

**ESSnet Big Data II**

**Grant Agreement Number: 847375-2018-NL-BIGDATA**

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata>

[<https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en>](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)

**Workpackage WPB**

**Implementation – Online Job Vacancies**

**Methodological aspects related to CEDEFOP data used for OJA statistics V.1.**

**Prepared by:**

**Ciprian Alexandru, Marian Necula (INS, Romania)**

**Date of present version: 20 August 2020**

Workpackage Leader:

Tomaž Špeh (SURS, SI)

e-mail address: tomaz.speh@gov.si

mobile phone: +38651672116

Contents

[Purpose and scope 4](#_Toc48712550)

[CEDEFOP system 4](#_Toc48712551)

[Methodological considerations 6](#_Toc48712552)

[Duplicates 6](#_Toc48712553)

[Lower rate of OJVs in some countries 7](#_Toc48712554)

[Comparisons between OJVs and Job vacancy Statistics (JVS) 7](#_Toc48712555)

[Rescaling the data 7](#_Toc48712556)

[Improving the comparability of the two data sources 8](#_Toc48712557)

[Understanding the OJVs business models 8](#_Toc48712558)

[OJVs vs. OJAs in National Employment Agency 8](#_Toc48712559)

[Vacancies by ISCO major groups 9](#_Toc48712560)

[Conclusions 9](#_Toc48712561)

[References 10](#_Toc48712562)

List of Figures:

[Figure 1 GDP versus population by country 12](#_Toc48712172)

[Figure 2 OJVs versus GDP by country 12](#_Toc48712173)

[Figure 3 Scatterplot of OJVs predicted (or fitted) by the model (vertical axis) against observed OJVs for the 13 countries of the CEDFOP data (log scales). 13](#_Toc48712174)

[Figure 4 Predictions of the model 13](#_Toc48712175)

List of Tables:

[Tabel 1 Structure of CEDEFOP data 4](#_Toc48712492)

# Purpose and scope

The purpose of this material is to describe the methodological aspects, at the conceptual and practical level, regarding the use of data produced by CEDEFOP as an input for producing statistical indicators in the area of online job adverts (OJA), starting from pre-processed data. It addresses key conceptual, methodological, and technical aspects in using CEDEFOP data in OJA statistics.

Methodological considerations from this material consider the results obtained by the countries members of WPB ESSnet Big data II, in the implementation of using the CEDEFOP data for producing job vacancies statistics based on OJA[[1]](#footnote-1).

The Work Package B, Implementation - Online Job Vacancies, has established a close working relationship with the European Centre for the Development of Vocational Training (CEDEFOP) in order to use the collected or/and processed data as a source for producing statistical estimates.

CEDEFOP is creating the infrastructure to gather information from the most important online job vacancy portals (OJV portals) in real time across the EU, into the project *Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis*. [1]

The CEDEFOP data were presented in June 2019, at the physical meeting in Thessaloniki, and all countries have started to use these data and the first results have come during the month of July, August, and September 2019.

# CEDEFOP system

The final processed data are presented in structured mode, as a table with 49 variables, accessible by CEDEFOP Data LAB (Table 1). The Data Lab environments are deployed in the EC (European Commission) Data Platform. In order to authenticate, the users receive a EC Data Platform account, which they will be able to use in any environment deployed in that platform. In this environment there are more options available for exploring the data: R and Python scripts, and Jupyter notebooks.

Tabel Structure of CEDEFOP data

| variable | type | example values |
| --- | --- | --- |
| general\_id | numeric | 157361354, 157355438, 157460283 |
| grab\_date | date | 17891, 17892, 17893 |
| year\_grab\_date | numeric | 2018, 2018, 2018 |
| month\_grab\_date | numeric | 12, 12, 12 |
| day\_grab\_date | numeric | 26, 27, 28 |
| expire\_date | date | 18011, 18012, 18013 |
| year\_expire\_date | numeric | 2019, 2019, 2019 |
| month\_expire\_date | numeric | 4, 4, 4 |
| day\_expire\_date | numeric | 25, 26, 27 |
| lang | character | ro, en, lu |
| idesco\_level\_4 | character | 4321, 8322, 7126 |
| esco\_level\_4 | character | Stock clerks; Car, taxi and van drivers; Plumbers and pipe fitters |
| idesco\_level\_3 | character | 432, 832, 712 |
| esco\_level\_3 | character | Material-recording and transport clerks; Car, van and motorcycle drivers; Building finishers and related trades workers |
| idesco\_level\_2 | character | 43, 83, 71 |
| esco\_level\_2 | character | Numerical and material recording clerks; Drivers and mobile plant operators; Building and related trades workers, excluding electricians |
| idesco\_level\_1 | character | 4, 8, 7 |
| esco\_level\_1 | character | Clerical support workers; Plant and machine operators, and assemblers; Craft and related trades workers |
| idcity | character | NA, 179132, 95060 |
| city | character | NA, Municipiul Bucureşti, Municipiul Iaşi |
| idprovince | character | RO424, RO226, RO414 |
| province | character | Timiş, Vrancea, Olt |
| idregion | character | RO42, RO22, RO41 |
| region | character | Vest, Sud-Est, Sud-Vest Oltenia |
| idmacro\_region | character | RO4, RO2, RO4 |
| macro\_region | character | MACROREGIUNEA PATRU; MACROREGIUNEA DOI; MACROREGIUNEA PATRU |
| idcountry | character | RO |
| country | character | ROMÂNIA |
| idcontract | character | 1, 2, 3, 4 |
| contract | character | Internship; Permanent; Self Employment; Temporary |
| ideducational\_level | character | 1, 2, 3, 4, 5, 6, 7, 8 |
| educational\_level | character | Primary education; Lower secondary education; Post-secondary non-tertiary education; Upper secondary education; Short-cycle tertiary education; Bachelor or equivalent; Master or equivalent; Doctoral or equivalent |
| idsector | character | 1,..., 97, A,..., U |
| sector | character | Accommodation and food service activities (General); Activities auxiliary to financial services and insurance activities; ... |
| idmacro\_sector | character | A,..., U |
| macro\_sector | character | Manufacturing; Other service activities, ... |
| idcategory\_sector | character | 1, 2, 3 |
| category\_sector | character | Agriculture; Manufacturing; Services |
| idsalary | character | 1,..., 13 |
| salary | character | 0 - 6.000 EUR; ...; > 90.001 EUR |
| idworking\_hours | character | 1, 2 |
| working\_hours | character | Part time; Full time |
| idexperience | character | 1,..., 8 |
| experience | character | No experience; ..., Over 10 years |
| source\_category | character | Classified ads portal; Company; Education; Employers organization; Employment agency; Employment website; Job search engine; Online newspaper; Public employment service |
| sourcecountry | character | FR, RO, US |
| source | character | ADECCO; ADZUNA; AT\_AMS; AT\_CAREERJET;... |
| site | character | academics; accenture; accesa; ... |
| companyname | character | cms careers; cmv; cn group; ... |

# Methodological considerations

## Duplicates

Duplicates occur when advertisements are taken from another job portal with identical values. For duplicated jobs the ‘general\_id’ have the same value. Sometimes, duplicated rows are completely identical and sometimes, duplicated rows indicate the same job announcement coming from different sources. In this case, in order to identify and eliminate duplicated records should be used ‘general\_id’ variable.

## Lower rate of OJVs in some countries

There is a bias or an artefact in the collected data: only a part of the OJVs have been screened and the main issues could be:

1. The OJVs channel is not yet much developed in some countries
2. The difference reflects a peculiarity of the labour market in these countries
3. Others?

Suggestions for testing those potential explanations[[2]](#footnote-2):

a) Bias or artefact:

* plotting time series for all countries to detect anomalies
* compare the proportion of the scraped sources over time
* search for duplicates

b) OJVs channel:

* time series might display an increase in OJVs with time
* comparisons with other indicators of IT technologies and internet development among countries

c) labour market:

* comparisons
* with employment,
* unemployment,
* activity rates (proportion of persons in employment in the population)
* Job vacancy rates

## Comparisons between OJVs and Job vacancy Statistics (JVS)

A major issue when dealing with OJVs is to understand how Online Job Advertisements are connected to actual job vacancies, in particular to the official Job Vacancy Statistics. For this purpose, CEDEFOP data are in a first step compared to the Eurostat JVS downloaded from: <https://ec.europa.eu/eurostat/databrowser/view/tps00172/default/table?lang=en>.

Rescaling the data: In an attempt to bring the data on a common basis allowing to compare the two sources, the total amount of CEDEFOP OJAs (a cumulated sum over a *n* months period) are transformed to monthly averages, as follows:

m\_ojv = total CEDEFOP OJVs / *n*

On the contrary, the quarterly Eurostat JVS correspond to a snapshot taken at the reference day (the last day of the quarter). For the same period, we have data of *q* quarters. The absolute amount of JVS are averaged as follows:

m\_jvs = (JVSQ1 + JVSQ2 + ... + JVSQn) / *q*

Regarding the CEDEFOP data, due to the availability of daily online job vacancies in each month, the daily average number of online job vacancies should be considered in each month, as it was a data collection on a continuous basis. As technical approach, the data ingestion is developed as sequential process, on daily snapshot of web scraping, crawling and API collected data. This must take into consideration if we want to compare the OJAs data with JVS on a reference day, because on that specific day data will be collected only for certain websites.

Improving the comparability of the two data sources

The comparability of the two data sources, OJVs resulting from OJAs CEDEFOP data and the JVS, may be improved in several ways [2]:

1. subsetting the OJVs to the corresponding months for which JVS results are available
2. subsetting the OJVs data to job advertisements still open for the JVS reference days
3. subsetting the OJVs to job advertisements corresponding to the NACE2 categories covered by the official Eurostat JVS (sections B-C), or to the NACE2 branches actually covered for each country individually
4. comparing the monthly OJVs (as above subsetted) data with the corresponding month of the quarterly JVS

## Understanding the OJVs business models

A good understanding of the “business models” underlying the development of job portals, as well as of the market of online job advertisements is crucial [2]. Understanding of their mechanisms is a key issue to understand the data collected and processed by CEDEFOP system and identify the best fit model to use them for statistical purposes, direct on JVS or in correlation with other socio-economic indicators. Beware of that simple aspect, the Online job portals have not been developed to produce statistics or indicators of the job market. Their goal is different from a portal to portal, some of them are focused on human resource aspects, in finding the right person for a specific job, but other portals are focused only in making money from promoted adverts on their web pages.

## OJVs vs. OJAs in National Employment Agency

The web portals of the National Employment Agencies are well structured and contains verified information declared by the companies, and most important, with official classifications, like NACE and ESCO. For some countries there quite big difference between the CEDEFOP data collected from National Employment Agencies and the data obtained directly from these agencies, some possible reasons could be [3]:

* the basic concept is a job vacancy (Employment Agency), not job advertisement
* depends on the country’s legislation, if the JVs that employers submit to the Employment Agency are declared voluntary (i.e.: BG situation) or as a request of the law (i.e.: RO situation); employers prefer to advertise their job vacancies mostly in online job portals, as it is faster and more convenient for them.
* from the moment of the declaration of the job vacancy in the Regional Labor Office until it’s actual publication on the Employment Agency's Internet portal, there is a certain lag of time (i.e. we can't say that these JVs are actual in real time)

## Vacancies by ISCO major groups

The share of the total amount of online job vacancies according to ISCO major groups differs in the CEDEFOP-data from data collected by countries, from web scraping or from data in JVS. The differences may be related to the difficulties in de-duplicating. Occupations with the lower skill levels are often published on multiple job portals, problems with de-duplicating can therefore be expected to become more apparent in major groups 8 and 9. [5]

As a future development of the CEDEFOP system, could be take into consideration, to add a new variable with the value of the ingestion source: web scraping, crawling or API collected data. [6]

A cross tabulation of “source” and ’source\_category’ shows possible wrong classifications of job portals: for example, the job board ’stepstone’ is wrongly classified as „employment agency“ as well as the job board ‘Monster’ is wrongly classified as ‘employment website’ (example from Germany). [7]

# Conclusions

The job vacancy advertisements on job portals cannot be considered similar to the job vacancies as defined in the EU Regulations. The online job vacancies cover only those job vacancies, as defined by EU Regulation, for which the active steps carried out by the employers to find a suitable candidate include also the advertising on internet job portals, *only* if all job portals used by the employers are covered by the web scraping and other data ingestion activities. However, even if all the job portals were considered, it is not clear if all the job advertisements posted on all job portals refer to total job vacancies that still exist at a specific reference date. [4]

The CEDEFOP data collected as the European scale seem to offer a promising basis for developing pertinent statistics and economic indicators.

A preliminary analysis shows that should be explored existing relationship between OJVs, national and regional population sizes, and national economic activity expressed as GDP. (Annex 1)

Therefore, model predictions can be drawn, which could be compared to actual observations for the other European countries for which no CEDEFOP data are available. On the same basis, OJVs could offer a tool for developing flash estimates of GDP among EU countries, since OJAs are potentially real-time available. The models and choice of variables should be evaluated in more detail and improved, by each country, according also with the OJA representativity among the sector activities of the economy.

# References

[1] Cedefop (2019). The online job vacancy market in the EU: driving forces and emerging trends. Luxembourg: Publications Office. Cedefop research paper; No 72., <http://data.europa.eu/doi/10.2801/16675>[[3]](#footnote-3)

[2] Francis Saucy, Sophie Schmassmann, Alberto Columbano, (oct. 2019), *Analyses of 1st CEDEFOP DATA (July-Oct. 2019)*, https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/ESSnet\_WPB\_OJV\_CEDEFOP\_data\_analyses\_20190729CH.docx?version=1&modificationDate=1564382647936&api=v2

[3] CEDEFOP data analysis: Bulgaria, September 2019, https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/BNSI\_CEDEFOP%20data%20analysis.docx?version=1&modificationDate=1570446081935&api=v2

[4] Domenico Aprile, Massimiliano Amarone, Diego Chianella, Annalisa Lucarelli, Marina Sorrentino, (2019), *First results on Cedefop data analysis*, https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/First%20results%20on%20Cedefop%20data%20analysis\_Italy\_luglio\_2019\_draft%20version.docx?version=1&modificationDate=1564491771323&api=v2

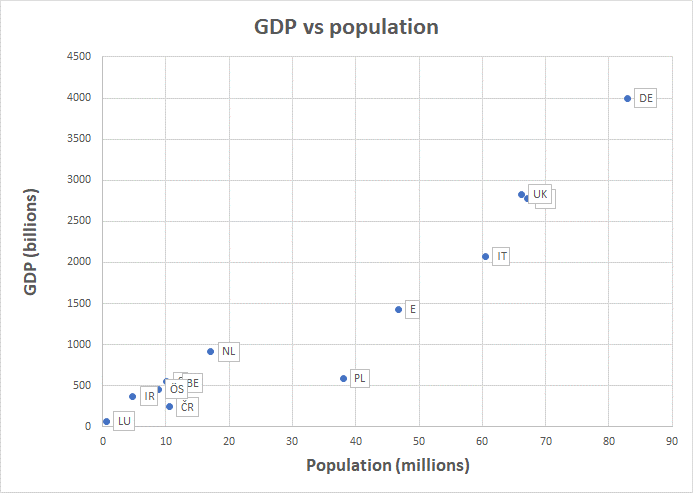
[5] Sue Westerman, (2019), *Main findings on navigating the public Cedefop dashboards*, https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/Main%20findings%20on%20navigating%20the%20public%20Cedefop%20dashboards\_Netherlands.docx?version=1&modificationDate=1562151646072&api=v2

[6] CEDEFOP data analysis Sweden, (2019), pg.4, https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/cedefop\_analysis\_SE.docx?version=1&modificationDate=1568710845794&api=v2

[7] Martina Rengers, Jakob De Lazzer (DESTATIS, Germany), *Open questions and requirements regarding CEDEFOP data set*, https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/DE\_open-questions-CEDEFOP\_answers-Matteo-Fontana\_within.pdf?version=1&modificationDate=1570537670254&api=v2

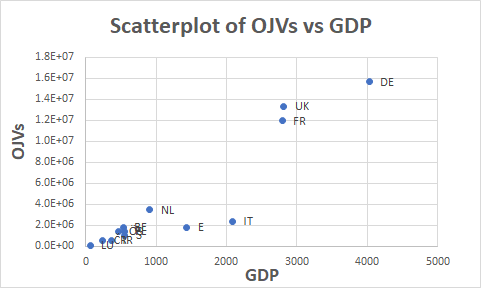
Annex 1

Figure GDP versus population by country



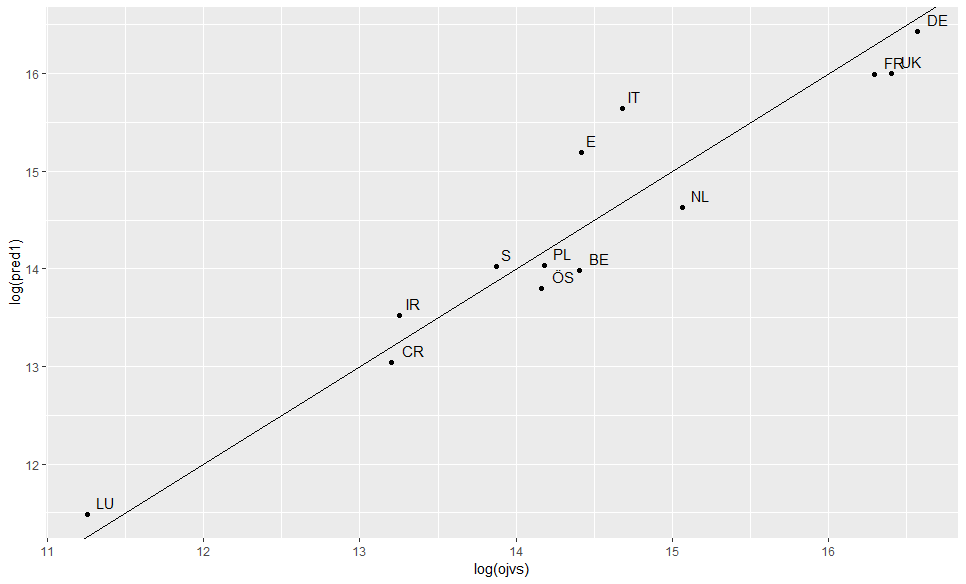
*Source*: Francis Saucy, Sophie Schmassmann, Alberto Columbano, Analyses of 1st CEDEFOP DATA (July-Oct. 2019), https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/ESSnet\_WPB\_OJV\_CEDEFOP\_data\_analyses\_20190729CH.docx?version=1&modificationDate=1564382647936&api=v2

Figure OJVs versus GDP by country



*Source*: Francis Saucy, Sophie Schmassmann, Alberto Columbano, Analyses of 1st CEDEFOP DATA (July-Oct. 2019), https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/ESSnet\_WPB\_OJV\_CEDEFOP\_data\_analyses\_20190729CH.docx?version=1&modificationDate=1564382647936&api=v2

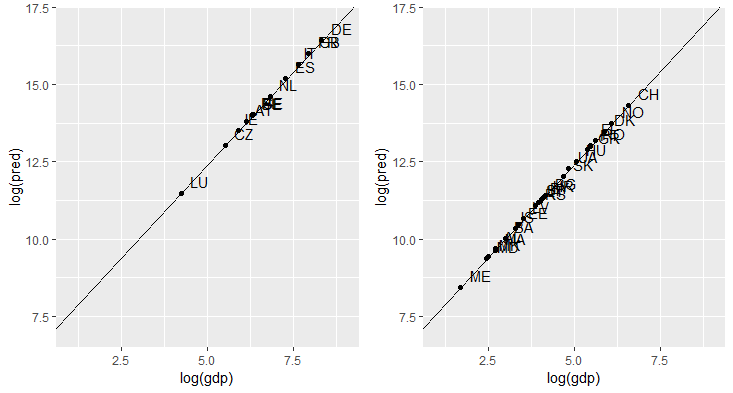
Figure Scatterplot of OJVs predicted (or fitted) by the model (vertical axis) against observed OJVs for the 13 countries of the CEDFOP data (log scales).



*Source*: Francis Saucy, Sophie Schmassmann, Alberto Columbano, Analyses of 1st CEDEFOP DATA (July-Oct. 2019), https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/ESSnet\_WPB\_OJV\_CEDEFOP\_data\_analyses\_20190729CH.docx?version=1&modificationDate=1564382647936&api=v2

Figure Predictions of the model

|  |  |
| --- | --- |
| 1. fit of the model for the 11 countries of the CEDEFOP data : predicted OJVs (vertical axis) against GDP (horizontal axis) | b) predictions of numbers of OJVS for the 13 remaining EU countries : predicted OJVs (vertical axis) against GDP (horizontal axis) |



*Source*: Francis Saucy, Sophie Schmassmann, Alberto Columbano, Analyses of 1st CEDEFOP DATA (July-Oct. 2019), https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/ESSnet\_WPB\_OJV\_CEDEFOP\_data\_analyses\_20190729CH.docx?version=1&modificationDate=1564382647936&api=v2

1. https://webgate.ec.europa.eu/fpfis/wikis/display/EstatBigData/Work+until+October+2019 [↑](#footnote-ref-1)
2. https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/364448484/ESSnet\_WPB\_OJV\_CEDEFOP\_data\_analyses\_20190729CH.docx?version=1&modificationDate=1564382647936&api=v2 [↑](#footnote-ref-2)
3. Cedefop (2019). The online job vacancy market in the EU: driving forces and emerging trends. Luxembourg: Publications Office. Cedefop research paper; No 72., http://data.europa.eu/doi/10.2801/16675 [↑](#footnote-ref-3)