ROMÂNIA

INSTITUTUL NAŢIONAL DE STATISTICĂ

R for web scraping Online Job Adverts and computing potential indicators

Contributions:

Prof. Phd. Bogdan OANCEA, National Institute of Statistics Romania /University of Bucharest, Romania

Prof. Phd. Nicoleta CARAGEA National Institute of Statistics/Ecological University of Bucharest, Romania

Prof. Phd. Ciprian ALEXANDRU, National Institute of Statistics/Ecological University of Bucharest, Romania

Marian NECULA, National Institute of Statistics Romania /The Bucharest University of Economic Studies, Romania

Phd. Ana Maria CIUHU, National Institute of Statistics Romania

Phd. Raluca DRAGOESCU, National Institute of Statistics Romania

Outline

1. Introduction – ESSnet on Big Data II – WPB Online job vacancies
2. Dataset description
3. List of potential indicators
4. Results
5. Other methodological contributions.
6. Conclusions.
7. Future: Eurostat's WIN/WIH project

Introduction

ESSnet on Big Data II – WPB Online job vacancies:

- ESSnet on Big Data II – part II of Eurostat's response on how to tackle/integrate big data sources into officials statistics production systems.

- Comprised of 11 work packages, ESSnet on Big Data II tries to provide a common methodological framework for adopting Big Data sources.

- Big Data sources targeted by work packages range from mobile network data, to satellite images data, to web data.

- The underlying principle of the work carried inside ESSnet on Big Data II is unification of statistical production tools (methodologies, software architecture, quality assessment).

Dataset description

ESSnet on Big Data II – WPB Online job vacancies:

- For WPB Online job vacancies packages a great part of the work done was to provide a common methodological framework for analyzing a large dataset collected by CEDEFOP/TabulaeX/CRISP.

- The dataset consisted, at the beginning of the year 2020, on approximately 64 millions observations of online job advertisements collected from a large sample of online job advertisements aggregators across all EU countries (also UK).

- Observations were broke down by 49 variables (dates, website, ESCO codes, education level, NACE codes,city, region, country, etc.)

Dataset description

ESSnet on Big Data II – WPB Online job vacancies:

- Apparently, the collection process seems uneven between countries and periods.

- For some countries (e.g. NL) a significant number of ads were collected from foreign websites. (a new possible statistical indicator, if not already calculated, could be the number/percentage of foreign workers per economy/country)

- Job ads per category sector are in line with the "economical reality", as described by the three-sector model. (EUROSTAT, 2019)
- Job ads per experience required suggests that „up to 1 year" of experience is over represented in the total number of job ads.

- Job ads per esco_level_1 shows that job ads addressed to professionals, technicians and associate professionals are over represented. This seems logical taking into consideration point c) above.

- In relation to the case study „Romanian websites as data sources" the data generally follows the real economic trends for migrating workforce, except the information for some countries (i.e. Spain, Italy) for which the well known high provision of Romanian workforce is clearly under represented by ads. A possible explanation for this is that for these countries the supply of workforce is directed towards agriculture and manufacturing, correlated with point c)

List of potential indicators

List of potential indicators (non-exhaustive, small selection)

- Number of observations collected in reference period
- Number of observations collected in reference period relative to the category sector.
- Number of observations collected in reference period relative to educational level.
- Number of observations collected in reference period relative to ESCO level 1.

Results

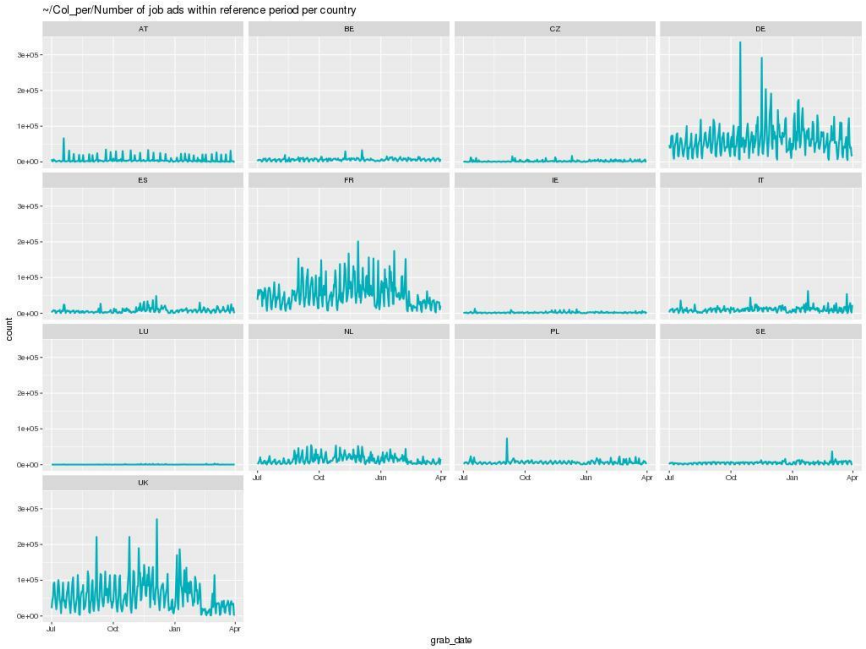

Figure 1. Number of observations collected in reference period.

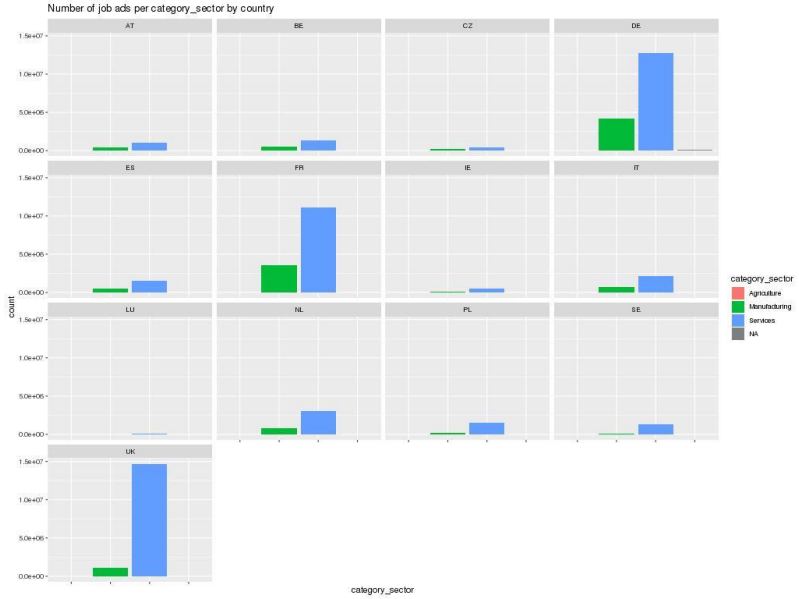

Figure 2. Number of observations collected in reference period relative to the category sector.

Results

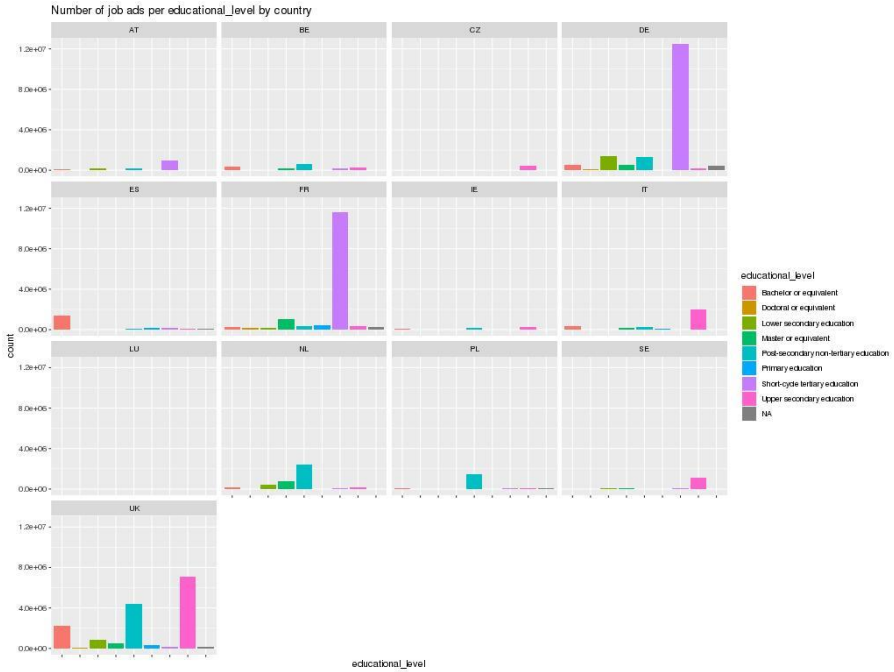

Figure 3. Number of observations collected in reference period relative to educational level.

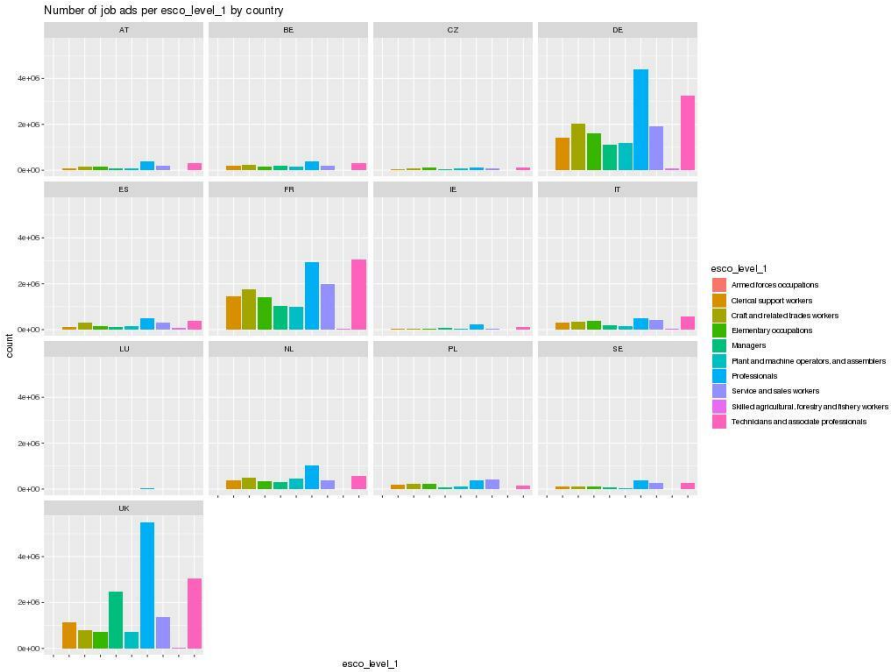

Figure 4. Number of observations collected in reference period relative to ESCO level 1.

Other methodological contributions.

Fitting discrete probability distributions to a national dataset combining jobs survey data with online job advertisements data. (Both dataset were aggregated by NACE codes)

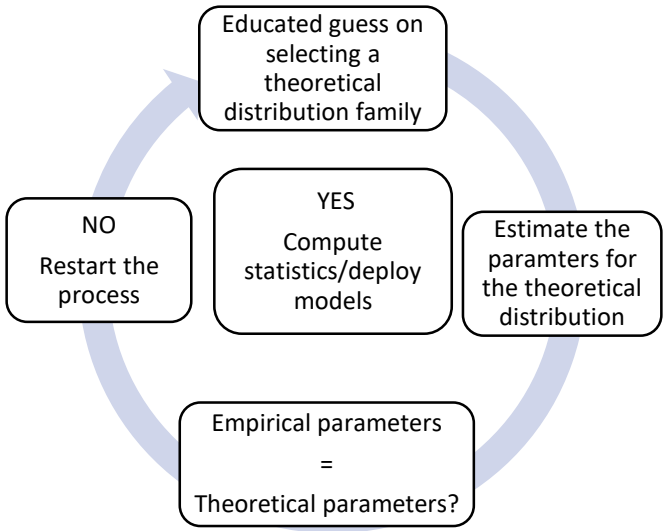

Figure 5. Description of distribution fitting process.

Other methodological contributions.

Main steps in MLE:

1. Make an educated guess about the probability distribution function (p.d.f.) which optimally fits the data.
2. Compute the likelihood function given by the formula:

$$\boldsymbol{L_n(\theta) = \prod_{i=1}^{n} f(x|\theta)},$$

Where

$$\boldsymbol{L_n(\theta; x) - likelihood\ function,}$$

$f(x|\theta)$- pdf

x – random variable or sample data,

$$\boldsymbol{\theta - paramter(s)\ of\ the\ p.d.f.}$$

Given the nature of this function in practice is simpler to compute the log-likelihood function, therefore the former can be expressed

$$\boldsymbol{ll(\theta; x) = \sum_{i=1}^{n} \log(f(x|\theta)),}$$

Where

$$\boldsymbol{ll(\theta) - log - likelihood\ function}$$

3. Estimate the parameter(s) θ which maximize $ll(\theta)$, formally

$$\boldsymbol{\hat{\theta} = \arg max\ ll(\theta; x)}$$

This is usually achieved by solving the Langrage multipliers system of equations.

Other methodological contributions.

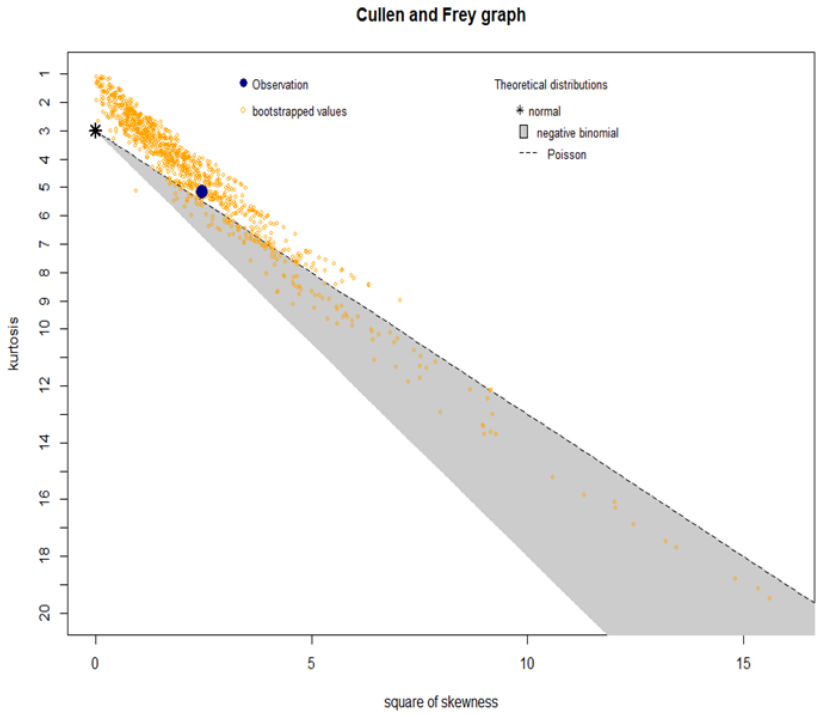

Figure 6. Cullen- Frey graph for job vacancies. Dataset: Romanian NIS Job Vacancies Survey.

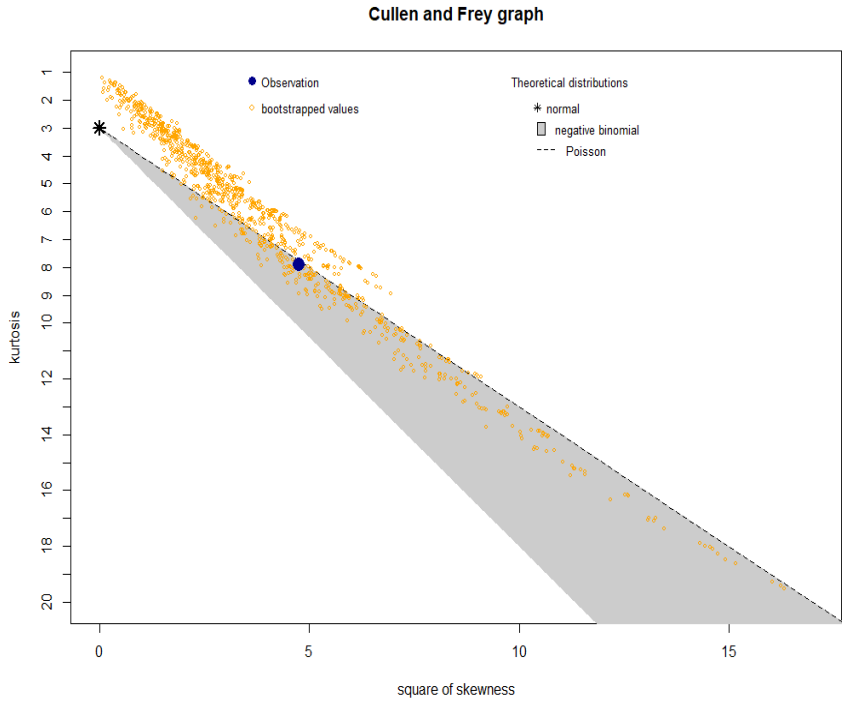

Figure 7. Cullen-Frey graph for job advertisements. Dataset: Romanian National Agency for Job Vacancies.

Other methodological contributions.

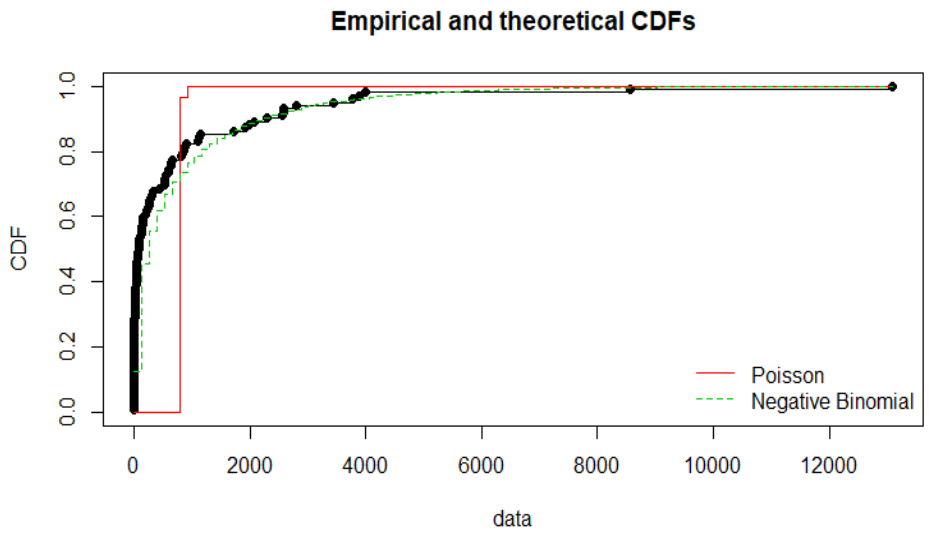

Figure 8. Empirical vs Theoretical distribution CDF.
Dataset: Romanian NIS Job Vacancies Survey.

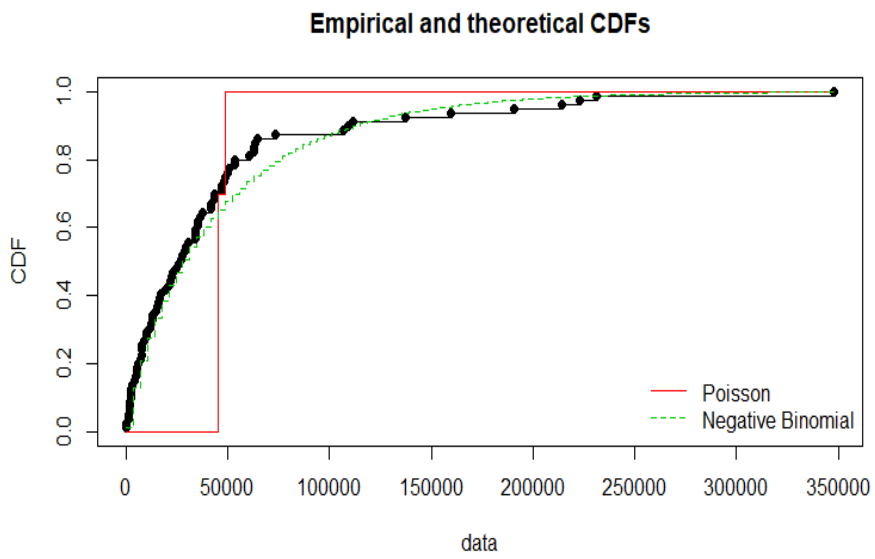

Figure 9. Empirical vs Theoretical distribution CDF.
Dataset: Romanian National Agency for Job Vacancies.

Conclusions

- We propose a possible set of variables seen as potential candidates for the convergence layer's output, if we work under the hourglass model assumptions.
- Other potential use of these variables is in a data forensics model to capture if something changed with the process which generates the data.
- New statistics can be developed by incorporating new data sources through statistical inference procedures, like empirical p.m.f./p.d.f. estimation.

Future: Eurostat's WIN/WIH project

- Eurostat will launch in 2021 a new project for developing a common statistical production architecture based on online data sources.
- Web Intelligence Network – a set of partners from statistical offices entrusted with expanding and developing an infrastructure for statistics based on Big Data.
- Web Intelligence Hub – a modern pipeline, based on cloud computing, were statistical offices can share and use common resources for producing statistics.