.

Instituto Nacional de Estatística
Statistics Portugal
eurostat
The conference is partly financed by the European Union

Main Reasons

- improve the quality of NACE characteristic in SBR
- get the new sources for detecting main economic activities (ENT webs and internet databases)
- decrease the response burden by substituting the questions on economic activities with machine learning
- replace the manual NACE assignment with automatic coding
- support the implementation of NACE revision at the CZSO.

The conference is
partly financed by the
European Union

The Cooperation with MODE Research

- insufficient experience of SBR staff with web scraping and machine learning (CZSO experts assigned to other tasks)
- cooperation with MODE Research was secured
- MODE Research = agency set up by the Faculty of Informatics and Statistics of the University of Economic and Businesses in Prague
- long-term experience with statistical processing and analysis, computational process automation, data mining and web scraping
- 2 experts assigned to the project.

Objectives

1. add the relevant web page to ENTs in SBR and scrap them to obtain information on economic activity
2. develop the model based on machine learning techniques to predict appropriate NACE codes
3. search for possibilities to use information on ENTs web pages and economic activity from internet databases
4. develop language model for automatic coding
5. train the internal CZSO staff in new techniques to secure sustainability of the project.

Sources of ENTs Web Pages

cz.nic | CZ DOMAIN REGISTRY

- association of internet providers
- provided us 2 mil. URLs
- identified 477.000 valid web pages (WP)
- linked 57.000 web pages to LEUs in SBR.

Firmy.cz

- internet database of companies in Czechia
- used Apify tool for web scraping
- identified 100.239 companies with WP
- linked 48.000 WP to LEUs in SBR.

NACE Prediction Model

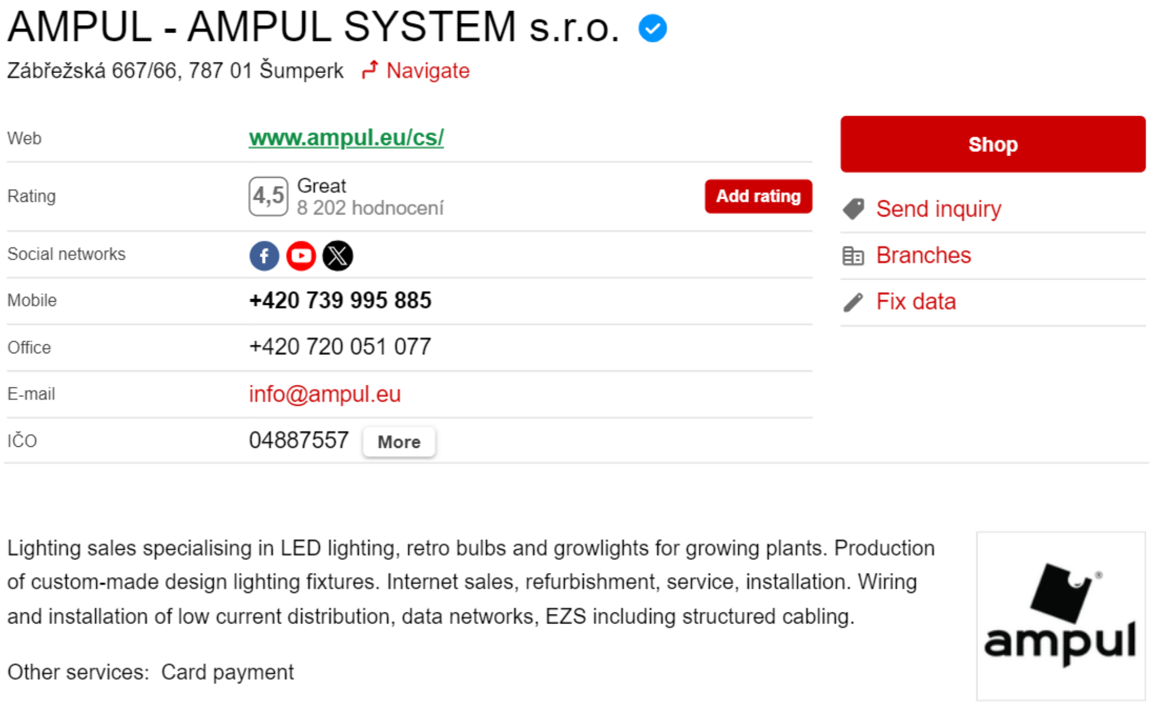
1) <u>defining the sample for verification</u>
   - LEUs with assigned WP were broken-down by employment, turnover, legal form and NACE
   - LEUs in sample have NACE confirmed from statistical survey (22.000 units).

2) <u>developing the model to predict NACE from ENTs WP</u>
   - 1st prototype uses random forest classifier with no hyperparameter tuning
   - tested on training sample of 15.000 units
   - 1st step is to predict correct NACE section - 56% accuracy at present
   - problem to feed model with appropriate words - intention to use NACE index in the future.

The conference is
partly financed by the
European Union

Using Economic Activities from Firmy.cz (1)

- the text description of economic activities from companies profiles was scraped



AMPUL - AMPUL SYSTEM s.r.o. ✓

Zábřežská 667/66, 787 01 Šumperk Navigate

Web	www.ampul.eu/cs/
Rating	4,5 Great 8 202 hodnocení Add rating
Social networks	f ▶ X
Mobile	**+420 739 995 885**
Office	+420 720 051 077
E-mail	info@ampul.eu
IČO	04887557 More

Shop

Send inquiry

Branches

Fix data

Lighting sales specialising in LED lighting, retro bulbs and growlights for growing plants. Production of custom-made design lighting fixtures. Internet sales, refurbishment, service, installation. Wiring and installation of low current distribution, data networks, EZS including structured cabling.

Other services: Card payment

Using Economic Activities from Firmy.cz (2)

- in total 276.726 profiles were scraped
- linking the data with SBR was based on ID numbers as the companies in firmy.cz are identified with the same unique ID as in SBR
- after exclusion of duplicities we identified 52.000 LEUs with textual description of economic activity.

Improving the Quality of SBR

- 6.500 statistically significant ENTs with unidentified NACE from administrative or statistical sources were selected from SBR and matched with scraped data from firmy.cz
- 1.500 ENTs were detected in both populations. For all these units the description of economic activities from Firmy.cz was converted into 4-digit NACE code by using LLM ChatGPT4
- The suggested NACE codes were manually checked by SBR administrators. They confirmed the correct code for 1.200 ENTs
- As a result the number of statistically significant ENTs with unidentified NACE was decreased by 18,5%.

The conference is
partly financed by the
European Union

Experience with Using LLM ChatGPT4

PROS

- very accurate – without additional learning was able to assign correct NACE for 80% of units
- cheap (total costs of our activities = 18$).

CONS

- expensive for large population of units - for more than 100.000 units the costs go to thousands dollars
- slow - allows processing only 150 requests per day.

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is
partly financed by the
European Union

Conclusion(1)

- at least one web page was assigned to 105.532 legal units in SBR – the searching for other sources of web pages still continues, possibility to buy these data from Duns & Bradstreet
- procedures for web scraping and text lemmatization were developed
- detection of NACE from web pages by using random forest classifier is currently tested
- focus on developing the appropriate dictionary to improve NACE prediction model

Conclusion (2)

- the data scraped from internet database firmy.cz helped us to decrease number of statistically significant units with unknown NACE code in SBR
- the same method is now used for checking the economic activities of LEUs that are expected to be misclassified in SBR (at present we verify the NACE codes for other 8.500 LEUs)
- the possibility to scrap data from other internet databases (e.g. yellow pages) is checked
- assigning NACE codes with using LLM ChatGPT4 is quite accurate but unfit for large populations
- as a result we are testing language model BERT, which is not so developed but can be improved by learning

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is
partly financed by the
European Union

Conclusion (3)

- BERT model is currently tested on textual descriptions of economic activities obtained from statistical surveys
- the possibility to use the project achievements for implementing smooth NACE revision is examined
- SBR staff is becoming trained in new techniques (some representatives of SBR are able to work with Python now)
- cooperation with other CZSO web scraping and machine learning activities to allow exchange of know-how was set up (e.g. representatives of economic surveys are part of project team).

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is
partly financed by the
European Union

THANK YOU FOR YOUR ATTENTION!