ESSnet Big Data

Specific Grant Agreement No 2 (SGA-2)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/.....

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2016.010-2016.756**

Work Package 1

Web scraping / Job vacancies

Deliverable 2.2

Final Technical Report (SGA-2)

Version 2018-05-30

Prepared by:
Nigel Swier, Frantisek Hajnovic (ONS, UK)
Thomas Declite (StatBel, Belgium)
Martina Rengers, Chris-Gabriel Islam (DESTATIS, Germany
Ingegerd Jansson, Dan Wu, Suad Elezovic (SCB, Sweden)
Crt Grahonja (SURS, Slovenia)
Christina Pierrakou, Eleni Bisotti (ELSTAT, Greece)
Maxime Bergat, Alexis Eidelman (DARES, France)
Rui Alves, Maria-Jose Fernandes (INE, Portugal)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

Contents

Executive Summary

Nine National Statistics Institutes (NSIs) within the European Statistical System (ESS) have been investigating the feasibility of using online job vacancy (OJV) data in the production of official statistics. OJV data contain information that are not generally collected by job vacancy surveys (JVS), such as the occupations of advertised vacancies, associated skills and their location. OJV data also offers the possibility of more frequent and near real-time data on the labour market. However, there are some important limitations of OJV data:

- Not all job vacancies are advertised on-line and some types of jobs are more likely to be advertised than others.
- There is no definitive source of OJV data. It is generated and managed by various and mostly commercial actors.
- Data about on-line job ads usually contain a mix of structured and non-structured elements, but the specific structure and variables may vary between sources.
- Some job ads are out of scope of the official definition of a job vacancy (e.g. student internships, international jobs, non-existent "ghost" vacancies).
- The official definition of a job vacancy does not correspond directly to the concept of a live job ad. Critically, a vacancy will usually persist after the ad closes.
- The specific OJV data landscape can vary considerably between countries, for example, in terms of the number and type of portals and use of on-line platforms. There may also be differences in terms of the role of the National Employment Agency and what type of information is contained in job ads. There may also be legal differences and finally, processing will often require language specific solutions.
- In summary, OJV data is not representative of the overall labour market and there are various definitional issues that make it difficult to compare directly with official statistics.

There are different routes for accessing data. Broadly, these are direct web scraping (for either job portals or enterprise websites) or arranged access (e.g. with a job portal, the National Employment Agency, or commercial suppliers). The specific job vacancy landscape varies between countries. There may be good reasons for direct web scraping depending on the specific aims of a project. However, in general the recommended approach is to focus on accessing data that has already been collected. Apart from the technical and legal challenges of web scraping, there is also the problem that it will take a long time to generate a sufficient time series to properly evaluate the data. Acquiring data directly from data owners may help circumvent these problems.

This work package has established a close working relationship with the European Centre for the Development of Vocational Training (CEDEFOP). CEDEFOP are developing a web scraping system for all Member states and it has been agreed that this should also aim to serve the long-term needs of the ESS. Therefore, NSIs should also generally avoid investing heavily in developing web scraping approaches as OJV data is expected to become widely available to EU member states via CEDEFOP by the end of 2020.

Several partners have had some success in using machine learning to derive structured variables (e.g. NACE and ESCO codes) from structured variables (e.g. from advertised job titles) or the whole text of the job advertisement. However, these methods are imperfect and often only give satisfactory results for parts of the classification.

Various comparisons have been made for comparing and OJV and JVS data including: total vacancies, comparisons by industry sector (by NACE), and comparison of vacancy counts by enterprise. The results of these analyses have been mixed with some analyses comparing reasonably well, with others showing only a very loose relationship between the OJV data and the JVS.

Slovenia has come closest to producing an end to end pipeline for producing estimates of OJV ads that can be approximately compared with official estimates. This suggests that only about 40% of all Slovenian job vacancies are published on-line. Although the total on-line coverage may be better in other countries, there are issues that would make the Slovenian approach difficult to replicate for larger countries. One issue is the greater number of important portals. A greater problem is that various matching problems (e.g. deduplication, and matching of OJV and business register or survey data) become more difficult as data volumes increase.

One area that shows some promise is to use the time series properties of OJV data to improve existing statistics. The pilot has had modest success in predicting survey values using OJV data, so these data could be used for producing flash estimates. It may also be possible to use these time series properties to produce more frequent estimates, or even possibly reduce the frequency of the survey. Other possibilities for statistical outputs include statistics about occupations, required skills and labour demand in local areas.

We conclude that OJV data cannot be used to directly replace the existing job vacancy statistics that are required by EU regulation. Indeed, the quality issues are such that it is not clear if these data could be integrated in a way that would enable them to meet the standards expected of official statistics. On the other hand, OJV data can provide more granular insights that official estimates cannot. As will as a need for further methodological development, there is a need to address the presentational challenges for how OJV data should be interpreted and used together with official estimates.

1 Introduction

This report is an update of the work by the Big Data ESSnet WP1 – Web Scraping for Job Vacancy Statistics for the SGA-2 period. This ran from August 2017 to the end of May 2018. The previous phase of the project (SGA-1) ran from February 2016 to July 2017 and delivered the following:

i. Qualitative Assessment of Job Portals (delivered July 2016)[1]
ii. Interim SGA-1 Technical Report (delivered December 2016)[2]
iii. Final SGA-1 Technical Report (delivered July 2017)[3]

1.1 Participation

Six countries participated in the work package for SGA-1

- Germany (DESTATIS)
- Greece (ELSTAT)
- Italy (ISTAT)
- Slovenia (SURS)
- Sweden (SCB)
- United Kingdom (ONS)

Four countries joined the work package for SGA-2

- Belgium (Statbel)
- Denmark (DST)
- France (DARES)
- Portugal (INE)

Denmark had to withdraw soon after the start of SGA-2 due to a lack of staff. Italy's involvement throughout the pilot has been limited to collaborating on the use of the methods developed by a separate ESSNet work package (WP2 - Web scraping for enterprise statistics), led by ISTAT. This work package is of interest to WP1 because of the potential to use these techniques to collect data about jobs advertised on enterprise websites. There has also been some collaboration between WP1 and WP2 on legal issues on web scraping, which was published as a WP2 deliverable[4].

[1] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_final.docx, retrieved on 8th May 2018.
[2] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/64/WP1_Deliverable_1_2_final.pdf, retrieved on 8th May 2018.
[3] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_3_main_report_final_1.0.pdf, retrieved on 8th May 2018.
[4] Stateva et al. (2016), "Legal aspects related to web scraping enterprise websites" (Section 4 p.17)

1.2 Format of the report

The purpose of this report is two-fold. The first is to report to Eurostat on the work achieved during SGA-2. The second is to provide information to assist future projects within the ESS and the wider official statistics community that aim to use on-line job vacancy (OJV) data. There are two main factors that have guided the presentation of this report:

First, OJV data is highly heterogeneous and the data landscape varies considerably between countries. Some countries have much bigger and well developed on-line channels than others. Also, while official job vacancy statistics are subject to EU regulation there are some differences between countries in terms of the range of variables, the frequency of the survey, and the availability of microdata. This means that approaches to data validation and integration may need to be different. In addition, legal barriers constraining an NSI in one country may not be an issue in others. Finally, a lot of the research in this pilot involves text processing, which is language specific and constrains the reusability of prototype solutions. Consequently, each country in this pilot has needed to find their own path and the work of each country naturally forms a distinct case study.

The second main factor influencing the structure of this report is that the work undertaken for SGA-2 is to a large extent a continuation of the work done during SGA-1. This creates a certain challenge the additional work undertaken in SGA-2 needs to be presented in a way that is coherent yet avoids unnecessary repetition.

For these reasons, the main part of this report is written in the form a summary and guide. It attempts to summarise the findings the country based studies along with recommendations and advice for NSIs wishing to do similar work. The guide then provides links to relevant country case studies in a series of annexes which provide more detail. Links will also be provided to relevant material in the SGA-1 reports and other studies where necessary to minimise repetition.

As part of this introduction we introduce some basic facts about OJV data that need to be understood before starting any project:

1.3 Basic facts about OJV data

i) Not all job vacancies are advertised on-line:

Although there is a general trend towards more job vacancies being advertised on-line, many continue to be filled through traditional channels, such as newspapers, employment agencies (who may or may not advertise on-line), noticeboards, or personal contacts (Carnevale, 2014). The results of the Slovenian study within the WP1 pilot suggest that only about 40% of all job vacancies in Slovenia are advertised on-line. A similar proportion has been reported for France, although this figure may be higher in other

countries. In addition, some types of jobs are more likely to be advertised on-line than others. This means that OJV data is not only missing many jobs, but is also not representative of the overall job market.

ii) There is no definitive source of OJV data

Although the situation varies between countries, OJV data is generally characterised by multiple job portals, with different business models and complex interplay between them. Job boards only publish original ads uploaded by employers. Job search engines republish ads from other portals. Hybrid job portals are a combination of both. Some portals advertise many different types of jobs while others may specialise in specific sectors. In addition, new job portals may appear while existing portals may decline in importance. All these factors mean that it is very difficult to capture all jobs advertised on-line and to reliably measure labour market trends in the real world. It also means that the complete set of all jobs advertised is a country will contain many duplicates.

iii) Data about on-line job ads usually contain a mix of structured and non-structured elements

Most on-line job advertisements have some structured elements that are separate from the full text of the job ad, typically, job title, job location, employer name. Further information, such as skills and education, is available in the full text of the job description, but needs to be extracted and converted into structured elements using natural language processing (NLP) and classification algorithms. Variables such as occupation and industry code will usually also be derived using text analysis and machine learning. Data derived in this manner will inevitably contain some processing errors.

iv) A job ad is only a proxy measure for the existence of a job vacancy within a company

There are additional factors that make it challenging to relate on-line job ads to established statistical concepts and definitions, such as those used by the Job Vacancy Survey (JVS):

- Some ads may not represent an in-scope job vacancy. These include non-existent vacancies (referred to as “ghost vacancies”), international jobs, and non-paying student internships.
- Some ads may contain more than one vacancy. The number of vacancies may sometimes be specified in the job ad, but often it is not. Even when the number is explicit, this is usually contained within the unstructured part of the job ad and is difficult to extract.
- The Job Vacancy Survey (JVS) is a stock estimate of the number of vacancies for which businesses are actively seeking recruits from outside their organisation. Online job ads represent a flow of new

vacancies but usually do not contain explicit information on when the recruitment activity will be concluded. Therefore, direct comparisons between these sources requires some assumptions about how long a company will be actively recruiting once an advertisement is published.

- Different levels of information will be available for different vacancies. For example, vacancies advertised through agencies will usually not contain the name of the employing business. This will both affect the quality of industry coding of job ads as well as the quality of any linkage with survey units.

General recommendations

All these factors together make it very challenging to use OJV data to produce estimates about job vacancies with known error characteristics. Indeed, a key conclusion of this work package is that OJV data does not on its own provide a complete picture of labour demand. It may be possible to use these data to measure trends. However, even this requires caution as there is no easy way of separating underlying trends in the labour market from the trends in how jobs are being advertised. Therefore, any analysis of OJV data needs to be sense checked and where possible, validated against other data sources (including other OJV sources).

Since the specific characteristics of the on-line job market vary between countries, it is recommended that NSIs start with a landscaping exercise. This should involve the collection of information about the national job portal market such as the total number of portals, a more detailed assessment of the largest portals, the role of the National Employment Agency (NEA), and other relevant information about the national on-line job market.

NSIs within the EU should first contact the European Centre for the Development of Vocational Training[5] (CEDEFOP). They already undertaken a comprehensive landscaping exercise for all EU member states and will be willing to share any information. They should also be able to provide contact names of the national experts involved in these landscaping activities.

[5] http://www.cedefop.europa.eu/en/about-cedefop/contact-us, retrieved on 8th May 2018.

2 OJV Data Use Cases

2.1 Improving current job vacancy statistics

In considering application of OJV data, an obvious starting point will be to consider how OJV data could be used to improve current job vacancy statistics. Official job vacancy statistics are subject to EC regulation No. 453/2008 and are collected primarily for the purposes of calculating the job vacancy rate, a key measure of labour market tightness. This harmonised approach enables these statistics to be compared between EU member states. Further details about the regulation and its definitions are provided in Section 5.1.

The key differences between what is required for official job vacancy statistics for EU member states and what is available within OJV data are summarised in Table 1.

Table 1: Differences between the Official survey based estimates of job vacancy statistics

Dimension	Official estimates (JVS based)	OJV Data
Frequency	Quarterly	Real-time
Industry Sector	Yes	Yes
Enterprise size	Yes	Yes
Job title / Occupation / Skills[6]	No	Yes
Sub-national	No	Yes
National totals (estimates)	Yes	No

The EC regulation requires estimates to be published every quarter and there is typically a lag of several weeks or months between the reference day and the publication date. In contrast, OJV data could potentially be used to produce high frequency statistics in near real-time. In terms of variables, the JVS collects data about industry sector and enterprise size, both of which can be derived (at least to some extent) from OJV data.

A key advantage of OJV data is that it contains information about job vacancies that are not mandated by the regulation, yet are often requested by users. This includes data on occupations in demand, information about skills as well as location information which could be used to gain insight into local labour markets. However, as explained, while the JVS produces estimates based on representative

[6] A survey of employer skills for all EU member states was carried on in 2014. Some member states have their own skills surveys, but these are usually infrequent.

survey samples, online job ads are a selective subset of all jobs advertised by employers. Since OJV data does not cover the full labour market, it cannot be used to directly replace the JVS.

The final SGA-1 technical report proposed a theoretical outline of how OJV and JVS data could be integrated by linking data at the enterprise unit level and then constraining to totals from the JVS (Swier et al, 2017, p.15). The aim would be to produce data that included more variables and granularity, but was also consistent with official estimates. However, the work undertaken during SGA-2 casts doubt on the feasibility of this approach since the job vacancy count trends between OJV and JVS data at the unit level are too volatile to produce reliable scaling factors.

One possible business benefit could be to reduce the number of surveys to reduce collections costs. For example, the quarterly survey could become an annual survey with the OJV data being used to estimate the remaining 3 quarters. If feasible, this could reduce survey costs. Belgium have done some concrete work in this area, but the findings are not yet ready for this report and so the feasibility remains unclear. The Belgium country report instead focuses on the feasibility of deriving NACE group codes from OJV data to replace the data collected from the survey (See Annex A).

Although OJV data has some distinct advantages, realising these improvements for official statistics purposes is very challenging for the reasons described in Section 1. There is not yet a clear pathway for using these data to produce statistics that meet the quality standards required for official statistics. Slovenia have been able to produce some estimates of the number of on-line job vacancies, but these are not comparable with the statistics produced by the JVS.

For these reasons, it is important to be mindful of the limitations of OJV data and to have realistic objectives of what is possible. It is recommended that NSIs focus on specific problems that would help with the production of experimental type indicators, with integration into statistical production being viewed as only a possible longer-term goal. The rest of this section suggests some broad research areas and some more specific use cases.

2.2 Classifying data from text descriptions

On-line job ads contain a combination of structured and unstructured text. Variables such the job title, location and employer/agency name are usually stored as separate fields while the full text description of the job ad will contain a wide range of different types of information. Another issue is that location information will usually not conform to standard geographical units. Therefore, the data needs to be classified in some way before it can be analysed. Usually, these classifications will take the form of a recognised nomenclature such as ESCO (for occupation and skills) or NACE (economic activity) However, OJV data could also be classified using unsupervised or semi-supervised clustering models, which do not use a pre-defined structure (e.g. Djumalieva et al., 2018). In all cases, this processing involves some combination of text pre-processing and classification methods, typically involving machine learning.

This is an aspect that has been explored in some considerable depth during SGA-2. Further details are provided in section 4.3.

2.3 Measuring OJV coverage

The most important methodological challenge in using OJV data for official statistics is to be able to understand the differences between what is represented by these data and what is measured by the JVS. WP1 has explores three different ways of doing this:

- Micro-level comparisons with the JVS
- Aggregate comparisons with the JVS
- A survey of advertising channels

This is discussed further in Section 5.3.

2.4 Time series analysis

Another strand of analysis involves using the time series properties of OJV data and explore how it relates to official job vacancy estimates. In this ESSnet, Sweden and the UK have considered the potential of using the near real-time availability of OJV data to produce nowcasts (or flash estimates) of job vacancies. A time series approach might be particularly useful for predicting turning points in the economy. This is discussed further in Section 5.5.

2.5 Data driven analysis

Traditionally, the development of official statistics has been driven by clearly defined user needs. Data is collected that meets these needs as closely as possible, either directly through a survey, indirectly through administrative data sources, or possibly some combination of both. In future, big data sources such as OJV data, are expected to become somehow integrated into the activities of NSIs.

Although, it is expected that official statistics will continue to be led by clearly defined needs, the complex nature of big data means that some analysis could be more data driven. Big data in combination with data science offers the opportunity to identify “unknowns, unknowns”, that is, new insights of policy relevance into the quality of current statistics that may have not previously been considered.

For example, analysis of OJV data shows that new job advertisements in the UK are less likely to be published on-line on a Friday compared to other week days. This may of some relevance as the survey is run monthly and the reference day is always on the first Friday of the month. This kind of analysis could possibly help inform the survey operation and estimation methodology.

2.6 Other potential use cases

2.6.1 International Labour Market

There is potential to use OJV data to produce new statistics on aspects of the labour market, which are excluded from current statistics. One example includes international jobs, namely, job advertisements that are advertised in locations that are outside of country in which the job portal is based. This itself could be a useful measure of labour market tightness and could help identify types of vacancies that difficult to fill. It may also be that advertisements of this nature are very likely to be advertised on-line.

2.6.2 Identification of new job titles

Occupation classification coding frames require regular maintenance to ensure that they capture new job titles and OJV data is an excellent source of such information. This has already been used to in the UK at a small scale to update the national coding frame for the national occupation classification (UKSOC). There is also some consideration as to whether OJV data could be used to support maintenance of the classification itself by reflecting up to date information about job titles and skills in the labour market.

3 Data Access

3.1 Introduction

When embarking on an OJV data project, one of the first steps is to consider options for data access. The approaches to accessing on-line job advertisement data can be divided into two broad types: direct web scraping and arranged access. The most appropriate type of access depends on exactly how the data will be used. The following questions may help in guiding decisions:

- How much data is needed?
- Can the data be a one-off supply or will it be needed on an ongoing basis?
- If required on an ongoing basis, is it required in real (or near real time)?
- Is historical data needed?
- Is the complete job advertisement required or just aggregated data?
- Is the aim to combine the OJV data with other data (e.g. survey data)?

However, it is also to consider practical issues around data access such as:

- Does your organisation permit web scraping and do your IT systems support it?
- Does the project team have to develop, and if necessary maintain, a web scraping system?
- How much project resource should be dedicated to web scraping compared to data processing and analysis?
- How easy is it to access job vacancy data that is already available?

In general, obtaining sample data from one job portal through web scraping is relatively quick and easy. However, scaling up to include regular scraping of multiple websites and related pre-processing (e.g. de-duplication) can consume a lot of resource. Therefore, for it is recommended that more substantial projects consider the feasibility of accessing OJV data that already exists from a job portal or other data provider.

3.2 Direct web scraping

Direct web scraping involves using web scraping techniques to collect data from on-line sources without an explicit data access agreement from the website owner. Target websites may include either job portals or enterprise websites. The main advantage of direct web scraping is that samples of data can be captured and analysed quickly. This may be achieved either through simple “point and click” web scraping tools or programmable web scraping tools. Direct web scraping also offers a high degree of control over what and how data is collected and provides an opportunity to produce near real-time.

3.2.1 Point and click web scraping

Point and click web scraping tools have user interfaces that are designed to build web scraping robots without the user needing to write any code. A web scraping robot API is built through the point and click actions of the user highlighting and labelling web page elements of interest to train the robot to recognise the page layout. The point and click tools used during the WP1 pilot include Import.IO[7] and Content Grabber[8]. These proved very easy to use and effective for small scale data collection.

One problem encountered during the pilot was that Import.io changed their business model so that functionality that was previously free charge became a paid for service. Another potential problem with Import.IO is that scraped data is physically held on their servers. However, the main limitation is that these tools are designed to stand alone and cannot be easily integrated into a production pipeline. Therefore, point and click tools are best suited for small scale experimentation.

3.2.2 Programmatic web scraping

Programmatic approaches to web scraping involve developing and deploy web scraping robots, usually using packages such as Python Scrapy[9], or Apache Nutch. These packages require some programming skills, but offer much better control and pipeline integration. For example, a robot can be deployed to scrape on a regular basis and load new data into a database for further processing.

While many web scraping projects will be able to meet requirements with just one of these packages, some scraping some websites may require additional tools. For example, for some websites additional content is loaded through scrollbar interaction. Accessing the full content of websites with this functionality requires the use of web site automation tools (e.g. Selenium). This requires a more advanced knowledge of software development. In contrast, some job portals[10] provide a public API in which case job advertisement data can be accessed without any specific knowledge of web scraping.

3.2.3 Web scraping enterprise websites

For the most part, job portals will be the target websites for scraping OJV data. However, as discussed in the introduction, a specific objective of WP1 SGA-2 was to further explore the feasibility of collecting job vacancy information from enterprise websites. The thinking is that this approach could have advantages over web scraping job portals as it would both avoid duplication and produce data that could be assumed to be the most accurate on-line measure of job vacancies by an enterprise. The proposal was to explore the feasibility of using the overall framework for producing statistics from enterprise websites developed by WP2 and to apply it to this use case.

WP2 established that it is possible to crawl enterprise websites and identify those that advertise their own job vacancies. However, reliably creating structured data, including identifying individual

[7] https://www.import.io/, retrieved on 8th May 2018.
[8] https://contentgrabber.com/, retrieved on 8th May 2018.
[9] This was the most widely used web scraping package within this pilot.
[10] Examples in the UK include Adzuna and Universal Job Match.

advertisements, from the relevant web pages is much more challenging. Work by Slovenia shows that it is possible to identify individual job ads for some enterprise websites but the highly variable nature of website design makes it very difficult to do this on a more representative basis.

A more "brute force" approach for collecting data from enterprise websites was also explored by the UK. This involved manually develop a small "army of mini-robots" to capture counts of vacancies from selected enterprise websites. This involved developing a framework of reusable components that could be assembled and applied for different enterprise websites depending on the specific website design. A test involving 150 websites showed that this approach of using reusable components allowed robots to be assembled relatively quickly and could be used to scrape vacancy counts for just about any type of website. However, ultimately the overall effort required to create and maintain the robots does not make this a viable approach for large scale data collection.

3.2.4 Web Scraping Legal Issues

The relevant legal issues are well documented in the WP2 report on web scraping (Stateva et al, 2016[11]). The main issue is that many websites have restrictions on what content may be scraped from a web site, and many prohibit web scraping entirely. The specific legal risk is around "sui generis" database rights, which are a form of ownership right pertaining to data that apply when scraping "substantial parts" of a website[12]. However, NSIs should also consider the ethical and reputational risks of web scraping. These risks can be managed by following principles of web scraping "netiquette", such as respecting the robots.txt exclusion protocol.

Despite efforts to develop a consistent approach within the ESSnet, legal departments in different NSIs may have different opinions about these matters. Ultimately, projects need to ensure that web scraping is compliant with their own organisational policies.

3.3 Arranged access

This general approach to data access involves an explicitly agreed data access arrangement with either the website owner, or other organisations holding these data including job portals, employment agencies (government and private) and data aggregators.

There are several advantages of accessing data through an explicit agreement from owners of job vacancy data rather than through web scraping. The most important reason is that developing and maintaining web scraping robots can be resource intensive activity that can tying up scarce data science resources. Another important advantage is that it removes any uncertainty over legal issues in accessing and using the data. An explicit agreement may also offer a route for accessing historical data, which are rarely available through web scraping since job ads are normally removed once they expire. Finally, the

[11] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/67/WP2_Deliverable_2_1_2017_02_15.pdf, retrieved on 8th May 2018.

data owner may also be able to provide metadata or other insights into the methods used to collect and process the data.

3.3.1 National Employment Agencies (NEAs)

Five countries in this pilot (i.e. Sweden, France, Belgium, Germany, Slovenia) managed to gain access to OJV data from their NEA. These agencies are typically the largest single source of OJV data. In some countries certain types of jobs (e.g. in the public sector) are required by law to be advertised by the NEA. Some, NEAs use common enterprise identifiers which greatly simplify the process of linking to business registers and survey reporting units. Another important advantage is that access arrangements are less likely be complicated by commercial considerations. Therefore, a recommended first step for NSIs looking to acquire data is to explore the possibility of gaining access to the job vacancy data from the NEA.

3.3.2 Private Job Portal Owners

Several countries (i.e. Sweden, UK, Slovenia) managed to secure some data supply arrangements with at least one private job portal owner. When approaching these organisations, it is important to consider what they might want out of a data supply arrangement and whether this is consistent with the policies of your organisation. Motivations could include payment for data, payment for data services or some in-kind benefit. An example of the latter could be an offer to link to their data to JVS data to provide an aggregate report to show some new insight into the coverage of their data. The main motivation for private job portals agreeing to partner with the NSI seems to part of corporate social responsibility but they are also likely to be interested in the publicity benefits of their data being used by the NSI.

In the UK, procurement rules required an open tender process to ensure that all potential providers had an equal opportunity to reap these benefits. This was required even though the tender made it clear that ONS policy is not to pay for data. Multiple teams were needed to be involved to support this process including finance and procurement, and the press office. The UK has recently established a commercial data acquisition team which has really helped with navigating the correct process. The process in Sweden and Slovenia was less formal and so therefore it is important to establish what steps are required by your organisation.

Slovenia also undertook some specific engagement of some private employment agencies. Many job advertisements are made by employment agencies where the actual employer is not identified. This is major problem for any analysis involving the linking of OJV to survey data and indeed is also a problem for correctly classifying job vacancies by economic activity. This engagement process resulted in some modest success in obtaining some additional information. However, the sheer numbers of private employment agencies in some countries[13] means that this is not always a practical option.

[13] For example, the UK has over 14,000 employment agencies.

3.3.3 Commercial suppliers of OJV data

There are several international companies that scrape job advertisement data from the web, process and enrich it to provide commercial data and analytical services (e.g. Textkernel, Burning Glass, CEB Talent Neuron). The advantages of these sources are that the time-consuming processing to prepare this data for analysis (e.g. deduplication, cleaning and enrichment) will have already been done. The disadvantages are that these methods will usually not be transparent and some form of payment will normally be expected. Another limitation is that these products are often only available in larger job markets (e.g. Germany and UK). The UK pilot managed to negotiate access to UK Burning Glass data without direct payment, although there are specific circumstances that may make this difficult to replicate elsewhere. For this reason, this is not generally considered a viable approach.

3.3.4 CEDFOP

In 2015, the European Centre for Vocational Training (CEDEFOP) funded a web scraping pilot of selected job portals in five EU member states (Germany, Italy, UK, Ireland, Czech). The pilot included development of an experimental system to remove clean data, remove duplicates and classify for analysis. A detailed assessment of the 2015 pilot data is made by Germany (Annex C, Section 3).

In early 2017, CEDEFOP launched the second phase which plans to develop a system for collecting on-line job advertisement data for all EU member states. In Spring 2018, a system of ongoing web scraping and processing commenced initially for eight countries with the first data becoming available by the end of 2018. Data for all member states will become available by the end of 2020.

It is becoming clear that different public organisations, including different parts of the European Commission have an interest in OJV data. It is also clear that it is not efficient to have duplicate systems to collect and process these data for different purposes. An agreement is now in place between CEDEFOP and Eurostat to facilitate collaboration and to ensure that the statistical requirements of the ESS are considered in the development of the CEDEFOP system. Expertise in the areas of quality and statistical measurement means that the ESS could play an important role in the appropriate use of these data for policy purposes. In March 2018, a joint validation workshop between representatives from the ESSNet, Eurostat, CEDEFOP and their contracting partners from the University of Milan to help achieve alignment with the requirements of the ESS.

It is important that any NSIs wanting to work with OJV data take account of these developments as part of their long-term planning. Specifically, it is recommended that NSIs avoid committing too many resources into the development of web scraping, data cleaning, deduplication and data enrichment methodologies since in the long term CEDEFOP data is likely to become the main source of OJV data to support the statistical activities of the ESS.

3.4 Summary

In general, options for acquiring data directly from suppliers should be explored before direct web scraping. Exploring options for acquiring data from the NEA is a good place to start. Depending on the specific aims of the project, there may be good reasons for a direct web scraping. In general, NSIs should generally avoid investing too heavily in web scraping and data processing since processed OJV data is expected to become available to all EU member states via CEDEFOP by the end of 2020.

All countries in the WP1 pilot have managed to gain some form of access to job portal data, either through direct web scraping, agreed access or both. The various avenues to data access by each country are summarised in Table 2.

Table 2: Investigation of On-line Job Vacancy Sources by Country

Country	Direct Web scraping		Agreed Access			
	Enterprise websites	Job Portals		National Employment Agency	Private Employment Agencies	Other data aggregators
Germany		Yes		Yes		Yes (CEDEFOP)
Greece		Yes				
Slovenia	Yes	Yes	Yes	Yes	Yes	
Sweden			Yes	Yes		
United Kingdom	Yes	Yes	Yes			Yes (CEDEFOP, Burning Glass)
France		Yes		Yes		
Belgium	Yes	Yes		Yes		
Portugal		Yes				

4 Data Handling and IT

4.1 Data storage and data handling software

For most countries in this pilot, the volumes of data that have handled so far are not so very large and can be processed on a single machine. For this reason, big data IT solutions have not been explored in any depth. The UK pilot is using NoSQL data storage (i.e. MongoDB), mainly because it had been established for other web scraping projects. The main advantage of this approach is that this will provide scalability if greater volumes of data are required in future. The UK web scrapers are hosted on a Google compute platform as this cannot currently be done from the main network.

Most countries used Python and related packages for data handling and machine learning although Belgium used R based packages.

4.2 Data cleaning and de-duplication

The raw data obtained from a job advertisement typically requires a lot of cleaning and pre-processing prior to analysis. For example, job titles often contain extraneous information, such as job location, key skills, and salary. This is because employers try to attract the attention of potential job seekers by stacking the job title field with other key information. In addition, OJV data will often include data that may be considered out of scope, for example, jobs based in another country or ads for unpaid student internships. These ads should to be identified and removed where possible.

Duplication of job ads is a key quality issue when combining data from multiple job portals. It can also be a problem within portals, particularly for job search engines where job ads from other portals. Job search engines may take steps to remove duplicates but the effectiveness of these procedures seems to be variable. Some duplicates within job search engines can be identified easily because the URLs linking back to the original ad will be identical. However, duplicates where a job ad has been posted on more than one job board are more difficult to identify.

Duplication methods were explored as part of a "virtual sprint" held during SGA-1[14]. These focused on matching common fields, comparing text content and then calculating a similarity metric to establish the likelihood that two job advertisements are the same. The first step is to prepare and standardize the data fields that are common and that can be compared to all data sources (i.e. job title, location, company name, date posted, job description). This involves text normalisation procedures such as, removal of white spaces, case standardisation and removal of stop words or other extraneous text, typically using regular expressions (regex).

[14] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP1_2016_07_28-29_Virtual_Sprint_Notes, retrieved on 8th May 2018. These are also documented within the WP1 - interim technical report

The next step involves calculating a similarity metric to identify any likely duplicate job ads. One approach involved using Python Dedupe, which is designed to identify duplicate records using supervised learning methods. This uses an initial match using logistic regression and then identifies marginal cases for clerical resolution. The decisions of this clerical process are then reincorporated into the machine learning algorithm, to be applied for automated removal of duplicates. Other duplication methods simply focused on the unsupervised probabilistic matching on the similarity of text strings. Methods explored included, Levenshtein distance and longest common substring distance, with Jaccard Similarity performing best.

The initial focus was on the structured data fields rather than the unstructured content of the full job description. This was mainly because this information is often difficult to scrape from websites in full and often only a "snippet" of a certain number of characters from the full job ads is readily available. This information would be needed to achieve a good quality de-duplicated data set, especially where there are many records.

Slovenia use three main sources of data, i) jobs that are advertised through the National Employment Agency (ESS), ii) Deduplicated job vacancies from the two largest job portals and iii) data scraped directly from enterprise websites. When combining these data there is a further duplication step, which involves matching each source to the business register and then using the source based on a priority principle. ESS data is used first. If an enterprise is found in the job portal data but not in the ESS data, then the job portal data is used. Last of all, vacancies are included that are found on enterprise websites where those enterprises cannot be found in the other two sources.

4.3 Text Analysis and Classification

Classification of OJV data is an important and complex topic with many different dimensions. OJV data contains information that can be used to derive occupation (e.g. ESCO), economic activity of the employer (e.g. NACE) and classifying to standard geographic units. It can also involve different elements of the job ad for related but distinct purposes. For example, it can be applied to matching occupation to job title but also to predicting occupation based on the text of the full job description. Finally, while official nomenclatures need to be used for official statistics purposes, it is possible to apply unsupervised clustering methods to gain new insights into the data.

Broadly, the methods for assigning classifications to OJV data comprise of two main steps:

i) Text pre-processing
ii) Classification methods[15]

In the context of the ESS there is additional complexity because these approaches require language specific lists, look-ups and training data. This is a particularly relevant issue in countries where there is

[15] While the first step clearly relates to data handling, the second could be considered more about methodology, which is covered separately in Section 5. However they are covered together here since they are so tightly integrated.

more than one official language (e.g. Belgium) or where the official language uses a combination of Latin and non-Latin scripts (e.g. Greece). In many countries it is common for jobs to be advertised in English, even where it is not an official language.

4.3.1 Text Pre-processing

Machine learning of text data requires some text pre-processing to ensure algorithms work effectively. This typically involves the combination of different sub-processes:

- Text standardization: This includes conversion to lowercase, removal of punctuation, repeated white spaces and non-alpha-numeric characters. This is normally implemented using regular expressions (regex), which is supported by all common programming languages.
- Stop word removal: Stop words are common words (e.g. “and”, “the”, “a”) that do not convey meaningful information in the context.
- White/black lists: These are lists of words that are applied as filters to be allowed or disallowed into the processed dataset.
- Stemming: This involves removing the end or the beginning of the word using a list of common prefixes and suffixes that are relevant to the language of the text. For example, the word “making” would be transformed to the stem “mak”
- Lemmatization: This involves taking into consideration the morphological analysis of words and requires more detailed dictionaries. For example, the word “making” would be transformed to the lemma “make”. TreeTagger is an example of a lemmatization tool that supports several European languages. Some lemmatization tools (e.g. Morphalou) can take account of different meaning depending on context. For example:

  Il conduit un **bus** (*He drives a bus*) -> Conduire **bus**
  Je **bus** une vodka (*I drank a vodka*) -> **Boire** vodka

4.3.2 Text classification methods[16]

Two broad classification methods were explored. A phrase based classification approach (PBC) was explored by Greece, while other countries explored machine learning approaches. Phrase based classification involves the creation of rules which trigger an action when a certain phrase is encountered. The main advantage is a high degree of precision and results which are easily explainable. A disadvantage is scalability for more complex classification problems.

Machine learning classification approaches exploit the relationships between the labelled features of a dataset and other features to build models that predict the values of unlabelled features in an unseen dataset. Machine learning is easier to apply on large scale data but these methods often lack transparency. The following is a brief description of the various WP1 SGA-2 studies:

- Belgium: Machine learning approach used to classify NACE group codes based on the full job description.
- France: Machine learning approach used to classify occupation codes based on the full job description, also string matching.
- Greece: Rules based approach using phrase based classification codes used to classify occupation codes for IT job advertisements in both Greek and English.
- Germany: Machine learning approach used to classify NACE group codes based on the full job description, also filtering ads from private employment agencies.
- Portugal: Machine learning approach used to classify occupation codes based on the job title only.

Further details are available in the relevant country Annexes.

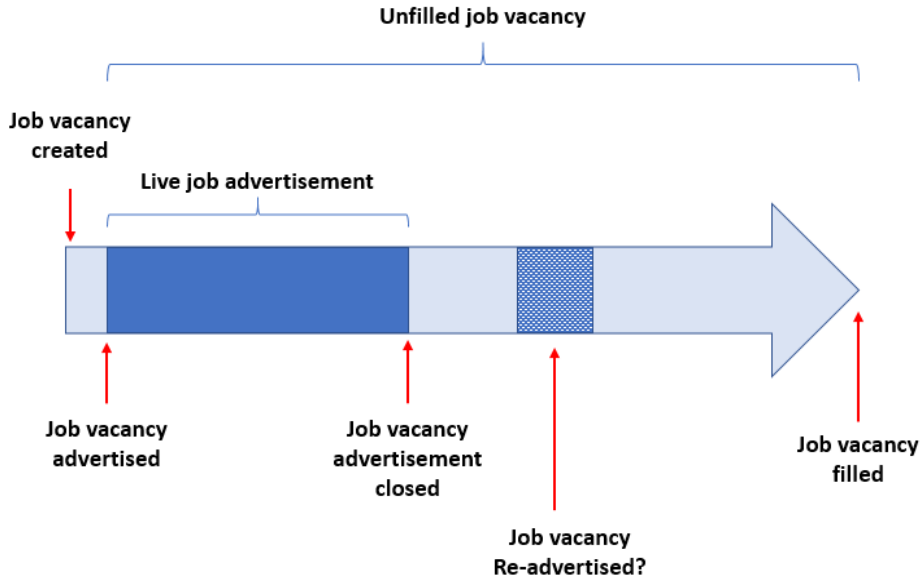
A general finding from these studies is that these machine learning classification methods usually work better for certain parts of a classification than others. For example, in the Belgian and German studies, the NACE categories I (Accommodation and Food Services) and O (Public Administration and Defence) were much more likely to be classified correctly than L (Real Estate activities) and R (Other Services). For some other NACE categories there was considerable variation in the results between the two countries and so there may be language specific or other technical factors that need further investigation.

An interesting finding from the German study was that the data from the German Federal Employment Agency (FEA) contains many more structured variables (e.g. occupation code, number of vacancies, type of degree) than the ads from private job portals. This makes the FEA data a feature rich source of training data which could be applied to the full text of job descriptions from private portals. This further strengthens the case for prioritising access to NEA data.

4.4 Flow to Stock transformation

When combining or comparing OJV data to the JVS it is important to consider the definitional differences between these two types of data. The JVS is a stock measure of the number of vacancies that employers are taking active steps to fill at a specific time point. OJV data may be collected as either a flow of new job ads, if collected continuously, or a stock of live job ads open on a specific date. Some job ads will have a closing date, and while some portals will remove jobs when they expire, this does not always occur. However, even with a known closing date the problem remains that an employer will typically still be taking active steps to fill a vacancy even after the job advertisement has closed. This may even include re-advertising the vacancy. These definitional problems are illustrated in Figure 1.

Figure 1: Job Vacancy Lifecycle:



OJV and JVS data can be made more directly comparable by estimating the time between when a job is advertised and when the vacancy is filled and then apply an adjustment factor to the OJV data[17]. Slovenia has approached this problem by using data on public sector jobs advertised through their national employment agency (ESS). These data can be linked directly to compulsory health insurance applications and so the difference between the start of the job ad and when a person starts the job can be measure directly. This means a distribution can be created and then applied to the entire flow of new job ads to estimate the stock of unfilled vacancies at any time point. Various methods were tested to establish how the time periods between advertising and filling a vacancy are distributed. The best distribution was geometric with a mean of 48 days, a median of 33 days and a standard deviation of 48 days. The distribution has a long tail because a small number of vacancies take a long time to fill. This is discussed extensively in Annex F, Section 3.3.

For the UK Burning Glass data, a transformation step was needed because the data only had a start date. As there is no equivalent data such as used in Slovenia, an approach developed that tested a set of stock time series based on different assumptions on the average length of time to fill a vacancy. The time series with the best trend alignment between the transformed data and the JVS, which used an assumption of 36 days. This assumption is applied to all subsequent analysis of Burning Glass data to derive an estimated daily measure of job vacancy stocks. This is discussed further in Annex H, Section 3.1.

[17] De Pedreza et al. (2017) take a different approach when comparing JVS and OJV data in the Netherlands. The JVS in the Netherlands asks some additional questions about new vacancies created in the previous quarter. This means it is possible to directly compare new vacancies measured in the JVS with new vacancies in the OJV data source.

Although these two approaches are not directly comparable, the average time periods are roughly similar, and intuitively seem "about right". It should be clear that while using these assumptions to transform data in this way will improve comparability between OJV and JVS data, these are coarse assumptions (especially in the case of the UK) and that will inevitably introduce some errors. It might be possible to refine this period value for different industry sectors, but this is limited by the problems of reliably disaggregating OJV data by industry.

4.5 Conclusion

The effort required to handle large amounts of raw data from job portals into a form ready for analysis, should not be underestimated. Raw OJV data invariably requires a lot of cleaning, deduplication and data reduction. Key variables of interest such as occupation and economic activity need to be classified from text. Finally, there are important definitional differences between OJV and JVS data and data may need to be transformed to make the data comparable.

For these reasons, NSIs need to consider what is the best use of their time and resources. NSIs should consider options that will minimise the amount of data handling required. NSIs should also keep in mind that CEDEFOP is expected to become a key source of OJV data for all EU member states over the long term and many of these data handling processes will have already been applied to the data. However, as CEDEFOP is focused on meeting EU requirements, only harmonised EU classifications are likely to be available (i.e. NACE and ESCO). Therefore, countries with specific national requirements may wish to focus on developing methods that are specific to their national classification systems.

5 Methodology

As explained in Sections 1, 2, and 4.4 there are fundamental issues around the quality of OJV data, what these represent, and how they compare with official estimates produced from the JVS. This provides some major methodological challenges in using these data to produce official statistics.

The section starts with a discussion of the definitions and target concepts used for official job vacancy statistics and this differs from corresponding OJV data. This is followed by a discussion on data quality frameworks and some steps that have been taken to compare OJV data against these frameworks. This is followed by a discussion on approaches for coverage assessment followed by data matching (which is one means of assessing coverage). The final section is a discussion of time series methods, which focuses more on how OJV and JVS data compare over time.

5.1 Definitions

There are differences between the target concept of official surveys and what can be practically measured from on-line job advertisements. Job vacancy statistics within the ESS are currently subject to EC regulation No. 453/2008. This defines a job vacancy as:

"… a paid post that is newly created, unoccupied, or about to become vacant:

(a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and

(b) which the employer intends to fill either immediately or within a specific period of time." [18]

EC regulation 453/2008 has several mandatory elements:

- Quarterly data that has been seasonally adjusted
- Data broken down economic activity (using NACE[19])
- Data is relevant and complete, accurate and comprehensive, timely, coherent, comparable, and readily accessible to users.

There are other elements that are optional, or subject to feasibility, including:

- Job vacancies in the agriculture, forestry and fishing sectors
- Job vacancies in public administration, defence and education
- Data on businesses with less than 10 employees
- Distinguishing between fixed term and permanent jobs.

[18] Regulation (EC) No 453/2008 of the European Parliament and of the Council http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:145:0234:0237:EN:PDF, retrieved on 8th May 2018.

[19] http://ec.europa.eu/competition/mergers/cases/index/nace_all.html, retrieved on 8th May 2018.

Member states are granted considerable flexibility regarding the implementation of regulation 453/2008 in the national statistical systems. Some countries use stand-alone surveys, while others combine the job vacancy survey with other business surveys. Some collect the minimum information required by the regulation while others collect data for their own national purposes. Although the regulation states that the data shall be collected using business surveys, the use of administrative data is equally permitted under the condition that the data are "appropriate in terms of quality" (according to the quality criteria of the European Statistical System).

The official definition of a job vacancy does not correspond exactly to the concept of a live job ad. Critically, a vacancy will normally persist after for a period after the closing date. In theory, this means that the stock of vacancies as measured by the JVS should generally be higher than corresponding OJV data.

5.2 Quality Assessment Frameworks

It should already be clear that there are many quality issues in consider when working with OJV data. During SGA-1, a "virtual sprint" was held by WP1 to consider two possible quality frameworks:

1. The Quality Assessment Framework[20] used by Statistics New Zealand as reporting tool for administrative data quality. The aim was to test the suitability of this framework for web-scraping for on-line job advertisements. This was in part in response from some initial proposals put forward by WP8 for approaching big data quality.
2. The UNECE framework for the Quality of Big Data[21] developed by the UNECE Big Data Quality Task team.

A summary of this work is available in the final SGA-1 technical report[22] (Section 4.2) and a more detailed report is available on the ESSnet wiki[23]. The general conclusion was that the UNECE framework seems to be more intuitive and so it could be a more useful starting point for documenting issues on quality on on-line advertisement data. However, the Statistics New Zealand Quality Framework is designed to support a total survey error approach which could further deepen the accuracy and selectivity dimension of the UNECE framework. These elements would become more important when considering how on-line job advertisements could be moved into statistical production.

[20] http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality/sources-of-error.aspx, retrieved on ???.

[21] http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2, retrieved on 8th May 2018.

[22] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/20/Deliverable_1_3_main_report_final_1.0.pdf, retrieved on 8th May 2018.

[23] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Virtual_Sprint_1_February_2017

Further to this, a more comprehensive review against the UNECE framework was done in preparation for a March 2018 workshop to validate the CEDEFOP web scraping system and its suitability for official statistics purposes. The aim was to explore the feasibility of incorporating this, or a similar, quality framework into the CEDEFOP production system. One of key outcomes of this process was identifying the importance of having metadata and where possible, meaningful quality measures, of all data processes. This would help enable official statisticians to make appropriate judgements about the data coming from the CEDEFOP system. This is also available as a separate report[24].

5.3 Measuring Coverage

As previously discussed, one of the fundamental quality issues in using on-line job advertisement data is that not all job vacancies are advertised on-line. Understanding these issues of coverage has been a key focus for this pilot and three distinct approaches have been identified for trying to better understand and measure these differences:

5.3.1 Micro-level comparisons

This involves linking OJV data to either the reporting units of the JVS or the Business Register. The UK have linked JVS data to vacancy counts by company for a range of OJV sources. This has revealed that the pattern of job counts between the JVS and other sources by reporting unit/company is typically very messy and difficult to understand. Slovenia have also undertaken this kind of microdata analysis as part of the production of experimental estimates.

For more details see: Section 5.5; Annex H, Section 3.1 (UK); Annex F, Section 4. (Slovenia)

5.3.2 Aggregate comparisons

This approach involves comparing JVS aggregates by industry sector (i.e. NACE) with the equivalent taxonomies from job portals. This approach has been used by Germany, who do not have access to JVS micro data. While this is quite a straightforward approach, private portals usually have their own taxonomies, which are often only approximately comparable with NACE.

For more details see: Annex C, Sections 2 and 3 (Germany) and Annex G, Section 4. (Sweden)

5.3.3 Measuring use of advertising channels via the JVS

A third approach has involved surveying enterprises and asking specific questions about advertising channels. Surveys of this type have been carried out by both Germany and Slovenia. The German study found that large companies were more likely to advertise on-line, with small companies more likely to use other channels.

[24] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/7d/WP1_Quality_Framework_v1.1.pdf, retrieved on 8th May 2018.

For more details see: SGA-1 Final Technical Report[25], p.34 (Germany) and p.58 (Slovenia)

5.4 Matching and linking

Several country pilots have explored the matching of on-line job ads with their own JVS micro data or Business Register data as a means of understand coverage issues. The results have been somewhat mixed. Linking between at the enterprise level between the Swedish Employment Agency (PB) and the Business Register is straightforward since they both use a common enterprise identifier. However, there is no such identifier for local units and so linking data at this level requires probabilistic matching using variables including enterprise name and location. This approach in produces lower quality results.

In the UK, enterprise level matching is done on solely company name only, which has proven very problematic. Common problems include use of abbreviated names, trading names rather than the legal enterprise name, and misalignment between the company names and the JVS reporting unit. Germany encountered the same problem for their Business Register. Also, there were difficulties in obtaining JVS microdata, since this is administered by another agency. In the Slovenian study, there are several enterprise matching steps that are made as part of the final deduplication process. This involves matching jobs from the national employment agency, deduplicated jobs from job portals, and additional jobs found on enterprise websites.

A major issue in trying match JVS reporting units to company names in job portals is that many jobs are advertised through private employment agencies and the employer is not usually identified in the advertisement. In some cases, there may be clues in the job ad about the type of business, or its location. Also, if matching counts between the JVS and direct employer counts from the on-line sources, then any shortfalls in the on-line data may provide further clues as to which employers and what type of jobs are being advertised through employment agencies. Work undertaken by Slovenia reveals that slightly less than 10 per cent of job offers are made through employment agencies. However, this proportion may well be higher in other countries[26].

5.5 Time series analysis

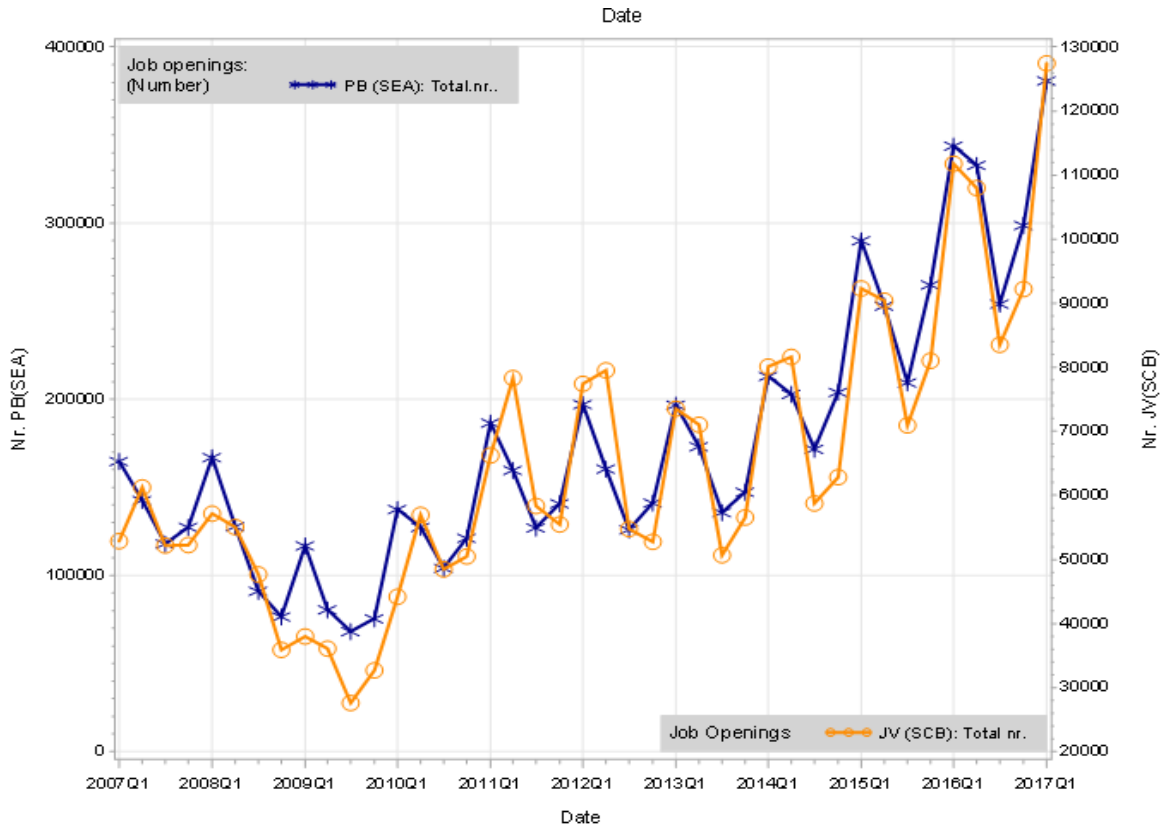
Despite the many issues with OJV data one promising approach is to exploit the time series relationships with the corresponding JVS estimates. A recent study using ten years of OJV data from a data aggregator in the Netherlands identified a clear relationship between macro-level trends in OJV data in the Netherlands and the official job vacancy estimates (de Pedraza et al, 2017). Although these relationships are not strong enough to suggest that OJV data to replace the survey, this does suggest that OJV data could be incorporated into modelling approaches.

[25] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/20/Deliverable_1_3_main_report_final_1.0.pdf, retrieved on 8th May 2018.

[26] In 2015, the UK had 14,280 business classified as activities of employment placement agencies.

Sweden has a continuous source of on-line vacancy data from their NES, Platsbaken (PB), going back to 2007. While, the overall levels of job openings are much higher than the Swedish JVS, a standardized view of the data shows that the data has similar time series properties, including a very distinct seasonal pattern (Figure 2).

Figure 2: Job openings from Swedish National Employment Agency (PB) and JVS (2007-2017)



Disaggregation of these series into public and private sector jobs shows greater trend correspondence for the public sector (especially in recent years) with a correspondingly weaker pattern for the private sector. Further details are available in the Annex G, Section 4.

The UK have explored the idea of using the time series properties of multiple OJV sources using a machine learning approach to nowcast job vacancy counts at the level of the individual enterprise or reporting unit. The advantage of this highly disaggregated approach is that it removes the problem sampling variance and uses the actual reported survey values as the training data. This uses a Long-Short Term Memory (LSTM) neural network algorithm using the JVS as training data[27]. The basic idea is that

[27] The UK JVS runs monthly which obviously provides more data points than the quarterly time series available in most other Member states.

different OJV sources may work better for different companies and that the model will choose the specific time series the works best. An example for a specific company is shown in Figure 3.

Figure 3: Example of a LSTM nowcasts for a specific company

The performance gave a modest improvement over a base-line persistence model (i.e. the predicted value based on the previous value) with the model being able to predict the correct direction of the trend in about 70% of cases. The limitations with this kind of approach is the short available time series for most of the OJV data an also it can only be applied to the approximately 25% of survey units that are always in the sample. However, this kind of approach may be a plausible solution to the problem of a highly dynamic data environment where different sources may be growing or declining in importance. Further details are available in Annex H, Section 4.

5.6 Summary

There are several major methodological challenges in using OJV data for statistical purposes. Some of these relate to differences in what an online job ad represents and how it corresponds to the target concept of a job vacancy. While these are difficult to solve, there are different approaches for evaluating quality and to better understand the extent of these differences. Other methods may be required to support these approaches, such as record linkage to compare counts for statistical units. Time series analysis is another useful perspective for better understanding OJV data. It may be possible to exploit the time series properties of OJV data by using nowcasting methods to produce more timely statistics.

6 Statistical Outputs

Most of the results of this work package are intermediate analyses and cannot be classified as statistical outputs, or even experimental statistics. However, there are a few results of the latter as well as some other examples of types of analysis that may be possible with this kind of data.

6.1 Estimates of on-line job vacancies

Slovenia have come the closest to developing an approach for producing experimental statistics based on OJV data (Table 3). These figures are the fully deduplicated detected job ads from the National Employment Agency, the two largest job portals and enterprise websites. The "detected job ads for the quarter" include data from before the reference period where a distribution has been applied to adjust for unfilled jobs after the jobs have closed (see Section 4.4). A comparison with the official job vacancy estimates show that only about 40% of all Slovenian job vacancies can be found on-line.

Despite the large difference with the official estimate, the greater frequency of OJV data offers the possibility of variations to these statistics. These include job ads that are available (i.e. advertised) both during the reference month and reference day, as well as new ads for both the reference month and the reference day. These are also shown in Table 3.

Table 3: Experimental on-line job vacancy statistics for Slovenia

Estimates	Type	28 August 2017 (Q3)	30 November 2017 (Q4)
Detected job ads for quarter	Stock	6849	6327
Official JVS estimate	*Stock*	*17221*	*15243*
Available in reference month	Stock	3542	4493
Available on reference day	Stock	1368	1285
Newly available on reference month	Flow	1984	2115
Newly available on reference day	Flow	123	76

6.2 Indicators or nowcasts of labour market activity based on OJV data

Given, the complexities of duplication and of producing figures that are directly comparable with official JVS estimates, an approach could be to produce a simple indicator or index. Where there are number of different OJV sources available one very simple approach could be to produce an average. Figure 4a shows eleven UK job portals averaged by month, Figure 4b shows a daily average of the eleven job

portals. In both cases these series are scaled to the JVS. These show that the OJV sources can detect a similar seasonal pattern seen in the non-seasonally adjusted survey.

Figure 4a: Time series of the total JV counts, averaged per month (scaled to the JVS scale).

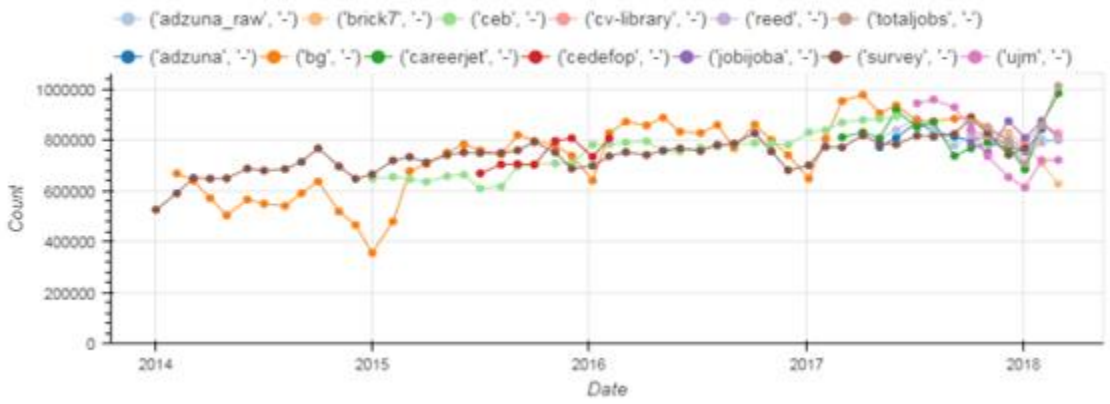
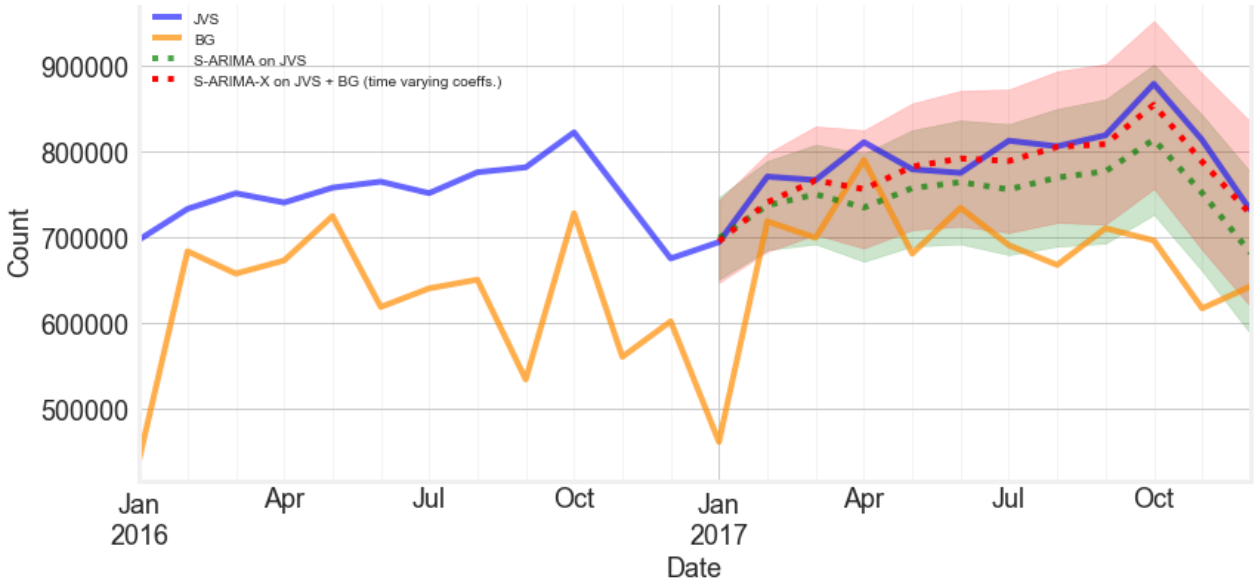


Figure 4b: Time series of the JVS and daily average of the online sources (scaled to the JVS)

The UK tried compiling the LSTM-based nowcasted estimates by company to produce an aggregate nowcast indicator (which was also scaled to the JVS). Given the model's weak predictive power and the small sub-sample (100 companies) it is perhaps not surprising that this aggregate nowcast indicator did not produce very good results. However, very late in the ESSNet, a different nowcasting approach was explored based on the S-ARIMA-X time series model. The green dotted line represents the nowcasts based only on JVS, while the red dotted line shows the (more precise) nowcasts that include the aggregate Burning Glass data as an exogenous variable. The shaded areas of respective colours indicate respective 95% confidence intervals. The inclusion of time varying coefficients made a noticeable improvement to the model and so this seems a promising area worthy of further exploration.

This is discussed further in Annex H (UK), Section 5.4

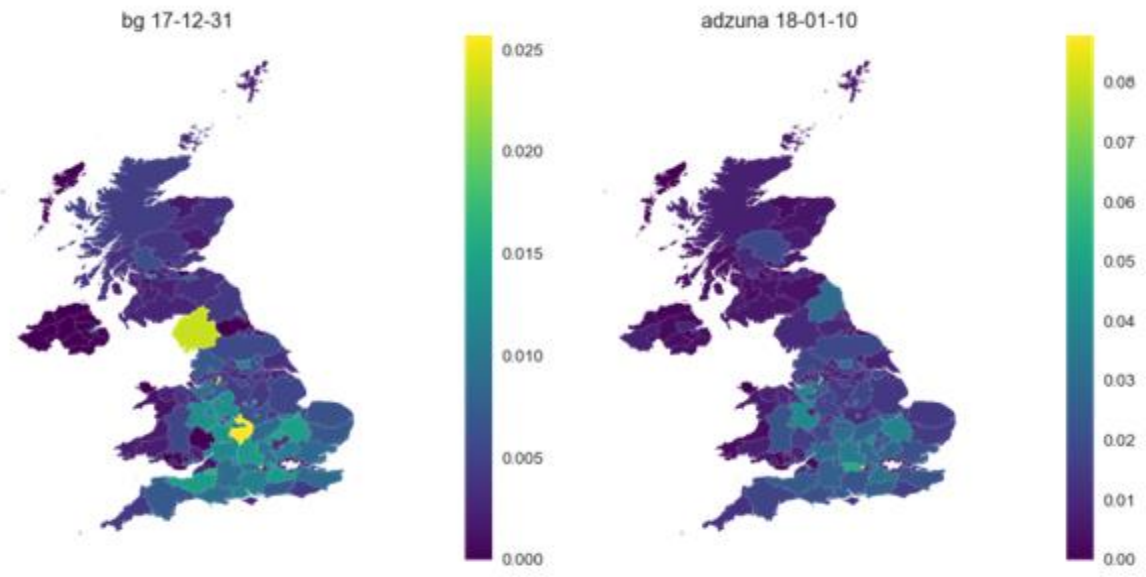
Figure 5: Nowcasts based on the S-ARIMA-X time series model.



6.3 Geographic indicators:

Figure 6 shows an example of the type of geographic indicators that could be possible with OJV data. Two different set of indicators are shown based on the two UK data sources for which location information is available (i.e. Burning Glass and Adzuna). These indicators represent the number of online vacancies as a proportion of the working age population in each local authority. Both sets of data show similar patterns with higher rates in Central and South England and generally lower rates in Scotland, Wales, Northern Ireland and peripheral areas of England. However, there are also some differences between the two sources, which illustrate the problems of relying on any one source of OJV data. There is then also the problem that these data may be giving a distorted picture of how local labour markets are performing. For example, in predominantly rural local authorities, employers may rely less on on-line channels. This is discussed further in Annex H (UK), Section 5.4

Figure 6: Number of job vacancies as a proportion of working age population[28]



6.4 Concluding Remarks

The results from Slovenia raise some fundamental questions about the OJV data and whether and how it should be used for official statistics. While this has shown that it is feasible to produce estimates of on-line job vacancies, there is a big difference between these and the official job vacancy estimates. It is therefore clear that these could never replace the official estimates. Further, the benefits of these estimates for policy making are not clear as they only give a partial (and not easily defined) view of overall labour market demand.

It therefore seems that the role of OJV data within official statistics is more likely to be as the basis for producing supplementary indicators. These could include indicators of local labour market demand (as shown above) and/or indicators of occupation groups and associated skills. However, rather than measuring absolute levels, these would be more useful for measuring change over time. Using OJV data for nowcasting purposes is another promising application. We also cannot yet rule out the possibility of using OJV data in conjunction with the JVS to reduce the frequency of the survey or reduce the size of survey samples and thereby reduce sampling costs. Another possibility could be using these data for imputing non-response in the JVS. However, it is important to recognise the considerable differences between different countries in terms of the OJV landscape. Thus, what is feasible in one country may not necessarily be reproducible in others.

[28] Data for London has been removed due to the distortive effect on the scales

7 Future Perspectives

The ESS Big Data Steering committee has agreed that on-line job vacancies will be one of four implementation work packages as part of the second Big Data ESSnet starting in early 2019. Therefore, this will be the organisational framework for taking this work forward within the ESS over the next few years.

It is expected that the current CEDEFOP web-scraping project will form a common supporting infrastructure to support the adoption of OJV data or the ESS during the second ESSnet and beyond. This means that NSIs may be able to reduce their activities around data access and data handling and focus more on the challenges around further methodological development. The ESSnet will need to continue to work with CEDEFOP to ensure that the data meets the needs of official statistics as far as possible.

However, one important limiting factor is that this system will only hold data from 2018 and so it will take several years at least to collect a reasonable time series. This will constrain what could be delivered within the timescales of the next ESSnet. This, coupled with the various challenges in using OJV data for official statistics, means that we need to be realistic about what is achievable in second Big Data ESSnet. In addition, there is a need to clarify what is meant by “implementation”, what could reasonably satisfy this expectation within the stated timescales.

One final future perspective is to consider how technology and recruitment trends may impact on OJV data. For example, there are now websites which allow business to create video job ads that can be targeted to individuals based on their browsing history. If this were to start displacing traditional job portals, then this would raise a whole new of technical and methodological challenges.

References

Carnevale A., Jayasundera T., Repnikov D., 2014, "Understanding on-line job ads data: A Technical Report", Georgetown University; Available at: https://cew.georgetown.edu/wpcontent/uploads/2014/11/OCLM.Tech_.Web_.pdf (Accessed 25 June 2017)

Djumalieva, J., Lima, A., Sleeman, C., 2018, "Classifying occupations according to their skill requirements in job advertisements", NESTA, Available at: https://www.nesta.org.uk/sites/default/files/classifying_occupations_according_to_their_skill_requirements_in_job_advertisements_28-03-2018.pdf (Accessed 09 April 2018)

De Pedraza, P., Visintin S., Tijdens, K., Kismihók G, 2017, "Survey vs scraped data: Comparing time series properties of web and survey Institute for Advanced Labour Studies, Working Paper Series, Available at: https://aias.s3-eu-central-1.amazonaws.com/website/uploads/1499760002407WP-175---de-Pedraza,-Visintin,-Tijdens,-Kismih%C3%B3k.pdf (Accesses 09 April 2018)

Stateva, G., Ten Bosch, O., Maslankowski, J., Righi, A., Scannapieco, M., Greenaway, M., Swier, N., Jansson, I., Wu, D., 2016, "Legal Aspect to Web Scraping of Enterprise Websites", Eurostat; Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WP2_Deliverable_2_1_15_02_2017.pdf (Accessed 29 June 2017)

Ketner, A. and Vogler-Ludwig, K., 2010, "The German Job Vacancy Survey: An Overview" in "1st and 2nd International Workshops on Methodologies for Job Vacancy Statistics, Proceedings", Eurostat; Available at: http://ec.europa.eu/eurostat/documents/3888793/5847769/KS-RA-10-027-EN.PDF/87d9c80c-f774-4659-87b4-ca76fcd5884d (Accessed 24 Oct 2016)

Körner, T., Rengers, M., Swier, N., Metcalfe, E., Jannson, I., Wu, D., Nikic, B., Pierrakou, C., 2016, "Inventory and qualitative assessment of job portals, Eurostat; Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_draft_v5.docx (Accessed: 1 November 2016)

Swier, N., Metcalfe E., Jansson, I., Wu, D., Nikic, B., Pierrakou, C., Körner, T., Rengers, M., 2016, "Interim Technical Report (SGA-1)", Eurostat; Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/64/WP1_Deliverable_1_2_final.pdf (Accessed: 31 July 2017)

Annex A: Belgium SGA-2 Report

7.1 Introduction

We consider and overall aim of reducing the number of surveys from four to just one annual survey during the first quarter to reduce the administrative burden on enterprises and the survey costs. To achieve this goal, we want to see if it is possible to predict the number of job vacancies and occupied posts for the other quarters with a prediction model. A prediction model of the job vacancies for 2017 is currently being developed based on data from the four employment public services (Actiris, ADG, Le Forem and the VDAB) for the period 2013-2016.

However, as this work is not yet complete, this report focuses on solving a specific challenge with this kind of approach namely, how industry codes could be derived from the content of web scraped job ads, in place of data collected directly from the survey. If some labour market developments are not properly reflected in the data collected and disseminated by the employment public services (EPS), the analysis of job vacancies available on the Internet may complete the prediction model for certain economic sectors.

We describe in this report the web scraping procedure tested out by the DG Statistics and we present the data collected in this way. As data from the Internet are not sufficient as such to estimate the development of job vacancies by economic sector directly, we present a machine learning technique to model the economic activity of an enterprise from the detailed description of the text of its job advertisements. To test this technique, we rely again on the administrative data provided by the EPS.

7.2 Data Access

We already receive administrative datasets from regional employment agencies containing job vacancies between 2013 and 2016. Based on the unique and official register number of each enterprise, we can be confident about linking a job description, the NACE code of the enterprise and the job location. This linking has already been completed for Brussels and is ongoing for Flanders and Wallonia. From now, we used data from Actiris (Brussels employment agency), with 90 372 job vacancies with the NACE code of the company, the regional location and the description of the job. We advise to take into account the administrative procedures and the delay. Using these datasets in statistical production system require conventions to avoid any problem in the future.

We also scrape job vacancies from 10 job portals. We choose these job portals from our job sector knowledge. We aim to expand to more portals and to scrape each trimester all job vacancies. We also plan to obtain agreements with some portals to obtain the job ads more easily and to avoid technical problems during web scraping. We have been blocked by two job portals, but we plan to contact them to obtain an agreement. This web data set takes time to be collected, and any change of the web scraping methodology requires time to collect a new set of job vacancies.

Table 1. Job vacancies from web scraping by job portal

Job portal	URL	Number of job vacancies collected				
		Sept-16	Apr-17	Aug-17	Nov-17	Feb-18
Indeed	http://emplois.be.indeed.com/	16010	15950	16010		8573
Jobat	http://www.jobat.be/fr/	13586	13590	41310	39745	45540
Jobinga	http://www.jobinga.be/	2066				
Jobrapido	http://be.jobrapido.com/	11929	6588	420		
Monster	http://francais.monster.be	15145	14791	14323	14914	16125
Optioncarrière	http://www.optioncarriere.be/	13399	23497	11551		9989
References.be	https://references.lesoir.be/	1260	1240	1710	1827	1220
Vacature.be	http://www.vacature.com/	7250	1688	3366		
Stepstone	https://www.stepstone.be/fr/	4300	4900	100		2200
Selor	http://www.selor.be/	0	55	102		
Total		84945	82299	88892	56486	83647

Source : Statbel

We also tried to collect job vacancies directly from company website and to compare with the survey data. We scraped 10,000 companies using only a basic spider which searches for the name of the company on Google, then search for a job section, and then scrape any figures near to such terms as "job". The correlation between these results and the survey is less than 2% and so we didn't investigate further in this direction.

7.3 Data Handling

The first task was to clean up the web scraping datasets. Even if we collect only job vacancies from job portals, we need to check for duplicate inside a job portal and between them. For now, we only collect the url address of the job vacancies to check the duplicates. Table 2 show the rate of duplicates by job portals (in red the duplicates inside a job portal, in black the duplicates between two job portals). We noticed that some job portals have a lot of duplicates and so provide limited additional data. Some job portals are larger and can include a lot of vacancies from other smaller job portals.

We tried to detect the company inside the job offer without success. Job portals provide often the name of the company, but it is difficult to match this name with our register. Even if this was possible, the most frequent company in our dataset is a temporary employment agency. These advertisements may be for temporary jobs with the agency, sometimes because this agency recruits for another company. So we cannot use this information to link with the NACE code. We tried to detect the economic sector of the job offer, in vain too because the description is present for a very few part of our dataset.

Table 2. Rate of duplicates job vacancies by job portals

	indeed	jobat	Jobrapido	monster	optioncarriere	stepstone	references	vacature	selor
indeed	56%	69%	55%	34%	35%	47%	76%	52%	56%
jobat	69%	72%	71%	56%	58%	69%	77%	70%	72%
jobrapido	55%	71%	51%	22%	20%	25%	86%	38%	50%
monster	34%	56%	22%	19%	18%	17%	53%	19%	19%
optioncarriere	35%	58%	20%	18%	16%	14%	57%	16%	16%
stepstone	47%	69%	25%	17%	14%	0%	80%	8%	0%
references	76%	77%	86%	53%	57%	80%	90%	85%	90%
vacature	52%	70%	38%	19%	16%	8%	85%	21%	20%
selor	56%	72%	50%	19%	16%	0%	90%	20%	0%

Source : Statbel

Our dataset cannot identify the NACE code of each job vacancy but we need this information to replace (partially) the survey. We plan to use the administrative dataset to model a link between the description of a job and the NACE code of the company. The dataset from Actiris provide of lot of textual information about the job: description of the job, of the hiring condition, of the task asked, of the educational background required and of the description of the company. Unfortunately, even the last description is the more important for us, only 23% of the job vacancies provide this information. We decided to concatenate all the description to get a global description of the offer.

7.4 Methodology

The aim is to build a machine learning process to test the link between the job description and the NACE code of the enterprise, based on the databases from regional agencies. According to our first test, done during 2017, we expect to arrive at a robust model for some NACE codes, but probably not for all of them. We have to configure a model for the languages used in Belgium (tree official languages Dutch, French and German, plus English), and to make several tests in order to obtain stable and comprehensive models. From now, we only use the dataset from Actiris in French.

For each job vacancy, we want to determine the NACE code of the enterprise from the detailed description of the job vacancy. For this, we use the package RTextTools[29] from the software R. This package provides several functionalities to set up our machine learning programme. The package proposes the most commonly used machine learning algorithms in the field: « support vector machine », « glmnet », « maximum entropy », « scaled linear discriminant analysis », « bagging », « boosting », « random forest », « neural networks » and « classification tree ». It also makes it possible to apply machine learning with several of these algorithms, and then to choose as categorisation result the most

[29] https://cran.r-project.org/web/packages/RTextTools/index.html, retrieved on 8th May 2018. A detailed presentation of the package and algorithms is available here: https://journal.r-project.org/archive/2013/RJ-2013-001/RJ-2013-001.pdf, retrieved on 8th May 2018.

relevant result according to the algorithms or the result for which there is consensus. For the results that are presented here, we have used all the algorithms and proceeded by consensus. Some of them only have a limited categorisation power, and in the future we will be able to delete them to speed up the programme.

7.5 Statistical outputs

We notice that 65 % of the NACE codes have an F-score above 60 % and that 95 % of the NACE codes have a precision higher than 70 % (see table 3). The precision of our modelling is very high. This means that we are very often right when we assign a NACE code to a job vacancy. However, the recall is not as high, which means that we often "forget" many vacancies for a given NACE code.

Table 3. Percentage of NACE by recall, specificity and precision.

Critical value	Percentage of NACE codes for which the *recall* is equal to or higher than the critical value	Percentage of NACE codes for which the *precision* is equal to or higher than the critical value	Percentage of NACE codes for which the *F-score* is equal to or higher than the critical value
50%	60%	100%	80%
60%	55%	95%	65%
70%	35%	95%	35%
80%	25%	60%	30%
90%	5%	35%	0%
100%	0%	15%	0%

Source: Statbel, Actiris data processing, job offers 2013 to 2016

The model is very effective for the NACE codes D, H, O, P and Q. For these job vacancies, we can properly assign the NACE code without setting too many aside. Conversely, for NACE codes L, S, T and U, the model performs poorly and the predictions are not of good quality. Some NACE codes get a 100 % precision, that should be interpreted with caution (NACE A, E, U). Indeed, the number of job vacancies in this case is low. Consequently, determining the NACE code for these few job vacancies is not a sufficient quality sign. Besides, the recall rate for these NACE codes is relatively low, which proves that we often wrongly set aside job vacancies related to these codes.

Table X in annex makes it possible to analyse the model results in greater details and to understand the distribution of NACE codes assignations. For example, we notice that we assign many job vacancies from all economic sectors to the NACE code G. This may be due to the fact that the NACE code G concerns trade activities, which can concern many other economic sectors, at least in the terms used for a job vacancy. On the other hand, job vacancies from sector N (administrative and support service activities,

including temporary positions) are regularly assigned to other economic sectors. The reason for this may be similar.

In order to analyse these results, we can also get the terms used by the machine learning programme to determine the assignation of NACE codes. A term will be divisive if it often appears in job vacancies of one NACE code and, at the same time, cannot be found in job vacancies of other NACE codes. This way, the terms "employment", "qualifications", "enterprises" will often be set aside, because they are too present in all job vacancies. Conversely, we notice that the term "Selor" is divisive for the job vacancies under the NACE code O (Public administration). Indeed, the public administration job vacancies are disseminated via Selor and the description of functions almost always refers to it. Such an analysis will be conducted for each economic sector in order to interpret the results in greater details.

This first test of machine learning is conclusive for some NACE codes. The prediction capacity for these economic sectors is strong and the risk of error is relatively low. Conversely, for other sectors, machine learning does not seem to be able to predict the NACE code from the detailed description of the post. Therefore, for these NACE codes, it will be difficult to use data from web scraping. This analysis was conducted on Actiris data in French only. We still have to run the programme on all data available for all EPS, in French and in Dutch, before being able to conclude on the overall efficiency of machine learning.

We will publish a report on the correlation between words in job ads and NACE codes, and a quality analysis of the machine learning process to predict NACE codes for the Job vacancy survey (JVS)

Table 4. Precision matrix of NACE codes classification

		NACE code by Machine Learning																			Sensitivity
		A	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
Real NACE code	A	3	0	0	0	1	16	0	0	0	0	0	1	0	0	0	1	0	0	0	0	14%
	C	0	469	1	0	41	287	2	27	16	0	1	44	18	14	2	22	0	0	0	0	50%
	D	0	0	219	0	1	16	0	0	6	0	0	7	5	7	0	3	0	2	0	0	82%
	E	0	0	0	16	3	9	0	1	1	0	0	2	3	8	2	0	0	0	0	0	36%
	F	0	4	0	0	671	200	7	21	6	1	6	45	24	42	3	24	0	3	0	0	63%
	G	0	26	0	0	39	4320	29	120	85	1	2	118	37	29	10	44	1	18	0	0	89%
	H	0	4	0	0	11	213	1499	3	18	0	1	25	5	57	0	89	0	4	0	0	78%
	I	0	10	2	0	6	355	1	1879	3	0	1	23	19	40	3	37	2	4	0	0	79%
	J	0	14	2	0	14	386	6	11	1096	1	3	152	40	61	4	28	0	5	0	0	60%
	K	0	0	0	0	1	138	0	1	23	524	8	73	22	24	1	5	0	1	0	0	64%
	L	0	2	0	0	12	118	1	21	4	1	249	48	20	32	8	69	1	6	0	0	42%
	M	0	44	0	0	45	876	10	38	134	29	10	2847	130	103	14	106	2	14	0	0	65%
	N	0	18	1	0	49	699	14	35	106	8	8	313	2116	108	12	107	2	6	0	0	59%
	O	0	4	1	0	5	51	4	12	6	2	12	33	20	6950	362	156	1	6	0	0	91%
	P	0	3	0	0	7	59	2	10	7	0	1	57	14	295	5141	276	8	21	0	1	87%
	Q	0	8	0	0	10	167	9	56	12	0	4	42	60	375	187	5499	6	45	0	0	85%
	R	0	3	0	0	2	103	0	7	9	0	2	22	7	91	26	158	213	15	0	0	32%
	S	0	6	0	0	3	193	3	16	13	1	8	86	58	232	39	389	16	601	0	0	36%
	T	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	1	0	0	2	0	25%
	U	0	0	0	0	0	12	3	2	4	0	0	4	2	5	0	2	1	2	0	20	35%
Specificity		100%	76%	97%	100%	73%	53%	94%	83%	71%	92%	79%	72%	81%	82%	88%	78%	84%	80%	100%	95%	

Source: Statbel, Actiris data processing, job offers 2013 to 2016

Annex B: France

7.1 Data Access

Scraped data can be used to get an overview of job vacancies. In France, there is another way of accessing data with the French National Employment Agency ("Pôle Emploi"). They are collecting data thanks to partnerships with online job boards. For now these two possibilities are explored separately and should lead to slightly different results. During the production phase **using both** scraped data and administrative data from the French National Employment Agency could be an option.

a) Scraped data

During this project we have scraped **4 job boards**. For now, **no detailed preliminary study** was led in order to build a list of websites to scrape. The choice of scraped websites depended on:

- **Agreement** of websites (at least non-rejection)
- Technical **feasibility** of scraping
- **Number of job ads** in the website

Websites **are warned** that we plan to scrape information from their pages before scraping. In case of no answer or acceptance websites are scraped. In case of disagreement there is no scraping at all.

Technical feasibility of scraping implies that so far only pages with **html structuration** are potentially scraped. When data are already (semi-)structured in the webpage (e.g. different html tags that define different parts in job description like "enterprise who is hiring", "profile of the ideal candidate") we consider this structuration for data scraping: several columns are filled in.

We have chosen for now 4 websites with **many job ads**. Some selected websites have a partnership agreement with the French National Employment Agency (enabling comparison between scraped data and structured data received by "Pôle emploi"), some others have not. Scraping system is coded in pure Python. We only have collected data once in **July 2017**. Complete data from one website and partial data from 3 other websites were scraped. Work to get an automatic collection is in progress (see Section 4).

b) Data collected by the French National Employment Agency ("Pôle Emploi")

From 1998 till mid-2017, quarterly statistics concerning **labor market tightness were published**[1]. The main indicator computed was the number of new job ads over the number of new jobseekers registered in the French National Employment Agency, which is defined as:

$$\frac{\#(\text{flow of job ads})}{\#(\text{flow of new jobseekers})}.$$

Only **job ads collected** by "Pôle Emploi" were considered. Use of this data can be improved because of **coverage issues**:

- Not all job vacancies are advertised online. According to a French survey about recruitment processes[2] only 41 % of job vacancies are advertised online.
- Not all online job ads are collected by "Pôle emploi". We can estimate from 25 % to 40 % rate of online job ads that are collected by "Pôle emploi".
- Coverage of job ads by "Pôle emploi" may fluctuate due to changes in business models of online job ads providers. That may impact labor market tightness indicator in an unwanted way.

Since 2015 the French National Employment Agency has been developing "partnership agreements" with other websites that publish job ads **including job portals and enterprise websites**. For now ~140 websites are covered. Partnership agreements imply:

- Daily transmission of data to "Pôle Emploi".
- Dissemination of collected job ads on the Employment Agency website **after de-duplication and occupation coding**.

These agreements mean that we can now consider **job ads published** by "Pôle emploi" (higher coverage rate of online job ads) instead of **only job ads collected** by "Pôle emploi". Unfortunately partnership agreements don't cover **statistical purposes** and are meant for operational purposes only. Consequently:

- List of websites with agreements is evolving constantly.
- All job ads where occupation cannot be coded are not available for data analysis (and they are not published online).

For now we are studying for experimental work all **published job ads by the French National Employment Agency** in 2016, including job ads collected (directly) by the agency other job ads available due to partnership agreements. There are some issues about data quality:

- We have data **after de-duplication and occupation coding**. Job title is not available and replaced by occupation code. So far we don't have details about methodology used for de-duplication and occupation coding.
- We don't know where (which website(s)?) every job ad is published.
- Job description is truncated after 1024 characters.

This data can be used as **training and test data** when coding occupation (see Section 3).

7.2 Data Handling

After data collection first step before data structuration consists of data preprocessing. This is mainly based on **text cleaning**. So far text cleaning is processed with two different tools for job title and job description.

a) Job title

Preliminary steps (deletion of phone numbers, punctuation deletion, and removing stopwords) are made in Python with **regular expression detection associated to a dictionary of stopwords**.

After preliminary step job title is cleaned with **Morphalou lemmatizer**[3]. This approach is based on a dictionary with ~540 000 words and ~70 000 associated lemmas. Other pieces of information (e.g. word grammatical type, word gender…) are available within Morphalou but they are not used for text cleaning for now. Morphalou lemmatizer is only available for **French language**.

b) Job description

Job description often consists in sentences and some words may mean different things depending on their position in a sentence. That's why we use the TreeTagger[30] lemmatizer for job description. The TreeTagger is a tool for annotating text with **part-of-speech and lemma information**. The TreeTagger can tag **many languages** including German, English, French, Italian, Danish, Dutch, and Spanish. Basic idea of this tool is to associate a lemma to each word depending of part-of-speech tagging (see Fig. 1). Consequently it enables to give a different lemma to each word depending of position (~grammatical category) in the sentence.

Input of the lemmatizer is text without preliminary cleaning (because punctuation and stopwords are important for part-of-speech tagging). Removal of stopwords and punctuation is made **after lemmatization** for job description. At first glance, lemmatization with the TreeTagger seems to perform quite well.

Fig.1: Example of lemmatization - "The TreeTagger is easy to use"

[30] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, retrieved on 8th May 2018.

Sample output:

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

.

7.3 Methodology

In this part we present preliminary work concerning **job coding** in unstructured data. The goal is to predict occupation code for job ads posted online. Two methods based on respectively **string matching** and **machine learning** algorithms are tested depending on data availability. We have mostly used **jellyfish** and **sklearn** Python libraries.

Occupation code we want to predict consists in **22 categories** (most aggregated level of a French-specific nomenclature).

a) String matching for job title – Scraped data

First method is based on **string matching**. After cleaning of job titles (see Section 2), a Jaro-Winkler based distance is computed between each job title and a list of ~11 000 occupation titles defined by the National employment agency ("Pôle emploi"). The list of occupation titles is also cleaned with the same approach. The computed distance is:

$$distance(A,B) = \frac{half\_distance(\mathrm{A},\mathrm{B}) + half\_distance(\mathrm{B},\mathrm{A})}{2a}$$

Where:

$$half\_distance(A,B) = \frac{\#(Words\ a \in A, \exists b \in B, a \sim b)}{\#(Words\ a \in A)}$$

$$half\_distance(B,A) = \frac{\#(Words\ b \in B, \exists a \in A, b \sim a\ )}{\#(Words\ b \in B)}$$

And we consider that:

$$\forall a,b \in A \times B, a \sim b \Leftrightarrow JW\_distance(a,b) > 0.9$$

We use the **Jaro-Winkler distance** at **a word-level** in order to avoid too long computation times.

To find one match for each job title:

- The pool of potential donors is **firstly limited** with occupation titles (among the list of 11 000) with at least one (exact) word in common with job title of the ad.
- In this pool the best match is given by the occupation title with the highest *distance*. In case there is one perfect match (*distance* = 1), the algorithm stops.
- We also store the corresponding *best distance* for all matches.

Job occupation with **title matching** was made once for job ads scraped from one website representing ~60 000 job ads. With this method computation time is **around 2 hours** for title matching of all job ads.

Results of coding are **not evaluated** precisely so far as there is no available test dataset when using scraped data (see Section 1). Scraped data contains only job title information but there is no occupation code comparable with coding system used by "Pôle emploi".

b) Predictive algorithms – Data published by the French National Employment Agency

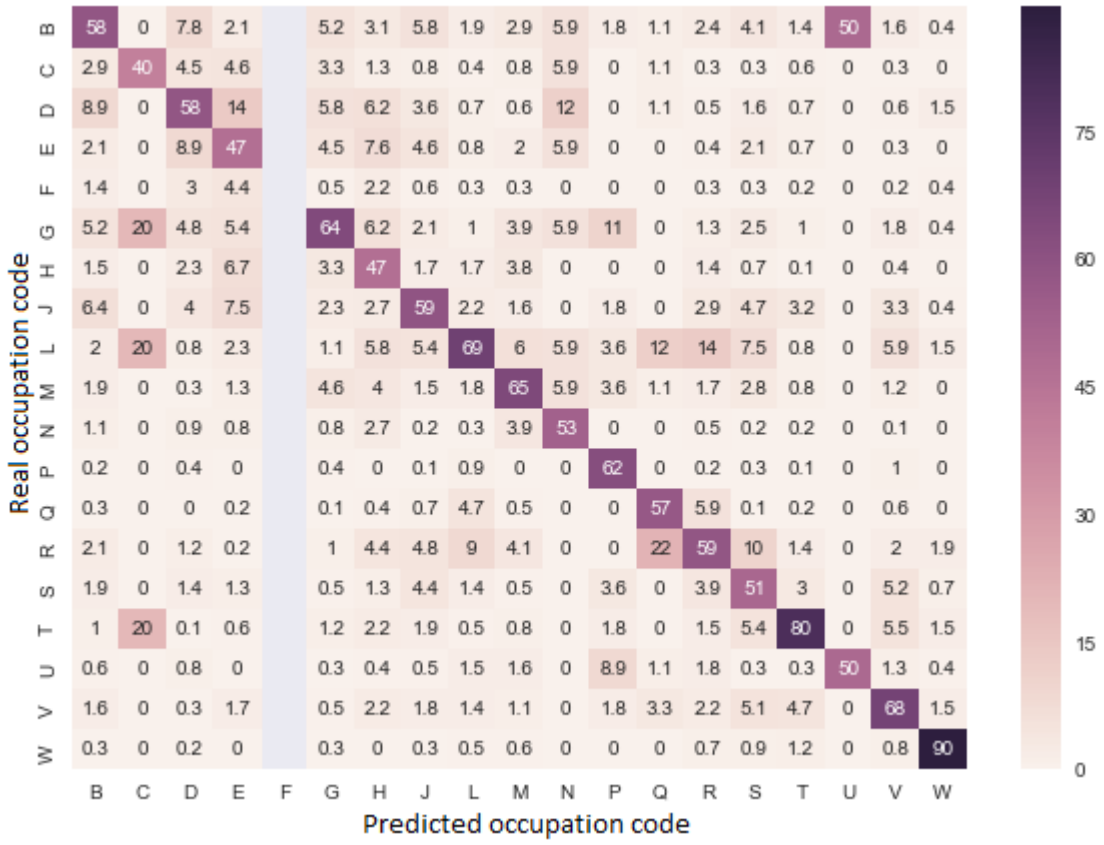
We have also tested predictive algorithms to code occupation for data collected by the French National Employment Agency. We use job description to predict occupation code.

The main steps include:

- Text description cleaning and lemmatization (see Section 2)
- Separation of job ads published by "Pôle emploi" between a training dataset and a test dataset. We consider a sample of 100 000 job ads.
- Computation of a document-term matrix with most frequent words. In the experimentation presented here we have kept ~1500 words. These words are used as **predictors**. We do not consider **additional variables**.
- Different supervised machine learning algorithms are then estimated and compared:
    - Logistic Regression
    - Random Forests
    - Neural network (multilayer perceptron)
    - Gradient boosting
    - Support vector machine

Initial results give a global accuracy rate of **~70%** for the test dataset. Fig. 2 presents results for precision detailed by category (confusion matrix) when estimation a logistic regression with L1 penalty.

Figure 2: Precision results when estimating occupation code with logistic regression



(F code is never predicted)

Further methodological work will involve:

- Improvement of **selection of words** used as predictors. The idea is to choose more occupation-specific words. So far we only have considered frequent words that can have a lower predictive power compared to words which are specific for the variable we want to predict.
- **Coding of other variables** like industry sector.

7.4 Statistical outputs

For now all results are **still experimental**. In a short term, we won't be able to publish statistics based on scraped data. Apart from further methodological work (see Section 3), next steps will consist in definition of **a more stable system** to collect scraped data. This will imply:

- Defining **a pool of websites** that will be scraped. The report from Cedefop project concerning France could be used to define this list.
- A more **frequent** and automatic collection.

7.5 References

[1] Bergeat M. (2017), "Les tensions sur le marché du travail au 2e trimestre 2017", *Dares Indicateurs*, available http://dares.travail-emploi.gouv.fr/IMG/pdf/2017-056.pdf, retrieved on 8th May 2018.

[2] Bergeat M., Rémy V. (2017), "Comment les employeurs recrutent-ils leurs salariés ?", *Dares Analyses*, available http://dares.travail-emploi.gouv.fr/IMG/pdf/2017-064.pdf, retrieved on 8th May 2018.

[3] Romary L., Salmon-Alt S., Francopoulo G. (2004), "Standards going concrete: from LMF to Morphalou", *20th International Conference on Computational Linguistics*, available http://www.cnrtl.fr/lexiques/morphalou, retrieved on 8th May 2018.

Annex C: Germany

Authors: Martina Rengers, Chris-Gabriel Islam

Table of contents

List of figures

List of tables

7.1 Introduction

The country report for Germany follows the basic framework – (1) data access, (2) data handling, (3) methodology, (4) statistical output – only very loosely. This is because the data sources have not been web scraped. Instead data provided by the European Centre for Development of Vocational Training (CEDEFOP: Centre européen pour le développement de la formation professionnelle) and the German Federal Employment Agency (FEA) are examined. Thus, the first point of the framework, data access, is covered in both of the subchapters belonging to each one of the data sets. The same holds for the third point, methodology. There the two data sets are mainly benchmarked against the job vacancy survey (JVS). The second point, data handling, is put into the annex, due to its very technical characteristics. Lastly, chapter 7.4 deals with theoretical statistical output. Even though there is none yet, possibilities and ideas are shown.

7.1.1 Data provided by FEA and CEDEFOP

In general, the FEA plays an important role in Germany: it hosts the largest and most important online job portal and is, furthermore, also responsible for the German Job Vacancy Survey (JVS).[31] Destatis and the FEA agreed to collaborate and one data set of the FEA job portal was submitted via a special web interface.

Another data set came from CEDEFOP which was formed in 1975 and is a decentralised agency of the European Union. Its main purpose is to support the European vocational education and training. Their focus lies on the mismatch of skills in the labour market. Therefore, they started a project that deals with scraping Online Job Vacancies (OJVs).[32] The goal is to measure employers' requirements and by that give advices to employment services and policy makers.

As CEDEFOP and the ESSnet are both European institutions there are ambitions to synergize. Like already proposed in Deliverable 1.1 (SGA-2)[33] the latter could use the data provided by the former in their paneuropean approach to scrape OJVs. Although the actual focuses of the two institutions differ from one another, both need a clean, deduplicated, representative data set for the whole European Union. That is the reason that CEDEFOP sent Destatis a prototype of their data in order to check and help improving it.[34] The transaction was very fast and uncomplicated.

---

31 For more details see Swier, Nigel et al. (2017), p. 28ff.

32 http://www.cedefop.europa.eu/en/events-and-projects/projects/big-data-analysis-online-vacancies, retrieved on 6th April 2018.

33 Swier, Nigel (2018), p. 1f.

34 This was a main point of the following conference: CEDEFOP (2018).

7.1.2 Measuring data quality of OJV data

In this report the data quality of OJV data is measured by a comparison with the current available job vacancy statistics.[35] For Germany this is – besides the other major data source, the statistics of registered job vacancies – the JVS which is carried out by the Institute for Employment Research (IAB – Institut für Arbeitsmarkt- und Berufsforschung). The survey is done by a quarterly collection and the written questionnaire of a fourth quarter of a year is followed by a short telephone survey in the three subsequent quarters. These follow-up telephone surveys include only a smaller number of variables than the written questionnaire of the fourth quarter and they are based on a smaller sample size as they are conducted as a sub-sample of the first wave participants. So the results of the fourth quarter are the most powerful results of the JVS.

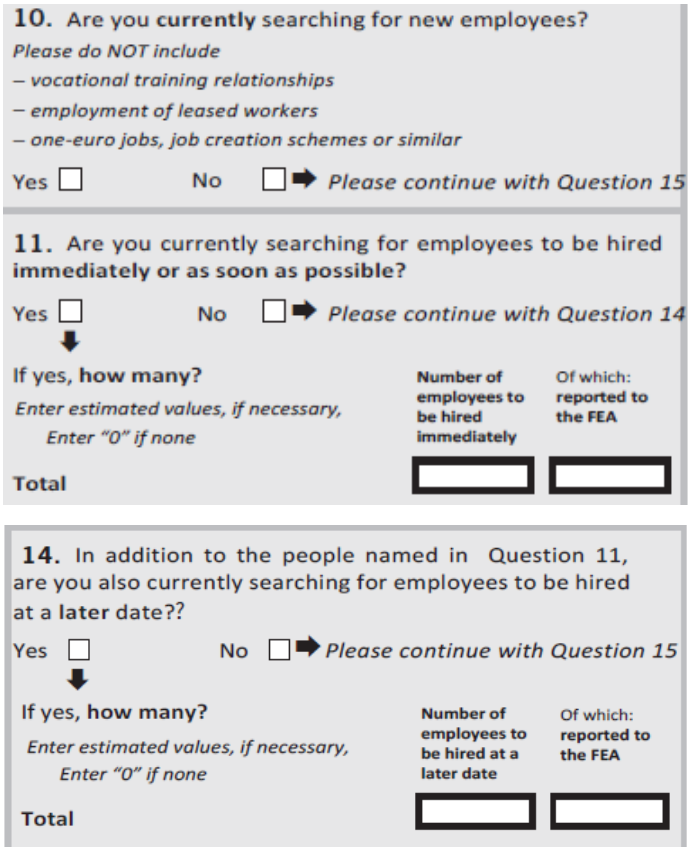
Compared on an international level the German JVS is the most detailed JVS of them all. Parts of its results are reported in Eurostat's job vacancy statistics. Other parts are published only on national level.

Figure 1 shows the relevant questions for statistics on job vacancies in the German JVS. There are two aspects one should have a closer look on:

1. **Currently searching:**
   In the JVS questionnaire personnel representatives and/or business managers with personnel responsibility were asked if they are *currently* searching for employees. As a consequence the quarterly results of JVS can more or less be interpreted as an average stock of job vacancies within a quarter.
2. **To be hired immediately or as soon as possible OR at a later date:**
   Question number 11 and 14 in Figure 1 show that it is not only asked for employees to be hired immediately or as soon as possible (number 11) but also for employees to be hired at a later date (number 14). The number of job vacancies is therefore the sum of the two figures given in both answers. This number is published by Eurostat but on a national level the IAB also publishes the number of jobs to be filled immediately.

[35] A description of the current German Job Vacancy Survey/Statistics can be found in Swier, Nigel et al. (2016) p. 28f.

Figure 1: Selected questions of the JVS



Source: JVS 2014q4 questionnaire.

When combing the OJV data to the JVS the above mentioned interpretation of JVS data as a stock measure has to be kept in mind. The difference between job vacancies and job advertisements and especially the fact that a job advertisement can contain more than one vacancy must also be taken in consideration [36]

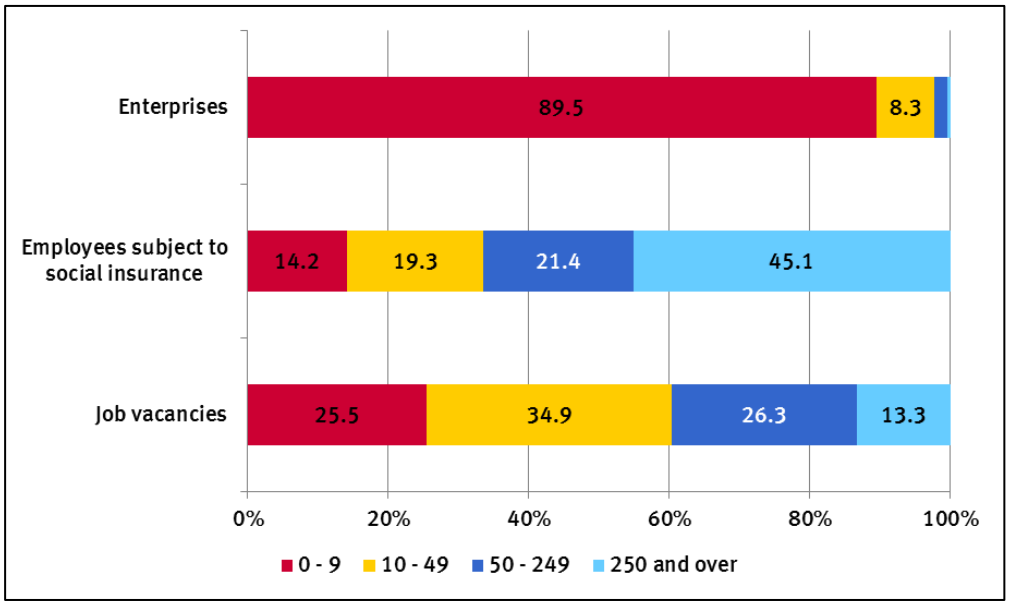
There is another data quality problem of OJV data that we do not handle in further detail in this report: the coverage problem. The coverage problem appears in two ways: first, the representativeness of online job vacancies for the whole job vacancies in the labour market and second, the representativeness of selected job portals for all online job vacancies. Although the total number of job vacancies in the labour market is unknown, the first coverage problem can be estimated by the German JVS as it also measures advertising channels used by employers. Furthermore there are existing studies from other institutes and it was found that larger enterprises are more likely to use online channels whereas small businesses are more likely to use traditional channels, such as print media.[37]

[36] Chapter „4. Job Vacancies versus Job Advertisements“ of Swier, Nigel et al. (2016) describes this in more detail.

[37] See Swier, Nigel et al. (2017), p. 34f.

Keeping this in mind it is important to have a short overview of the number and the size of enterprises in Germany. Due to the Business Register (BR)[38] there were almost 3.5 million enterprises in 2016. Most enterprises (89.5 %) have fewer than ten persons employed as Figure 2 shows. Altogether the number of employees subject to social insurance had a high of 29.5 million where 45.1 % of these employees worked in enterprises with a size auf 250 employees or more and 14.2 % in enterprises that have less than ten persons employed. The JVS counted in 2016q4 nearly 1.1 million job vacancies from which only 13.3 % come from larger enterprises with 250 and more employees but 25.5 % came from the smallest enterprises. This means that enterprises with 250 and more employees have a lower job vacancy rate than smaller enterprises.

Figure 2: Selected data by number of employees of enterprises 2016



Source: Own calculations on Business Register 2016 and JVS 2016q4 data.

7.2 OJV data from FEA

7.2.1 Data Access

The data Destatis got from the FEA represent the whole stock of job ads on their portal from the 26th of February 2018. The data was delivered in two files, where one file was the main document with over 75 variables and the other document contained the locations of the job ads. Both files could be linked via a common ID, but since sometimes in one ad several persons in several cities are searched for, it was a one-to-many-relation.

[38] More information about the German Business Register can be found in Swier, Nigel et al. (2017), p. 33f.

Besides the full text of the job ad a number of other variables were delivered, e.g. job title, industry sector, occupation, number of persons searched for, size of company and working hours. A full description of the data can be found in the annex of the German Country Report on Table 6 to Table 8.

There is a variable for the earliest possible date of entry by which it can be seen that there are ads from June 2016 and possibly beyond (since the date of entry should be scheduled after the publication date). This means that very old data exist in the data set which has to be handled and filtered for. In the next chapter follows a detailed explanation about the created subset which was analyzed. It is worthy to note, that not all jobs ads are original from the FEA. Some ads are provided by allied partners, mostly via a special portal. These partners are either companies or other job portals.

7.2.2 Methodology

7.2.2.1 Comparison with JVS

In order to guarantee a sound analysis one has to focus on a certain time period and filter for ads that are too old. Like already written in Deliverable 1.1 (SGA-1)[39] assuming a maximum duration of 30 days is a valid assumption for OJVs. Unfortunately, the data set does not contain a variable for the publication date of a job ad. So as to shrink the data set the only possibly useful variable is "EintrittVon" which represents the earliest possible date of starting the work offered in the job ad. By filtering for job ads with an earliest possible date of job entry between 27th January 2018 and 26th February 2018, around 34 thousand job ads were eliminated, as can be seen in Table 1.

Table 1: FEA-OJV versus JVS: number of job ads, job vacancies and jobs to be filled immediately

	FEA 26.02.2018	FEA 30 days	JVS 2017q1
number of job ads	470,374	436,506	–
number of job vacancies	748,901	694,820	1,064,000
number of jobs to be filled immediately	–	–	824,000

Source: Own calculations on FEA data and JVS data.

As a result 436,506 job ads and 694,820 job vacancies of the FEA data set still remain, i.e. there are 1.6 job vacancies per ad on an average. The amount of job vacancies is not by far the amount of job vacancies in the JVS. Note that the FEA data is compared to the first quarter of 2017 of the JVS data since there is no data for 2018 yet and the first quarter is seasonally spoken comparable to the FEA data set shrank to 30 days. This means a little difference could be explained by the time lag which on the contrary is rather unlikely since the demand for workforce is steadily increasing over the years. Interestingly enough, the number of job vacancies in the FEA data is not even equal to the number of job vacancies written on their online job portal (748 k VS ~1500 k). Hence, it is more likely that just some data is missing in the data set. Also, some differences occur due to the fact that there is a time lag between the expire date of a job ad and the entry date, meaning that companies in this time period would still answer in the JVS that they are searching even though there is no online job ad anymore.

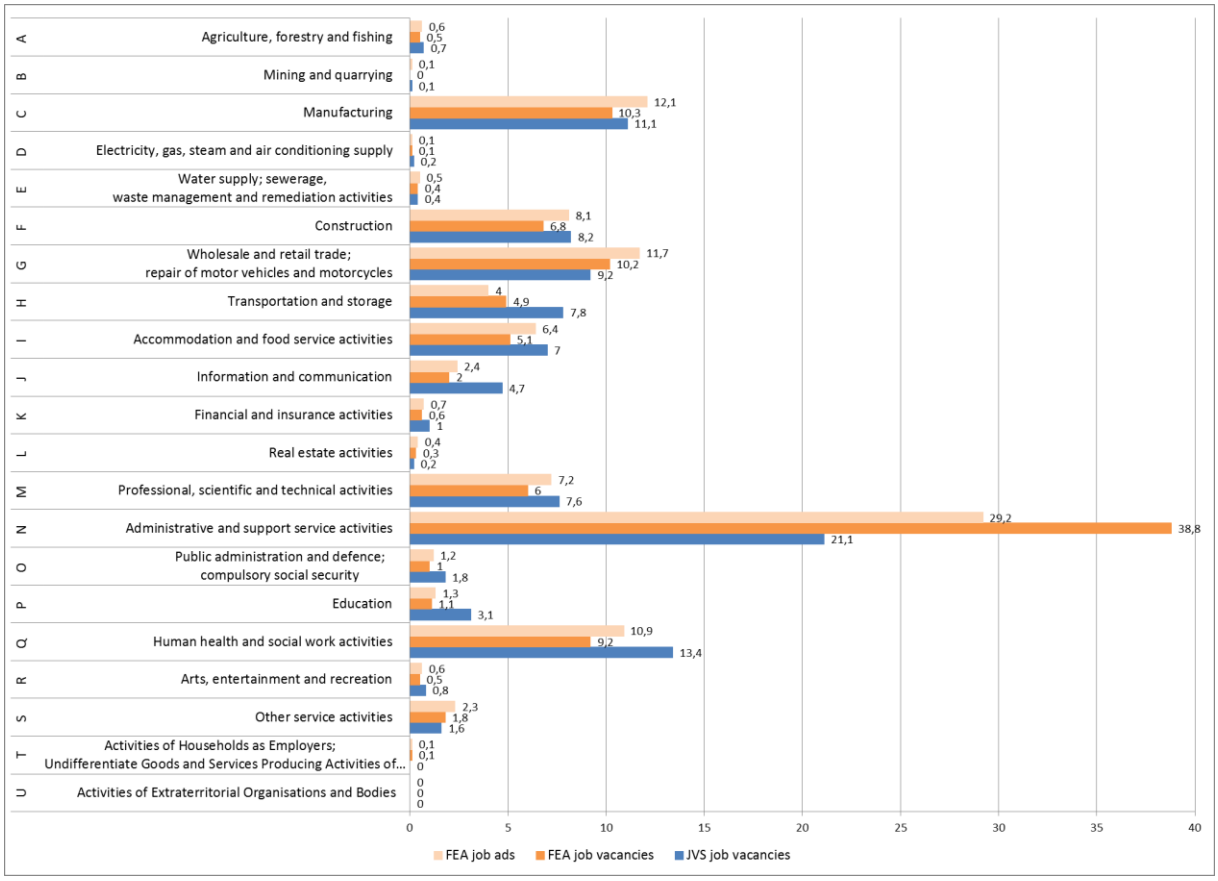
[39] Körner, Thomas, Martina Rengers et al. (2016), p. 15.

For all comparisons in the following sections, the data set in Table 1 named “FEA 30 days” is used and compared against the first quarter of 2017 of the JVS.

7.2.2.1.1 NACE

The classification of the industry sector of the employer in both data sets is coded by NACE although some text extraction had to be done in the FEA data first. Figure 3 shows the percentage distribution of the job ads and job vacancies by their industry sector. Missing values can be ignored since there are none for this variable. In general, all values seem to be somewhat close to each other. Sometimes even the distribution of the job ads is closer to JVS’ job vacancies than the job vacancies of the FEA. Note the varying differences between job ads and job vacancies per industry sector. This behaviour occurs because the number of jobs per ad varies between the industry sectors, meaning that for some sectors it is more likely to formulate job ads in which more than one person is searched for than for other.

Figure 3: FEA-OJV versus JVS: job ads and job vacancies by industry sector NACE (30 days and 2017q1), in %



Source: Own calculations on FEA data and special analysis on JVS data by IAB.

The biggest dissimilarity is found in sector N “Administrative and support service activities”. Having a more detailed look the origin of this problem can be located. In Table 15 (annex of this country report) the distribution of the FEA data is split into original data from the FEA and data that is provided by other allied partners. It becomes clear that the difference origins from the other allied partners.

So, as it will also be shown in chapter 7.2.2.1.3 for the ISCO structure, the NACE structure of the original FEA data is actually very close to the JVS data. Even more, considering that the time window is actually another one. Having such a good result, the question about the computation of this variable arises. This is actually unknown. It might be assumed that the FEA uses a sort of classification server based on the employer's name. Hence, there are no missing values in the data. Either case, having access to such a classification server would be a good base for machine learning, which is described in chapter 7.3.2.2.

7.2.2.1.2 ISCED

Comparing the two data sets on educational level is difficult, because there is no single variable for this but there are multiple. Following the numbers of Table 6 variable 16 to 18 and 62 to 64 would be suitable as they resemble the required academic degree or vocational training. Sadly, they do contain too much missing values, e.g. the variable "Hochschulabschlussart" for the required academic degree with 93 %. Furthermore even by combining all the variables you would not get a variable that is comparable to the variable in the JVS which is following the ISCED-11-Classification. In the end, there is only one possibility to compare FEA and JVS data which would be by translating the occupation coded in ISCO-08 to ISCED-11. This is possible, but results in strange numbers. By now, it cannot be said if the method is just wrong or if there is something odd in the data set. Anyhow, since this comparison would just be a proportional translation of the ISCO-comparison, the additional value is questionable. For this reason, this comparison is skipped at this moment.

7.2.2.1.3 ISCO

Comparing the FEA data with the JVS on the occupation level requires numerous steps beforehand. This is due to the several different classifications. Since the JVS is following ISCO-08 and the FEA data only shows the occupation in a FEA-intern classification called DKZ ("Dokumentenkennziffer") a transformation is needed. First, the transformation of DKZ to BKZ ("Berufskennziffer") is performed by sticking to a transformation table from the FEA website[40]. This BKZ then can be translated without problem to a code called KldB 2010 ("Klassifikation der Berufe"). In the last step, a transformation from KldB 2010 to ISCO-08 must be done. The FEA provides a table[41] for this, but unluckily the transformation is not unique. One occupation of KldB may belong to several of ISCO (around 25 %). Now, one could propose a number of solutions like e.g. distributing these cases equally on each occupation slot. In the end, the method already used in the Labour Force Survey was chosen assigning just the first option if several options exist.

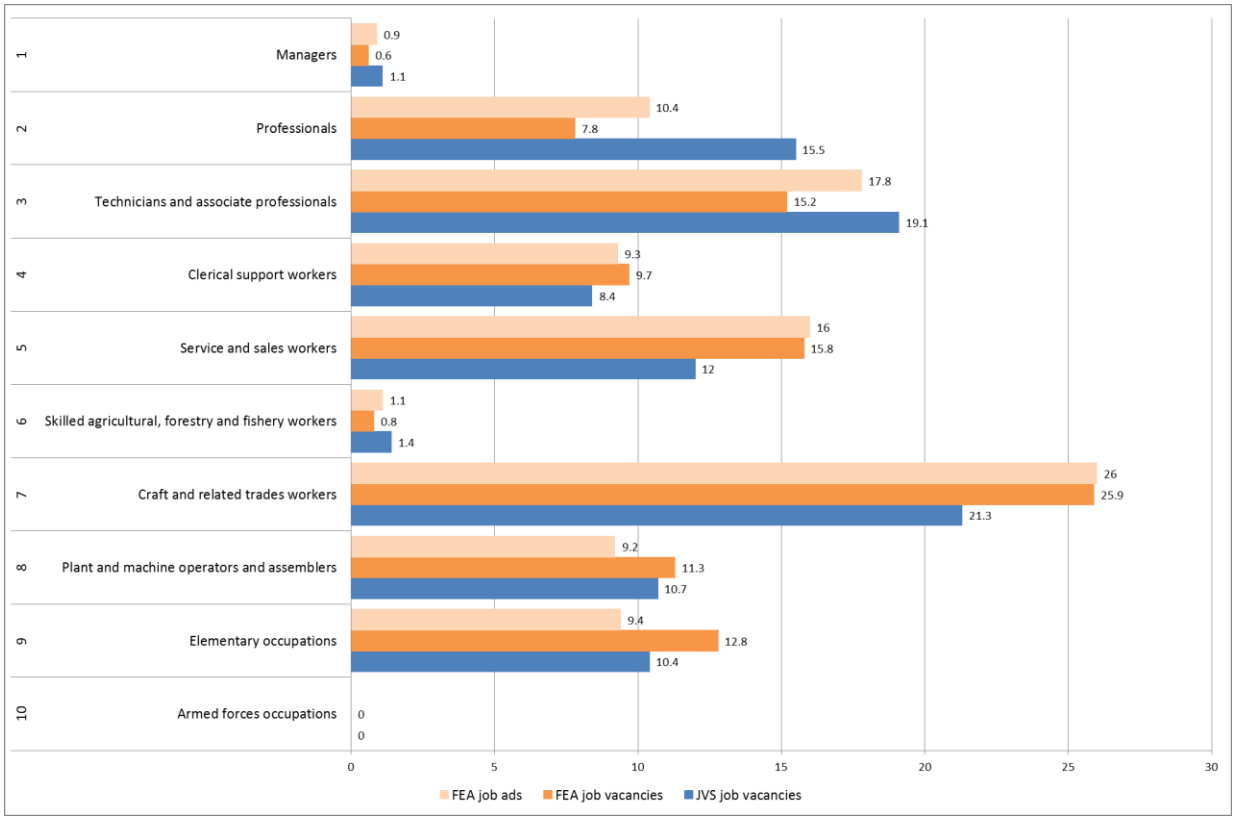
The result of this transformation and comparison can be found in Figure 4 or in a more detailed way in Table 14 in the annex of this country report. It can be seen immediately that the FEA data benchmarks quite well against the JVS data. Sometimes even the ISCO-structure of the job ads in the FEA data approximates better the ISCO-structure of JVS job vacancies than the job vacancies in the FEA data do. This may indicate that the distribution of the variable for job vacancies per job ad ("AnzStellen" in Table 6 in annex of this country report) is not similar to the real distribution. But in general, besides the group of professionals (ISCO group 2), neither the count for job vacancies nor for

40 FEA (2018a).

41 FEA (2018b).

job ads performs too badly having a maximum difference in the shares of around 5 %. Looking at the split between original FEA data and data from allied partners in Table 14, it is notable that the same statement even holds for the group of professionals. This means that the proprietary FEA data correlates very well with the JVS data on this feature.

Figure 4: FEA-OJV versus JVS: job ads and jobs to be filled immediately by ISCO (30 days and 2017q1), in %



Source: Own calculations on FEA data and special analysis on JVS data by IAB.

7.2.2.2 Strengthening

Wrapping it all up, the FEA data is quite good even though it was compared with JVS data from the previous year and there was no true variable for the publication date in order to divide the data set correctly. Having this said, it would be nice to have more data like this from previous or even future years in order to do more comparisons.

Nevertheless, the FEA data is far from perfect. Actually, there are some important variables that were ordered in the cooperation process but which could not be delivered at that moment. These variables are listed below:

- Publication date
- URL of job ad
- Work experience

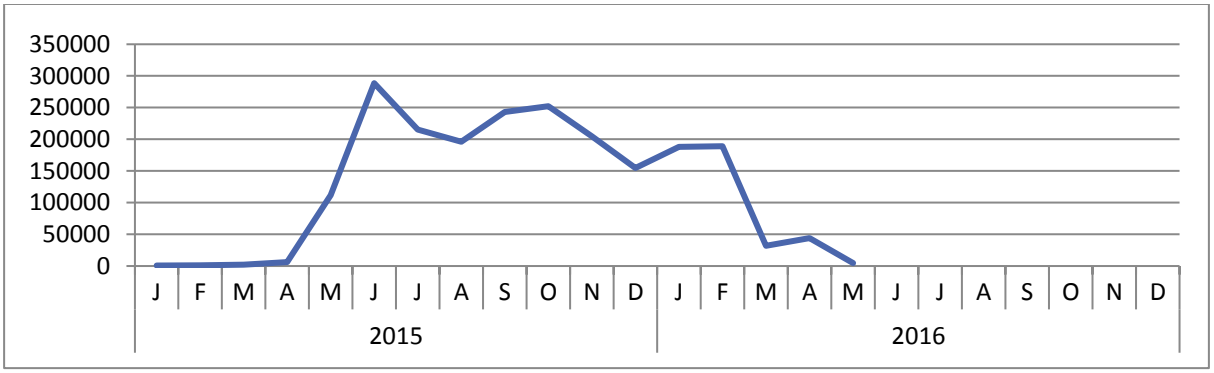
Moreover, some variables have too much missing values or do not follow an explicit categorization. This is due to the fact that an employer while posting a job ad has to fill out forms some of which are not obligatory and do allow entering free text. So, in order to get more structure in this sense, the FEA would have to change the interface of their portal. And indeed, at the moment there are two online job portals provided by the FEA since a newer one is ran parallel to the old one. However, especially the fact that employers have to type in manually their ads makes this data set of higher quality and therefore a very good training set for Machine Learning (cf. chapter 7.3.2.2).

7.3 OJV data from CEDEFOP

7.3.1 Data Access

Figure 5 gives a little more details about the time sequence of the scraping process. In the main file there is a variable called 'GrabDate' that is either the publication date of an OJV if it is available or otherwise the actual scraping date. Obviously, this double definition of a variable without a flag indicating if the date represents the publication date or the scraping date is unsuitable for further, more detailed analysis. But nonetheless, the variable is a vague indicator – or better: an upper bound – for the publication of an OJV.

Figure 5: CEDEFOP-OJVs in Germany per GrabDate



Source: Own calculations on CEDEFOP data

It can be easily seen that there are some problems with the scraped OJVs. Despite the fact that the scraping progress took only place between June 2015 and February 2016 the data set comprises OJVs from before that time until even from 2008 (not shown). This means that the scraped job portals include very old OJVs. But on the contrary they also include OJVs from May 2016 which should be impossible.

The main issue concerning the CEDEFOP data set is its unstructured form. Although the main file has 17 variables, the first 3 are just identifiers. Then, each date variable consists out of four variables: the actual date, the year of the date, the month of the date and the day of the date. Following the list of important features in Deliverable 1.1 (SGA-1)[42] that can be found on job portals the only useful and trustworthy information are the job description and the job title since the other variables hardly

[42] Körner, Thomas, Martina Rengers et al. (2016), p. 8.

contain any values. Unfortunately the job description is not structured. It is only a string with mostly several thousand characters. On the upside, doing web scraping in a way like this is very easy to program. But on the downside, this makes an analysis really difficult afterwards.

All data in the main file is collected via web scraping. The data in the other files are computed with the aid of Machine Learning Algorithms or are just metadata information. A more detailed data set description can be found in the annex of this country report on Table 9 to Table 11.

Table 2: Job ads of CEDEFOP data by job portal

Source	Frequency	Percent
DE_GIGAJOB	417,862	16.98
DE_MEINESTADT	1,036,645	42.12
DE_ONLINESTELLENMARKT	1,006,886	40.91

Source: Own calculations on CEDEFOP data.

Table 2 shows how many job ads come from which job portal. In total only three job portals were scraped. From MeineStadt and OnlineStellenmarkt around 1 million ads were scraped where Gigajobs contributes with less than a half of it. However, according to Deliverable 1.1 (SGA-1)[43] there were more than 1,600 job portals in Germany in 2016. Considering this, it is more than obvious that the data set has a coverage problem. Since the FEA is the biggest player in Germany's online job market according to the number of own job ads, a major issue is that data of the FEA is missing. This is due to the ranking CEDEFOP performed. The job portal of the FEA is only ranked as number 4 but they decided to scrape only the first 3.[44] As all scraped job portals are hybrid portals and therefore additionally use search engines some OJVs from the FEA are also listed. Nevertheless, it is not enough to cover the whole stock and therefore when producing the next prototype CEDEFOP should scrape the job portal of the FEA by all means. Alternatively, CEDEFOP could also arrange a cooperation with the FEA similar to ours (cf. chapter 7.2).

7.3.2 Methodology

Likewise to chapter 7.2.2, the focus has to be not on the full data set but on the restricted subset. Looking at Figure 5 it is clear that the data of June 2015[45] serves as a good foundation since with 288,585 entries this is the biggest amount by month (see also Table 3). Similarly, one has to harmonize this time period with the time period of the benchmark which is the JVS. Since the JVS is a quarterly survey, the time period to compare would be 2015q2. But this is problematic in two ways: First, one would compare data based on a monthly measurement period with data based on a quarterly measurement period and therefore ignore two whole months, i.e. April and May 2015. Second, many data in the JVS are only available for the fourth quarter of a year, because some

43 Körner, Thomas, Martina Rengers et al. (2016), p. 12.

44 CEDEFOP (2016), p. 2.

45 Here and for any other subset of the CEDEFOP data the variable GrabDate was used in order to separate the data. On the one hand, this is questionable since the GrabDate is not well defined and either is the publication date (if available in the ad) or the scraping date. On the other hand, this was the only option.

questions are only asked in the written questionnaires of the fourth quarter and not in the short follow-up telephone interviews in the following three quarters (cf. chapter 7.1.2 or more detailed in Deliverable 1.2 (SGA-1)[46]). That is why some more subsets of the data were analysed. The focus was on Q4/2015 in the JVS and it was compared to the CEDEFOP data of the same time period. Actually the way of gathering this data was not unique considering that there exist stock and flow data. So, one possibility would be just to accumulate all job ads that have a grab date[47] that lies in Q4/2015. This would resemble flow data. Another possibility would be to point out the currently active job ads of last day of each of the three months in Q4/2015 and divide the result by 3.[48] This would resemble stock data. In the end, analysis on both of them was done even though the German JVS (not the European in general) is more like a stock data.

Table 3: Job ads of CEDEFOP data by job portal and data subset

CEDEFOP	June 2015		2015q4 accumulated		2015q4 3 end of months average	
Source	freq	%	freq	%	freq	%
DE_GIGAJOB	33,203	11.51	99,983	16.34	42,149	18.17
DE_MEINESTADT	161,099	55.82	305,553	49.94	117,544	50.66
DE_ONLINESTELLENMARKT	94,283	32.67	206,317	33.72	72,339	31.18
Total	288,585	100	611,853	100	232,032	100

Source: Own calculations on CEDEFOP data.

As shown in Table 3 the distribution between the different job portals does not vary too much from subset to subset. Nevertheless, the total amount indicates that accumulating might not be the right way since it is triple the amount the other sets. Therefore, the following analysis focusses on the 3 end of month average which resembles the way the JVS is done the most since it is stock data.

7.3.2.1 Comparison with JVS

In order to be coherent to the previous chapter, the same comparisons with the CEDEFOP data as with the FEA data was done. Note in Table 4 that neither the stock version nor even the accumulated version is nearly close to the total amount of OJVs following the JVS, which reveals that the CEDEFOP data does not cover the online job market at all, also considering that there is a difference between job ads and job vacancies. But even if the total amount is not comparable there might be a distributional correlation.

46 Swier et al. (2016), p. 28ff.

47 Cf. footnote 45.

48 Keep in mind that the variable 'GrabDate' contains either the publication or the actual scraping date. So in order build this kind of stock data set the only possibility is to gather all data that has a grab date before and an expire date after one of these days.

Table 4: CEDEFOP-OJV versus JVS: number of job ads, job vacancies and jobs to be filled immediately 2015q4

	CEDEFOP 2015q4 3 end of months average	JVS 2015q4
number of job ads	232,032	–
number of job vacancies	–	1,047,079
number of jobs to be filled immediately	–	822,800

Source: Own calculations on CEDEFOP data.

7.3.2.1.1 NACE

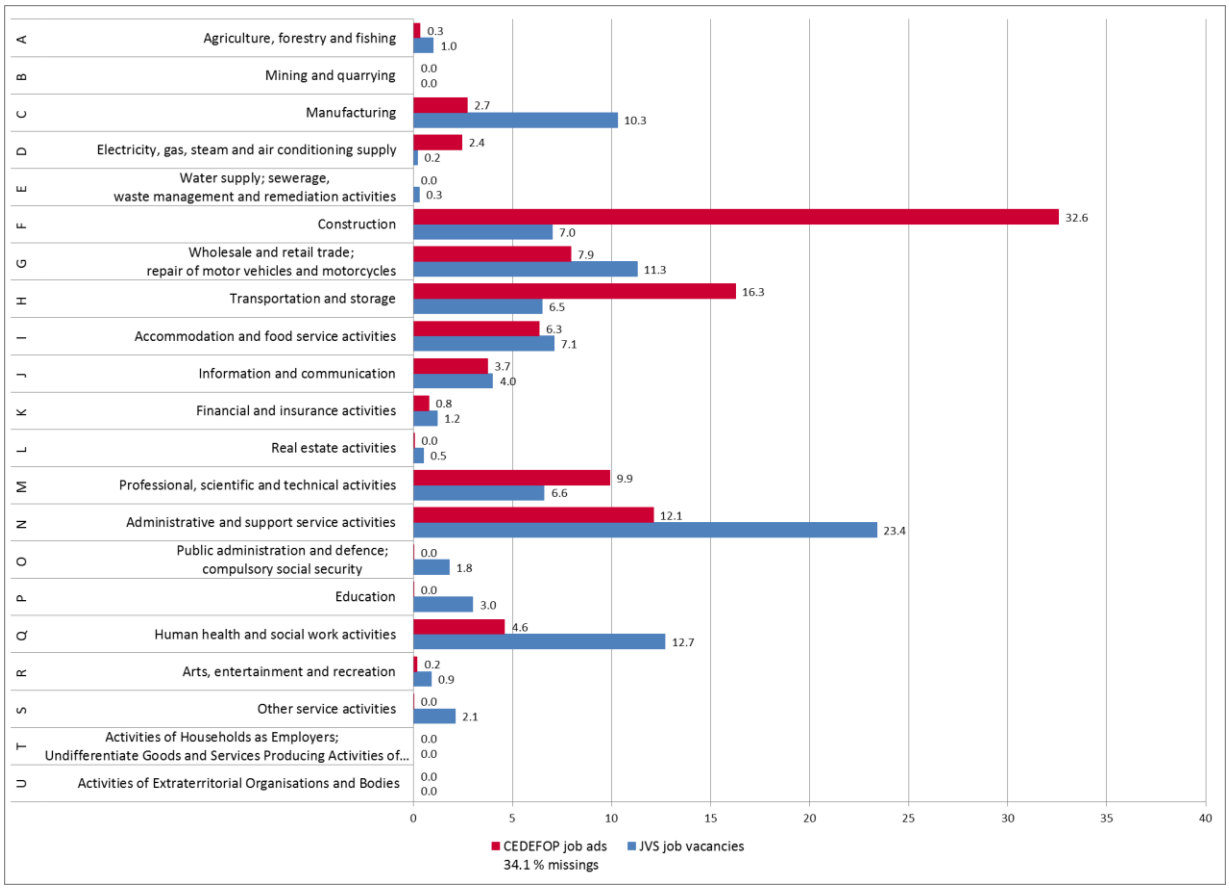
Comparing CEDEFOP data with JVS data based on the industry sector of the employer reveals several problems as can be seen in Figure 6 and –with detailed figures– in Table 15 of the annex of this country report. First, more than a third of CEDEFOP's NACE data is missing (34.1 %). Second, there are three big gaps regarding the NACE structure in the sectors F (25.6 percentage points), N (11.3 percentage points) and H (9.8 percentage points). Third, although the percentage of job vacancies due to JVS in sectors B, E, L, O, P and S is also very low job ads do not occur at all in the CEDEFOP-OJV data. Nonetheless, do not forget that CEDEFOP only provides job ads and no number of vacancies, unlike the JVS. In the same way as when comparing the FEA data to JVS in Figure 3 some differences can just be explained by this fact. Digging further, it can be discovered that no job ad scraped from OnlineStellenmarkt is classified and that most of the job ads in sector F come from the job portal MeineStadt and most of the job ads in sector H from Gigajob (cf. Table 15).

Last but not least, one should not forget that this is a comparison between online job ads versus job vacancies in general, meaning that there is a difference between the demand of the online and the whole job vacancy market. Carnevale et al. compare this difference per industry sector, discovering that for transportation and storage (H) the number of job vacancies on the whole labour market is lower than the online similar to our comparison in Figure 6. But unfortunately this does not hold for construction (F) and manufacturing (C) where the result of Carnevale et al. is just vice versa to our findings.[49] Remember here, that in earlier publications from Germany a comparison with data from the job portal StepStone and the JVS data was done. It was revealed that scraped data contains three times more ads for Information and Communication (J) than there were vacancies in the JVS data.[50] This is not the case for this comparison. It is difficult to say, which result is the true one. Most probably both of them are wrong due to missing coverage by the scraped job portals.

[49] Carneval, Anthony P. et al. (2014), p. 14.

[50] Rengers, Martina (2018), p. 79.

Figure 6: CEDEFOP-OJV versus JVS: job ads and job vacancies by industry sector NACE (2015q4), in %



Source: Own calculations on CEDEFOP data and special analysis on JVS data by IAB.

7.3.2.1.2 ISCED

The comparison of the CEDEFOP data with the JVS data regarding the educational level following the classification ISCED-11 is shown in Table 5. The classification was grouped into five classes due to the questionnaire of the JVS which did not ask for the educational level in a more detailed way. Once again, a lot of missing values in CEDEFOP's data had to be dealt with (here: 46.2 %). On the one hand, it can be seen that the missing values origin mainly from Gigajob and OnlineStellenmarkt. On the other hand, the ads of MeineStadt are nearly fully classified as Bachelor as educational requirement which is the reason for a gap of more than 90 % in comparison with the JVS data. Having such a big overestimation in one category means that there are also several underestimations in other categories like e.g. in short-cycle tertiary education or in post-secondary non-tertiary education.

This bad result is probably due to the classifier CEDEFOP used. Concrete, it is not yet evident, how they classified the educational level per job ad. Most of the times there were little to no information about it in the job ad. Considering that even the FEA has no true, uniform classification for the educational level, the result above becomes clear.

Table 5: CEDEFOP-OJV versus JVS: job ads and job vacancies by educational level ISCED (2015q4), in %

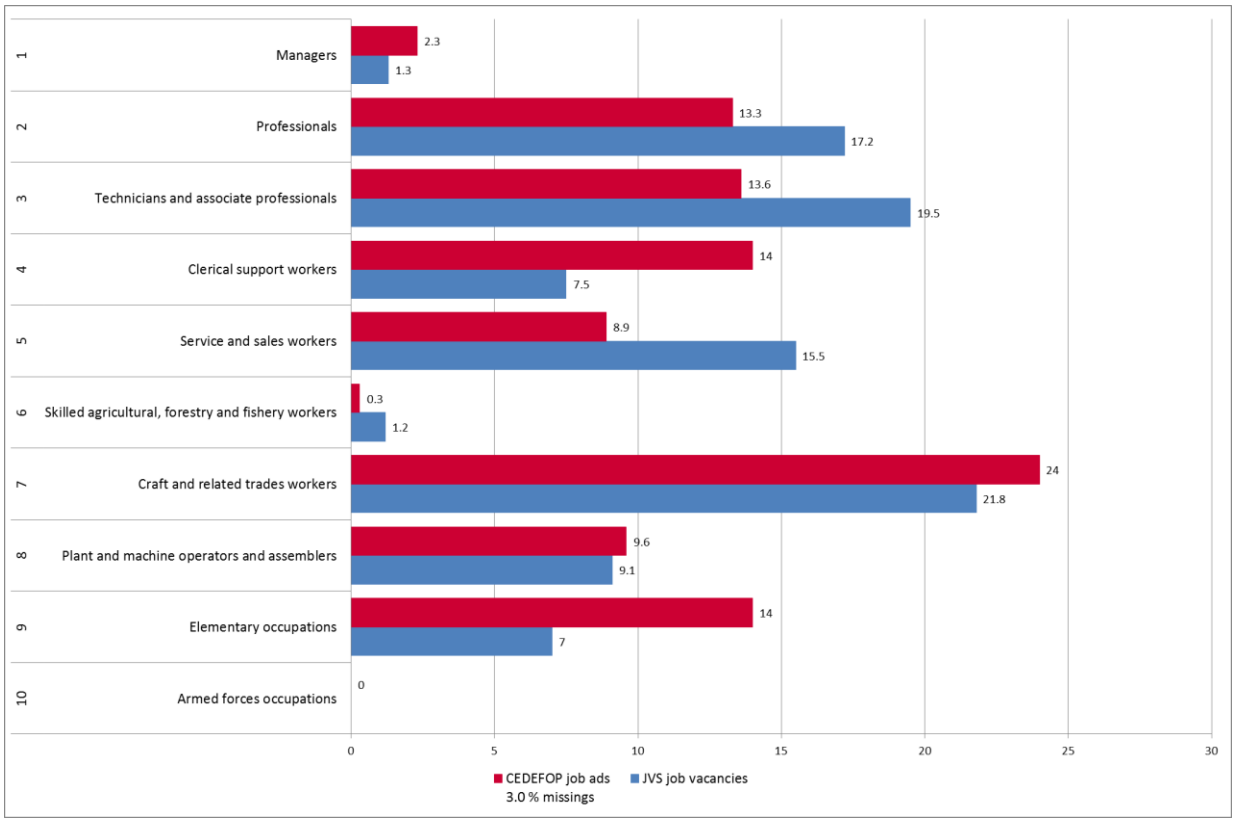
ISCED		CEDEFOP 2015q4 job ads 3 end of months average									JVS job vacancies 2015q4	
		Gigajob		Meine Stadt		Onlinestellen-markt		Insgesamt				
		freq	%	freq	%	freq	%	freq	%	% w/o @	freq	%
0-3	Without professional qualification/unskilled	3	***6.2***	0	***0.0***	1	***2.0***	4	***1.8***	***3.3***	208	***19.8***
4	Post-secondary non-tertiary education	0	***0.9***	0	***0.1***	0	***0.7***	1	***0.4***	***0.7***	617	***59.0***
5	Short-cycle tertiary education	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	58	***5.5***
6	Bachelor or equivalent	2	***4.8***	116	***98.5***	2	***2.8***	120	***51.6***	***96.0***	61	***5.8***
7-8	Master, Doctoral or equivalent	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	103	***9.8***
Missing		37	***88.2***	2	***1.5***	68	***94.5***	107	***46.2***	**–**	–	**–**
Total		42	***100***	118	***100***	72	***100***	232	***100***	***100***	1,047	***100***

Source: Own calculations on CEDEFOP data and special analysis on JVS data by IAB.

7.3.2.1.3 ISCO

The comparison based on the occupation following the ISCO-08 classification is shown in Figure 7 and –with detailed figures– in Table 16 of the annex. At first sight, it does not seem too imperfect. A reason would be the fact that the occupation is mainly based on the job title which is one of the few separate variables in the CEDEFOP data set. After asking them directly they stated that they used a Machine Learning algorithm for doing this classification. Applying a Machine Learning algorithm on such tiny strings seems like a rather easy task. Hence, the few missing values of only 3 % are very understandable. The main differences between the CEDEFOP job ads and the JVS job vacancies occur for the occupations 4, 5 and 9. Additionally, there is quite some variety between the distributions of the job portals, having differences up to 8 %. However, compared to the previous evaluations, CEDEFOP data performs rather well for the classification of occupations.

Figure 7: CEDEFOP-OJV versus JVS: job ads and jobs to be filled immediately by ISCO (2015q4), in %



Source: Own calculations on CEDEFOP data and special analysis on JVS data by IAB.

7.3.2.2 Machine Learning

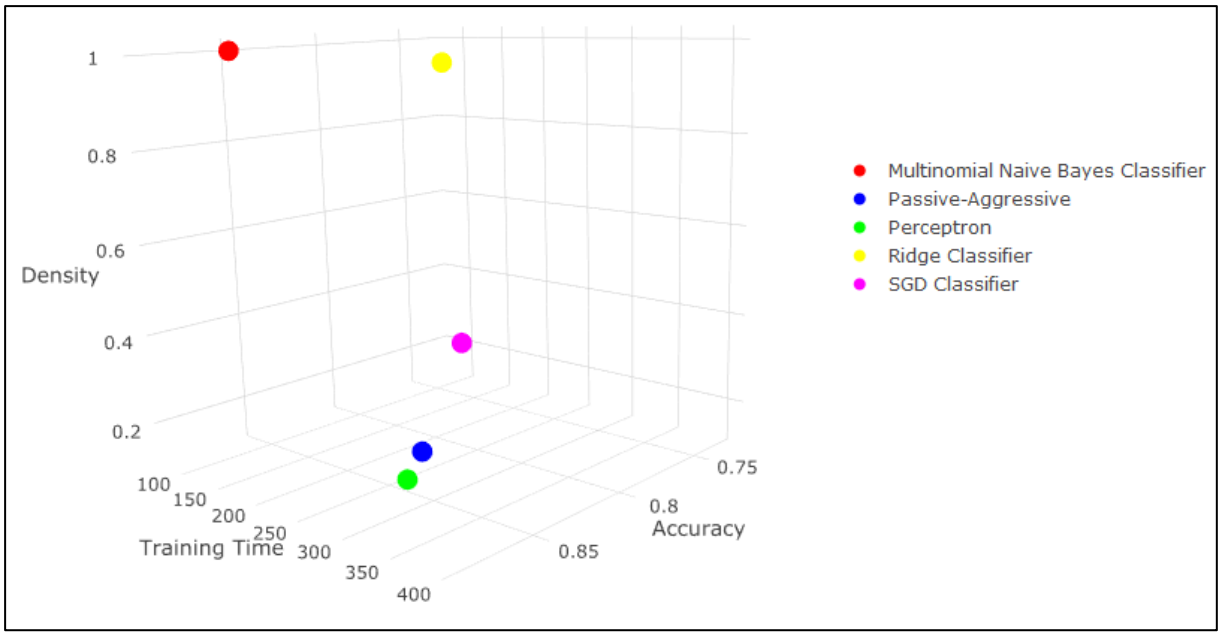
Having such a huge amount of data, it is nearly impossible not to automate some processes via machine learning (ML). CEDEFOP used ML mainly in order to categorize the job ads into their industrial sector or their occupation following ISCO-classification. The advantage of having good ML algorithms is that the scraping process can be performed without too much effort since the algorithms will even classify correctly just based on the full unstructured text. The difficulty is to find and train such an algorithm. So, our task was to examine the algorithms of CEDEFOP.

Keep in mind that, with all its downsides, the JVS is still a good benchmark. And having seen in the previous chapter that the FEA data correlates very well with the JVS data, using the FEA data in order to train algorithms is quite feasible. This holds even more since the FEA data is mostly produced and revised manually.

The task started by evaluating the industry sector per job ad following the NACE-classification. The technical companion was Python equipped with the packages NLTK and scikit-learn. In a first step, it had to be found out which algorithm performs well on classifying a job ad based only by the full unstructured text of a job ad. This was possible since the FEA data also encompassed this variable. A simple bag-of-words-model on the whole data set of the FEA data was used, getting a vector with 577,481 entries. The training set consisted out of 80 % of the original set, while the test set consisted out of 20 %.

Figure 8 shows the performance of the 5 algorithms used for testing purposes considering three dimensions: density[51], accuracy and training time. No algorithm executed in more than 8 minutes which is basically really fast. Nonetheless, keeping in mind that there might a production state with a lot more data in a more frequent way, each second is valuable. This means that the Multinomial Naïve Bayes Classifier is by far the best algorithm in this category. It also has the best density. With a density of exactly 1, it can be assumed that there is no overfitting. The only bad point is its accuracy. On the one hand, with only 85 % nearly every other algorithm, besides the SGD Classifier, performs better. On the other hand, the best algorithm, the Ridge Classifier, only performs 4 percentage points better which is not too much and can possibly also be achieved by the Multinomial Naïve Bayes Classifier by tuning the input parameters or increasing the size of the training set.

Figure 8: Performance of several ML algorithms on the FEA data classifying ads by NACE



Source: Own figure.

In the end, the Multinomial Naïve Bayes Classifier was chosen. A detailed confusion matrix can be found in Figure 9 in the annex of this country report. This classifier was used to test the classifier of CEDEFOP. The finding was that that our prediction of the NACE in the CEDEFOP data differed in 83 % of all cases with the prediction of CEDEFOP's classifier. Interpreting this result, it must be said that the classifier of CEDEFOP can be improved since the FEA data that is used as a training set is the more reliable data set.

Summarizing, although we just started to use ML, there are already some feasible results. That is why the focus should be on improving our skills in this area and do more testing. Other variables can be determined or OJVs can be filtered for private employment agencies by using ML. Also, like already

[51] The density measures the amount of cases per dimension, with dimension defined as the amount of different words in a bag-of-words-approach. Therefore, a low density is a good indicator for overfitting.

mentioned in chapter 7.2.2.1.1, the data of classification servers could be useful since they would provide a good dictionary and therefore a good base for textual machine learning.

Obviously, there are other ways how to categorize job ads. One that is already mentioned in Deliverable 1.2 (SGA-2)[52] is linking the job data to the national BR. Getting access to the BR was easy because it is situated at Destatis. The idea of using the BR was to get more information on the employer of a job ad. This could be the industry sector, the size of the company, the location or many more. In order to achieve this, the data needed to be linked. Except of the employer's name there was no real common variable or ID in both data sets, FEA and CEDEFOP. This made the linkage very difficult and nearly impossible.

Linking FEA's variable for the employer's name to the BR was problematic, because the employer's name in the BR was noisy[53] and had to be cleaned first. Also one company appeared several times in the dataset for only year. However, after cleaning the variable, some good results for test sets of tiny size were achieved by linking the two data sets using the minimal Levenshtein-distance. But using the whole data set was a time consuming task. After two weeks of computing, calculation was stopped. It is only understandable that doing the same thing for the CEDEFOP data is quite impossible, because one would have to extract the name of the employer out of the job ad beforehand. Nonetheless, in future projects some results using newer and better techniques and frameworks might be obtained. There are some promising developments concerning full text searches. Tools like Lucene could be used, but as always and as written in the annex of this country report installing new software is complicated at all times.

In general, the BR is a mixed data set. On the one hand, it is very comprehensive and contains a lot of variables. On the other hand, there are many missing values and the data is not very cleaned. Especially the name of a business is not uniform but maintained in the same way the person filled out its tax explanation. Nevertheless, since it encompasses the number of employees subject to social insurance the data can be used at least to calculate a job vacancy rate. Also, one could compare the structure of the companies that search staffing obtained by web scraping with the structure of companies in the BR in general. This should not differ too much.

7.3.2.3 Strengthening

Even though the CEDEFOP data set is just a prototype and there will be an update by the end of 2018[54] it is important to note how to strengthen the data set concretely. First, the findings of the previous subchapters are summarized, second, some more problems are pointed out and third, some suggestions for improvement are given.

Considering that the JVS data is a survey with a lot of experience, the big gaps between the JVS and the CEDEFOP data in the previous subchapters seems to be due to the bad quality of the CEDEFOP data. Here keep in mind that some of the gaps are explained by the fact that CEDEFOP counts job ads while JVS counts job vacancies. Since in CEDEFOP's data is no variable for the number of job

52 Swier, Nigel et al. (2016), p. 12.

53 This means that names and business types were not consistent due to the origin of the data: The tax office, where people write their company's name by hand.

54 Colombo, Emilio (2018), presentation slide no. 2.

vacancies per ad it had to be assumed that there is one vacancy per ad, as it was the first data received. Now, that the FEA data is obtained, there are actually two ways on how to improve the quality of the CEDEFOP data. Either, in the next data release they try to scrape or construct this variable, or they use a corrective factor calculated via FEA's data. This could work, because the FEA data does contain this variable and that is why one might calculate the number of vacancies per ad per selected feature.

Furthermore, some of CEDEFOP's classifier might be causing some mistakes, as well. A first approach would be to unify the classifiers used by CEDEFOP and by the IAB.

Nevertheless, there are still some more problems with CEDEFOP's data set. Like for example the problem with deduplication that is still not solved for sure. Although CEDEFOP did some deduplication and states that there are no duplicates any more, some duplicates were found manually. This means that there might be even more. Additionally, the CEDEFOP data contains a lot of private employment agencies. By that, the data is biased, because the job ad does not reveal the real employer. In numbers, approximately 75 % of the ads are ads from private employment agencies. One first would have to filter those ads and treat them later on in a special way. Actually, the same machine learning approach like in the previous subchapter could be applied, meaning FEA data could be used to train an algorithm which then is able to categorize unstructured text into ads from private employment agencies or from normal employers. This task can be fortified by using a dictionary containing typical words like "temporary employment" or "recruitment agency".[55]

In general, the big problem of CEDEFOP's data is that it is unstructured and that the few variables that do exist contain too much missing values, e.g. 98.3 % for work experience or no classification of industry sectors for data of the job portal OnlineStellenmarkt. But even the variables that do contain values are not fully valid. As for example the variable for the industrial sector sometimes hold two or more sectors for one job ad. Therefore, it misses some uniqueness. Another example is that sometimes there exist values for the area code but in the full text of the job ad no location is mentioned at all. So the question is where this information comes from. At the moment, evaluating the CEDEFOP data is still very dangerous since the data does not cover the job market at all and is studded with misclassifications. Hence, the current prototype of CEDEFOP is no representative data set for the online job market.

Anyhow, the data can be improved quite easily. A first step would be to scrape more than only 3 portals, like for example StepStone, Monster and last but not least the FEA. Keep in mind that in some cases it could be better to form a cooperation with the portal owner and furthermore that in general it is better to choose only job boards and not hybrid portals.[56] On top of that scraping the OJVs in a more structured way is also not too difficult. The distinction between publication date and scrape date is easy and should not be merged into one variable (grab date) in further approaches. Also, it would be nice to have some of the following variables in order to better analyse the data:

- Employer's name
- Amount of persons searched for in one ad
- URL

[55] Note that there are also ads where a private employment agencies searches for own employers but since this occurs very rarely, this case can be neglected.

[56] See Körner, Thomas, Martina Rengers et al. (2016), p. 6 for more information on hybrid boards.

- Indicator if an ad is owned by the scraped portal or if it is from a partner (for deduplication purposes within hybrid portals)
- Date, when data has to be occupied
- Duration of time limitation

So, if CEDEFOP considered these suggestions and scraped regularly, trying to reach an amount of approximately 1 million job ads per month, the data set would strengthen tremendously.

7.4 Theoretical Statistical Outputs

7.4.1 Current Publication of JVS

Eurostat publishes quarterly and annual data on job vacancies based on the Job Vacancy Survey. For some European countries the number of vacancies and occupied posts is available 45 days after the end of the reference quarter. For all other participating countries it is available around 75 days after the end of the reference quarter (the shorter period applies for countries whose number of employees represents more than 3 % of the European Union total). Quarterly data is broken down by economic activity and enterprise size. Annual data is suited for more structural detailed analysis as they are, in addition - where available, broken down by region and occupation. Data on annual job vacancies are collected on voluntary basis. This is what is said on the website, but de facto there are often missing data for a lot of countries and furthermore much information is not very detailed. For example for results broken down by enterprise size the database allows only a distinction between “total” and “10 employees or more”.

The German JVS is actually very detailed. Following the questionnaire it is possible to publish more on a national level. Even though 16 indicators exist, they are as well not very detailed and some like the required educational level are only given for the fourth quarter of a year. Furthermore, it is not possible to link the variables in order to get a cross table. This would require full access on the survey data which is only given by very restrictive conditions. Another critical point is that some variables do not follow directly an international classification system. The educational level is reported in three categories and unfortunately not in ISCED-11. Some of the industry sectors are merged and the distinction by the company size follows a very uncommon pattern. Nonetheless and also with regard to our comparisons above, the possibility to order special analysis of the JVS calculated by the IAB is possible.

7.4.2 Potential of OJV

In labour market statistics there is a serious lack of empirical information about labour demand as a whole and especially about recent trends of the employers demand for competences, skills and experience with technical equipment. Job advertisements in contrast are a rich source of

information. So OJV data is different to the usual survey data like JVS (or Labour Force Survey) in several points. A lot of them are already marked in previous reports. The main differences are the greater amount of features and their timeliness.

It is already pointed out in chapter 7.2 and 7.3 that OJV data does not really match JVS data. Thus, it is no alternative that may replace the JVS. But OJV data can be used as an additional source for supplementary and new indicators of the labour demand side. Furthermore they can maybe improve the JVS reporting. As OJV data may be collected continuously as a flow of new job ads or as a stock of ads open on a specific date relationships between stocks and flow can be analysed. Overall, the potential OJV data have depends extremely on the extent how the data quality problem of OJV data can be solved.

An idea for an improvement of the JVS reporting could be to use the FEA data as a preliminary result for the JVS data and so improve the timeliness. This whole idea should be possible, because as shown in chapter 7.2.2.1 the differences between these two data sets are not too significant.

The only problem is the coverage, meaning that absolute values of JVS and OJV data do not match. Nevertheless, first, the percentage distribution seems to match and second, it would be possible to adjust the absolute values of OJV data by some grossing-up factors.

Another option would be to focus on the online labour market. While the JVS represents the online and offline labour market, OJV obviously can just describe the online labour market. Of course, the popularity of posting job ads online becomes bigger and bigger, but the offline labour market will never vanish. By implementing an OJV index one would follow the lead of Australia[57] or America[58] who already implemented such an index for quite a long time with a good reputation.

Lastly, the greater amount of features in the OJV data could also be useful. The challenge here is to link this data to the JVS data. A way of solving this problem would be by adding the OJV data to the JVS data and by imputing the missing features for the JVS data with the help of the OJV data. Also, one could just evaluate the new features for the online labour market, like written above and mark them as online-only. By yearly having some special modules in the JVS, the IAB could test for the difference between the online-only and the normal JVS data.

In the end, there is no statistical output for Germany, at the moment, but in order to summarize this chapter, there are several ideas that can be implemented in the future.

[57] Australian Government (2017).

[58] HWOL (2011).

7.5 Conclusions and outlook

The FEA data set benchmarked quite well against the JVS data. It was also very useful as a training set for Machine Learning. However, the data could still be improved concerning the classification and the missing values, like in the case of the educational level. On the contrary, the CEDEFOP data did not perform too well against the JVS data. Looking at both comparisons and both data descriptions, one can observe that the FEA data contains more variables that can be compared to the JVS data than the CEDEFOP data, e.g. size of company. This reveals a main problem: The CEDEFOP data has too few variables and too much unstructured text. The hope is that in further development, the data becomes more structured. Then actually, a cooperation with CEDEFOP really might be beneficial. If an institution starts a project that deals with OJVs and does not have to scrape portals because the data is already provided by CEDEFOP, then progress can be done really fast. Nonetheless, this would just mean that therefore CEDEFOP or other institutions that do scraping would have to deal with the problems of a rapidly changing online job market.

Actually, there are quite a few problems concerning the gathering of OJV data. Some of them were already discussed in earlier reports, some of them are new. So for example the legal issue is still not resolved. The main consent is to inform the scraped portals beforehand and give them the chance to actively object on the scraping procedure. But there is still the question about the legality of the storage of the scraped data.[59] It can be assumed that those legal issues are not going to be solved in the short term. Even more, since the internet and the online job market in particular is changing in various ways. First, new players emerge on a constant base in the same way as old players cease. Hence, new cooperations or new scraping frameworks have to be built continuously. Second, the latter is also the case if the scraped job portal is just changing its layout or its structure of its website. But in order to control for this Destatis is currently aiming on implementing a generic approach that tries to adapt a scraping program easily to new structures.[60]

Furthermore, a few years ago, another type of job ads – video ads – emerged and gained more and more fame. Examples for job portals that include video ads are JOBBÖRSE.de or STELLENANZEIGEN.de. Portals like StepStone allow embedding videos in their job ad. In general, video ads can be distinguished by videos that are tailored for a certain vacancy, videos that just present the company or a certain type of employment (e.g. trainee) or videos that only support a text-based OJV. In either case, the video might contain information that is not written anywhere else and therefore scraping this kind of information would also be necessary or at least beneficial.[61] All in all, video ads are no issue that really have to be taken into account right now. Nevertheless, one should keep in mind that video ads do exist and observe the upcoming changes on the online job market.

In order to summarize the findings, it is left to say that, as shown, every online data set is fraught with problems when comparing it with the whole job market. This means that OJV data will never replace the JVS. Still it might enrich it, either by adding features, by giving preliminary results or

---

59 Stateva, Galia et al. (2017), p. 2f.

60 The approach will primarily be implemented for the price statistics, cf. Blaudow, Christian (2018), p.5.

61 Technically spoken it is not too hard to extract a video file. What is more difficult is the information extraction of the video. With the help of Optical Character Recognition (OCR) it is possible to retrieve the written text in a video. Obviously this text is unstructured and must be handled similar to the way the CEDEFOP data was handled. Additionally, an idea would be to convert the spoken words into text via Speech Recognition and analyse this unstructured text in a similar way.

maybe by having a special online index. As stated in the previous chapter new indicators that cannot be built by the JVS could be published. So, albeit all problems and obstacles, this report closes with a plea that OJV data should be used and also be published.

7.6 References

Australian Government (2017), Vacancy Report. September 2017, Department of Employment, ISSN 1446-9448.

FEA (2018a), Download-Portal, Bundesagentur für Arbeit, https://download-portal.arbeitsagentur.de/files/, retrieved on 18th April 2018.

FEA (2018b), Tabellarische Umsteigeschlüssel zur KldB 2010, Bundesagentur für Arbeit, https://statistik.arbeitsagentur.de/nn_237808/Statischer-Content/Grundlagen/Klassifikation_-der-Berufe/KldB2010/Arbeitshilfen/Umsteigeschluessel/Umsteigeschluessel.html, retrieved on 18th April 2018.

Blaudow, Christian (2018), Fortschritte und Herausforderungen beim Web Scraping – Automatisierung von Preiserhebungen im Internet, in: Statistisches Bundesamt (Ed.): METHODEN – VERFAHREN – ENTWICKLUNGEN. Nachrichten aus dem Statistischen Bundesamt 1/2018, Wiesbaden.

Carnevale, Anthony P., Tamara Jayasundera and Dmitri Repnikov (2014), Understanding online job ads data. A Technical Report, Georgetown University.

CEDEFOP (2018), Real-time labour market information and skill requirements: Setting up the infrastructure for EU system. Expert workshop, Agenda, Milan 20th-21st March, http://www.cedefop.europa.eu/files/agenda_ws_cedefop_crisp_ess_net_v31.pdf, retrieved on 25th April 2018

CEDEFOP (2016), Final Report, Project "Real-time labour market information on skill requirements: feasibility study and working prototype", Annex 2.

Colombo, Emilio (2018), Setting up the infrastructure for EU system. Current state of the art and next steps, presentation at Expert Workshop in Milan, CEDEFOP, CRISP, Tabulaex, 20th-21th March 2018 (unpublished).

HWOL (2011), The Conference Board Help Wanted OnLine™ Data Series. Technical Notes, https://www.conference-board.org/pdf_free/HWOLJan11_TN.pdf, retrieved on 16th April 2018.

Körner, Thomas, Martina Rengers et al. (2016), Deliverable 1.1. Inventory and qualitative assessment of job portals (SGA-1), ESSnet Big Data, Work Package 1: Web scraping / Job vacancies.

Rengers, Martina (2018), Internetbasierte Erfassung offener Stellen im Statistischen Bundesamt, in: König, Christian et al. (Eds.), Big Data – Chancen, Risiken, Entwicklungstendenzen, Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute, Wiesbaden, p. 61-86.

Stateva, Galia et al. (2017), Legal aspects related to Web scraping of Enterprise Web Sites, ESSnet Big Data, Work Package 2: Web scraping / Enterprise Characteristics.

Swier, Nigel (2018), Deliverable 1.1, Strategy for ongoing engagement (SGA-2), ESSnet Big Data, Work Package 1: Web scraping / Job vacancies.

Swier, Nigel et al. (2017), Deliverable 1.3. Final Technical Report (SGA-1), ESSnet Big Data, Work Package 1: Web scraping / Job vacancies.

Swier, Nigel et al. (2016), Deliverable 1.2. Interim Technical Report (SGA-1), ESSnet Big Data, Work Package 1: Web scraping / Job vacancies.

7.7 Annex

7.7.1 Description of OJV data from FEA

Table 6: 75 variables of FEA-OJVs data set „stea_20180226.csv“

No	Variable	Description	Features
1	Stellenid	Identifier for the job ad	19 digits, no missing
2	Arbeitgeber	Name of employer	Text
3	Straße	Street and house number of employer	Alphanumerical text
4	Postleitzahl	Post code of employer	5 digits
5	Ort	Location of employer	Text
6	Angebotsart	Type of workers who may apply	“only specialized personnel”, “only managers”, “only aid workers”, “workers”
7	AnzStellen	Number of vacancies per ad	Integer > 0
8	Veroeffstatus	Type of publicizing	„published anonymously“, „published“, „internal and published anonymously on job portal“
9	EintrittVon	Earliest date of staffing	Date
10	EintrittBis	Latest date of staffing	Date
11	Allianzpartner	Name of allied partner, meaning origin of job ad (either enterprise or portal)	Text
12	SozialversPflichtig	Flag for subject to social insurance contributions	0 or 1
13	Hauptberuf	Occupation coded by DKZ (Dokumentationskennziffer)	3 to 6 digits, no missing
14	Alternativberuf1	First alternative occupation coded by DKZ	3 to 6 digits, 181,254 missings
15	Alternativberuf2	Second alternative occupation coded by DKZ	3 to 6 digits, 309,743 missings
16	Hochschulabschlussart	Type of degree needed	13 different levels of degree, 438,157 missings
17	HochschulabschlussAlt1	First alternative type of degree needed	13 different levels of degree, 450,003 missings
18	HochschulabschlussAlt2	Second alternative type of degree needed	13 different levels of degree, 459,128 missings
19	SteaTitel	Title of job ad	Text
20	Fuehrungsverantwortung	Managerial responsibility by level	6 different levels of responsibility
21	Stellenbeschreibung	Job description as full text	Text
22	Befristung	Time limitation of job	limited, unlimited
23	BefristetBis	Date of time limitation	Date
24	BefristungMonate	Duration of time limitation in months	Integer
25	BietetUebernahme	Flag if transfer is offered (for temporary employment)	J (=Yes), N (=no) or empty
26	Arbeitszeit	Working period	Full time, part time (divided into more detailed periods), weekends, home office
27	Wochenstunden	Working hours per week	Decimal number
28	Arbeitszeitinfo	Additional information on working period, given by the employer	Text
29	Gehaltswunsch	Salary requirement	Alphanumerical text
30	Tarifvertrag	Name of collective labour agreement if applicable	Text

31	IstMinijob	Flag for minijob	0 or 1
32	BietetFirmenwagen	Company car is offered	Empty
33	BietetProvision	Commissions are offered	Empty
34	BietetBonuszahlungen	Bonus Payments are offered	Empty
35	BewerbungTelefonisch	Application via phone is possible	J (=Yes) or Empty
36	BewerbungSchriftlich	Written application is possible	J (=Yes) or Empty
37	BewerbungPerEmail	Application via email is possible	J (=Yes) or Empty
38	BewerbungPersoenlich	Personal application is possible	J (=Yes) or Empty
39	BewerbungFirmenFormular	Application via a company form is possible	J (=Yes) or Empty
40	BewerbungOnline	Online application is possible	J (=Yes) or Empty
41	BewerbungUeberBA	Application via FEA is possible	J (=Yes) or Empty
42	UrlBewerbungsFormular	Link to application form	URL
43	InternetAdresse	Internet address of company	URL
44	BewerbungZeitraumVon	Earliest possible date for application	Date
45	BewerbungZeitraumBis	Latest possible date for application	Date
46	GeforderteAnlagen	Required attachments	Text (mostly CV and certificates)
47	GeforderterAbschluss	Required degree	“not relevant” or Empty (1 case with “university” and 2 cases with “main school”)
48	KonkretisierungAbschluss	Concretisation of degree	Empty (2 cases “main school”)
49	GeforderteSchulart	Required type of school	Empty
50	GeforderteFachrichtung	Required speciality	Empty
51	Reisebereitschaft	Travel readiness	Permanently, timewise or never needed
52	GefordertKraftrad	Flag if motorcycle is needed	0 or 1
53	GefordertPKW	Flag if car is needed	0 or 1
54	GefordertLKW	Flag if truck is needed	0 or 1
55	GefordertOmnibus	Flag if bus is needed	0 or 1
56	Berufserfahrung	Work experience	Empty
57	AlterVon	Minimum Age	Empty
58	AlterBis	Maximal Age	Empty
59	Geschlecht	Required sex	Empty
60	Wehrdienstpflicht	compulsory military service needed	Empty
61	NurFuerBehinderte	Flag if job is only for disabled	0 or 1
62	Hauptausbildung	Main vocational training	3 to 6 digits, 356,940 missings
63	HochschulabschlussAusB	Educational level by category	11 different levels of education, 458,087 missings
64	AngabenZurAusbildung	Details to education	31,122 different features, 426,425 missings
65	GeforderteLeitungsart	Type of management required	Business or technical management
66	GeforderteVollmacht	Letter of attorney required	3 cases “general power, 144 cases “power of attorney”, 3 cases “procuration”, 470,224 missings
67	JahreFuehrungserfahrung	Years of management experience	968 cases “2 to 5 years”, 899 cases “up to 2 years”, 175 cases “more than years”, 468,332 missings
68	Budgetverantwortung	Responsible for budget	4 different features stating amount of money, 470,270 missings

69	Personalverantwortung	Personnel responsibility	4 different features, stating amount of coworkers, 468,187 missings
70	Betriebsgroesse	Size of company	(1) not specified (2) less than 6 (3) between 6 and 50 (4) between 51 and 500 (5) between 501 and 5,000 (6) between 5,001 and 50,.000 (7) more than 50,000 7,467 missings
71	Branche1	Industrial sector by NACE/WZ-2008	21 NACE-sectors A-U
72	Branche2	First detailed text for industrial sector	Text
73	Branche3	Second detailed text for industrial sector	Text
74	Branche4	Third detailed text for industrial sector	Text
75	Branche5	Fourth detailed text for industrial sector	Text

Source: Own presentation.

Table 7: 9 variables of FEA-OJVs data in supplementary data set „Stea_Lokation_20180226.csv“

No	Variable	Description	Features
1	LaufendeNummer	Increasing count for number of locations per ad	1 to 10
2	Straße	Street of place of employment	Name of street + house number
3	Adresszusatz	Addition to address, e.g. „z.Hd. v.“ (= for attention of: FAO)	Text
4	PLZ	Post code of place of employment	5 digits
5	Ort	Name of location of place of employment, where necessary inclusive federal state, name of river or other additional information	Text
6	Ortsteil	Name of district of place of employment	Text
7	Land	Name of country of place of employment (mainly Germany but also worldwide)	Text
8	Region	In Germany Federal State, for the rest cantons or regions	Text
9	OrtsID	ID of location	10 or 11 digits

Source: Own presentation.

Table 8: Size of data files of the FEA data set

No	File	Rows	Columns	Comment	Size
1	Stea_20180226.csv	470,374	75	Main file	935 MB
2	Stea_Lokation_20180226.csv	532,580	10		45.8 MB

Source: Own presentation.

7.7.2 Description of OJV data from CEDEFOP

Table 9: 17 variables of CEDEFOP-OJVs data set „st_document_de.csv“

No	Variable	Origin	Description	Features
1	GeneralId	Internal, increasing	Identifier of the job ad	Integer
2	LocalId	Internal, increasing	Identifier	2 to 7 digits plus “DE”
3	IdDocument	Internal, increasing, grouped by country	Identifier of the scraped document	Integer
4	Description	Scraped Web Page	Full text of the job ad	String
5	DescriptionCrypt	Calculated from Description	Encrypted description for deduplication purposes	Alphanumerical String
6	GrabDate	Scraped Web Page	Publication date if provided. Otherwise date of scraping (if same ad is scraped twice, the first date is kept)	Minimal Date: 2008, Maximal Date: 2015
7	YearGrabDate	From GrabDate	Year of variable GrabDate	cf. GrabDate
8	MonthGrabDate	From GrabDate	Month of variable GrabDate	cf. GrabDate
9	DayGrabDat	From GrabDate	Day of variable GrabDate	cf. GrabDate
10	ExpireDate	Scraped Web Page	Date when job ad expires if provided. Otherwise first date, when job ad was not available any more	2015 or 2016
11	YearExpireDate	From ExpireDate	Year of variable ExpireDate	cf. ExpireDate
12	MonthExpireDate	From ExpireDate	Month of variable ExpireDate	cf. ExpireDate
13	DayExpireDate	From ExpireDate	Day of variable ExpireDate	cf. ExpireDate
14	Title	Scraped Web Page	Title of the job ad	String
15	IdProfession	Scraped Web Page	Profession following ISCO/ESCO 4th digit codes	4 digits
16	Experience	Scraped Web Page	Information how much experience is needed for the job	0.5-1; 1-2; 2-5; 6 months experience in this job; no experience needed; more than 5 years of experience in this job; at least 1 year experience in this job; at least 2 of years of experience in this job; at least 3 years of experience in this job; at least 4 years of experience in this job; at least 5 years of experience in this job
17	Salary	Scraped Web Page	Information if salary is fix or variable	Fix salary; Variable salary

Source: Own presentation.

Table 10: 14 variables of CEDEFOP-OJVs data in supplementary data sets

No	Variable	Origin	Description	Features
1	Source	st_source_de.csv	Job Portal	Gigajob, MeineStadt, OnlineStellenmarkt
2	Nuts-0	st_area_de.csv	Countries	2 digits
3	Nuts-1	st_area_de.csv	Major socio-economic regions	3 digits
4	Nuts-2	st_area_de.csv	Basic regions	4 digits
5	Nuts-3	st_area_de.csv	Small regions	5 digits
6	GeneralNut	st_area_de.csv	-	empty
7	Contract	st_contract_de.csv	Type of contract	Permanent; Temporary; Self Employed
8	Educational Level	st_educational_level_de.csv	ISCED-Classification	1. Primary education 2. Lower secondary education 3. Upper secondary education 4. Post-secondary non-tertiary education 5. Short-cycle tertiary education 6. Bachelor or equivalent 7. Master or equivalent 8. Doctoral or equivalent
9	Industry_Level 1	st_industry_de.csv	NACE-Classification	1 letter
10	Industry_Level 2	st_industry_de.csv	NACE-Classification	2 digits
11	IdEsco_Profession	st_skill_profession.csv	Profession following ESCO-Classification	4 digits
12	IdEsco_Skill	st_skill_profession.csv	Skill following ESCO-Classification	5 or 6 digits or „new skill“
13	NGram	st_skill_profession.csv	Skill written-out	String
14	WorkingHours	st_working_hours_de.csv	Full or Part Time	Part Time; Full Time

Source: Own presentation.

Table 11: Size of data files of the CEDEFOP-OJVs data

No	File	Rows	Columns	Comment	Size
1	st_area_de.csv	2,028,370	6	Sometimes more than one entry corresponds to one entry in the main file	77.2 MB
2	st_contract_de.csv	938,074	2		13.5 MB
3	st_document_de.csv	2,140,547	17	Main file	3.71 GB
4	st_educational_level_de.csv	1,074,694	2		15.5 MB
5	st_industry_de.csv	1,658,762	3	Sometimes more than one entry corresponds to one entry in the main file	30.7 MB
6	st_skill_profession.csv	12,450,036	4	More than one entry corresponds to one entry in the main file	488 MB
7	st_source_de.csv	2,461,393	2	Sometimes more than one entry corresponds to one entry in the main file	70.3 MB
8	st_working_hours_de.csv	1,388,913	2		20.1 MB

Source: Own presentation.

7.7.3 Data Handling

The following section deals with the internal problems we had to face at Destatis in order to handle and analyse the data. It can be seen as an experience report and an advice on how to prepare an institution for handling big data tasks in the area of OJVs.

Table 12: Problems in data handling

Topic	Problem	Solution
Loading Data	There are very big files that need much memory and much random access memory (RAM). Due to this loading the data is sometimes difficult or needs too much time or even some errors occur. Depending on the development environment there are different methods and tools to load data but sometimes these packages cannot be installed for institutional security reasons. Loading data with SAS is possible, but the SAS servers are very limited (100 GB per department).	Increase memory and RAM of local machines as well as the capacity of the SAS servers. Also allow freely to install packages e.g. in R and Python.
Web Scraping	At the moment there is now good IT environment for web scraping. Since it is not even allowed to update a browser it is quite difficult to even see the content of a page in the right way. Also one is not able to test several ways of scraping because one cannot just install software by himself. First, IT security must be asked.	Loosen some restrictions on installing third party software and/or enforcing development of a generic approach for web scraping that makes updating the program if a site has changed very easy.[62]
Text Mining	There is no special tool for text mining that is actually allowed at our institute. Since SAs is already used, SAS Text Mining could be used, but this is not available due to its high costs. Also one cannot use Python because the connection to pip – the package manager of Python – is blocked by the internal firewall.	Buy SAS Text Mining or allow connection to pip.
Installing new software	Several tools are essential for Big Data tasks. Programs like Python, R, Java, MySQL, Anaconda or Slack should be the basic equipment of a developer. Despite that, they are not preinstalled and are difficult to get. Asking the internal IT department is a time consuming task. It takes them a long time to do their services and to answer questions. Also they want to know, why a task cannot be handled by the standard equipment (like Excel or SAS). Since developing is an agile process with a lot of trials and errors, one sometimes cannot argue why a certain program is the one and only answer, because he has not even tested the software. There is a solution to just use a free notebook that is not connected to the normal IT infrastructure. But even such a notebook is not fully free since some installation files are blocked by the antivirus scanner which cannot be configured at all.	Allow more software to be installed. Answer requests by users more quickly. Have more software preinstalled or maybe even create a repository where a user can install software by his own. Allow configuration of antivirus scanner on free notebooks.
Data Protection	Like written in the previous point, sometimes one needs to develop on a free notebook in order to be able to install new software by his own. The trade-off is that it is not possible to put sensible data on a notebook like that.	Protect sensible data by anonymizing.

[62] This is already mentioned in the conclusions chapter.

Lacking IT skills	Usually at Destatis while working on a project, there are specialty departments and functional departments. The latter helps the former with certain knowledge, mostly IT knowledge. Concerning Big Data there is hardly any knowledge at our institute.	Strengthen skills for Big Data tasks.

Source: Own presentation.

By now, all the data collected is no Big Data in general, since it is not very big at all. However, several difficulties that are named above are already encountered. If the process goes on and data really becomes bigger and bigger in a more frequent way, a huge problem might appear. The only solution seems to be a fundamental change concerning the IT environment.

7.7.4 Others

Table 13: FEA-OJV versus JVS: job ads and job vacancies by industry sector NACE (30 days and 2017q1), in thousands

NACE	FEA 30 days										JVS 2017q1 job vacancies	
	FEA				Other allied partners				Total			
	job ads		job vacancies		job ads		job vacancies		job vacancies			
	freq	%	freq	%	freq	%	freq	%	freq	%	freq	%
A: Agriculture, forestry and fishing	3	*0.7*	4	*0.6*	0	*0.0*	0	*0.1*	4	*0.5*	8	*0.7*
B: Mining and quarrying	0	*0.1*	0	*0.1*	.	*.*	.	*.*	0	*0.0*	1	*0.1*
C: Manufacturing	52	*13.1*	70	*11.4*	1	*2.5*	1	*1.5*	71	*10.3*	118	*11.1*
D: Electricity, gas, steam and air conditioning supply	1	*0.2*	1	*0.1*	0	*0.0*	0	*0.0*	1	*0.1*	2	*0.2*
E: Water supply; sewerage, waste management and remediation activities	2	*0.5*	3	*0.4*	.	*.*	.	*.*	3	*0.4*	4	*0.4*
F: Construction	35	*9.0*	47	*7.7*	0	*0.0*	0	*0.0*	47	*6.8*	88	*8.2*
G: Wholesale and retail trade; repair of motor vehicles and motorcycles	48	*12.3*	68	*11.1*	3	*5.9*	3	*3.5*	71	*10.2*	97	*9.2*
H: Transportation and storage	16	*4.2*	33	*5.3*	1	*2.6*	1	*1.7*	34	*4.9*	83	*7.8*
I: Accommodation and food service activities	27	*6.8*	34	*5.5*	1	*2.8*	1	*1.7*	36	*5.1*	74	*7.0*
J: Information and communication	10	*2.6*	14	*2.3*	0	*0.1*	0	*0.1*	14	*2.0*	50	*4.7*
K: Financial and insurance activities	3	*0.7*	4	*0.7*	0	*0.5*	0	*0.3*	4	*0.6*	11	*1.0*
L: Real estate activities	2	*0.5*	2	*0.4*	0	*0.1*	0	*0.1*	2	*0.3*	2	*0.2*
M: Professional, scientific and technical activities	28	*7.2*	38	*6.1*	3	*7.6*	4	*5.2*	42	*6.0*	81	*7.6*
N: Administrative and support service activities	94	*23.8*	204	*33.0*	34	*77.6*	66	*85.5*	269	*38.8*	225	*21.1*
O: Öffentliche Verwaltung, Verteidigung; Sozialversicherung	5	*1.3*	7	*1.1*	0	*0.0*	0	*0.0*	7	*1.0*	20	*1.8*
P: Public administration and defence; compulsory social security	6	*1.5*	8	*1.2*	.	*.*	.	*.*	8	*1.1*	33	*3.1*
Q: Human health and social work activities	48	*12.2*	64	*10.3*	0	*0.1*	0	*0.1*	64	*9.2*	142	*13.4*
R: Arts, entertainment and recreation	3	*0.7*	4	*0.6*	.	*.*	.	*.*	4	*0.5*	9	*0.8*
S: Other service activities	10	*2.5*	13	*2.1*	0	*0.0*	0	*0.0*	13	*1.8*	18	*1.6*
T: Activities of Households as Employers; Undifferentiated Goods and Services Producing Activities of Households for Own Use	0	*0.1*	1	*0.1*	.	*.*	.	*.*	1	*0.1*	0	*0.0*
U: Activities of Extraterritorial Organisations and Bodies	0	*0.0*	0	*0.0*	.	*.*	.	*.*	0	*0.0*	0	*0.0*
Missing	0	*0.0*	0	*0.0*	.	*.*	.	*.*	0	*0.0*	0	*0.0*
Total	393	*100*	618	*100*	44	*100*	77	*100*	695	*100*	1,064	*100*

Source: Own calculations on FEA data and special analysis on JVS data by IAB.

Table 14: FEA-OJV versus JVS: job ads, job vacancies and jobs to be filled immediately by ISCO (30 days and 2017q1)

ISCO-08 First digit	FEA 30 days												JVS jobs to be filled immediately 2017q1		
	FEA				Other allied partners				Total						
	job ads		job vacancies		job ads		job vacancies		job ads		job vacancies				
	freq	%	freq	%	freq	%	freq	%	freq	%	freq	%	freq in K	%	% w/o @
1 Managers	3,706	0.9	4,171	0.7	220	0.5	248	0.3	3,926	0.9	4,419	0.6	8	1.1	1.1
2 Professionals	40,956	10.4	49,810	8.1	4,330	9.9	4,642	6.1	45,286	10.4	54,452	7.8	112	14.8	15.5
3 Technicians and associate professionals	72,129	18.4	97,475	15.8	5,558	12.7	8,069	10.5	77,687	17.8	105,544	15.2	138	18.2	19.1
4 Clerical support workers	34,790	8.9	57,190	9.3	5,704	13	10,066	13.1	40,494	9.3	67,256	9.7	61	8.0	8.4
5 Service and sales workers	65,598	16.7	103,725	16.8	4,223	9.6	5,974	7.8	69,821	16	109,699	15.8	87	11.5	12.0
6 Skilled agricultural, forestry and fishery workers	4,556	1.2	5,641	0.9	140	0.3	190	0.2	4,696	1.1	5,831	0.8	10	1.3	1.4
7 Craft and related trades workers	100,278	25.5	156407	25.3	13,207	30.1	23,662	30.9	113,485	26	180,069	25.9	154	20.3	21.3
8 Plant and machine operators and assemblers	34,765	8.9	67,245	10.9	5,295	12.1	11,132	14.5	40,060	9.2	78,377	11.3	77	10.1	10.7
9 Elementary occupations	35,808	9.1	76,536	12.4	5,236	11.9	12,635	16.5	41,044	9.4	89,171	12.8	75	9.9	10.4
10 Armed forces occupations	4	0.0	4	0.0	0	0.0	0	0	4	0	4	0	0	0	0
Missing	3	0.0	3	0.0	0	0.0	0	0	3	0	3	0	37	4.9	–
Total	392,593	100	618,207	100	43,913	100	76,618	100	436,506	100	694,825	100	759	100	100

Source: Own calculations on FEA data and special analysis on JVS data by IAB.

Table 15: CEDEFOP-OJV versus JVS: number of job ads and job vacancies by industry sector NACE (2015q4), in thousands

NACE	CEDEFOP 2015q4 Stichtag job ads									JVS job vacancies 2015q4	
	Gigajob		Meine Stadt		Onlinestellen-markt		Total				
	freq	%	freq	%	freq	%	freq	%	% w/o @	freq	%
A: Agriculture, forestry and fishing	519	***1.2***	0	***0.0***	0	***0.0***	519	***0.2***	***0.3***	11	***1.0***
B: Mining and quarrying	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	0	***0.0***
C: Manufacturing	4,134	***9.8***	0	***0.0***	0	***0.0***	4,134	***1.8***	***2.7***	108	***10.3***
D: Electricity, gas, steam and air conditioning supply	3,745	***8.9***	0	***0.0***	0	***0.0***	3,745	***1.6***	***2.4***	2	***0.2***
E: Water supply; sewerage, waste management and remediation activities	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	3	***0.3***
F: Construction	1,539	***3.7***	48,301	***41.1***	0	***0.0***	49,841	***21.5***	***32.6***	73	***7.0***
G: Wholesale and retail trade; repair of motor vehicles and motorcycles	1,858	***4.4***	10,288	***8.8***	0	***0.0***	12,147	***5.2***	***7.9***	118	***11.3***
H: Transportation and storage	17,928	***42.5***	6,943	***5.9***	0	***0.0***	24,871	***10.7***	***16.3***	68	***6.5***
I: Accommodation and food service activities	1,935	***4.6***	7,759	***6.6***	0	***0.0***	9,695	***4.2***	***6.3***	74	***7.1***
J: Information and communication	99	***0.2***	5,613	***4.8***	0	***0.0***	5,713	***2.5***	***3.7***	42	***4.0***
K: Financial and insurance activities	1	***0.0***	1,187	***1.0***	0	***0.0***	1,188	***0.5***	***0.8***	13	***1.2***
L: Real estate activities	60	***0.1***	0	***0.0***	0	***0.0***	60	***0.0***	***0.0***	6	***0.5***
M: Professional, scientific and technical activities	5,071	***12.0***	10,099	***8.6***	0	***0.0***	15,171	***6.5***	***9.9***	69	***6.6***
N: Administrative and support service activities	442	***1.0***	18,097	***15.4***	0	***0.0***	18,539	***8.0***	***12.1***	244	***23.4***
O: Öffentliche Verwaltung, Verteidigung; Sozialversicherung	11	***0.0***	0	***0.0***	0	***0.0***	11	***0.0***	***0.0***	19	***1.8***
P: Public administration and defence;compulsory social security	2	***0.0***	0	***0.0***	0	***0.0***	2	***0.0***	***0.0***	32	***3.0***
Q: Human health and social work activities	120	***0.3***	6,923	***5.9***	0	***0.0***	7,043	***3.0***	***4.6***	133	***12.7***
R: Arts, entertainment and recreation	288	***0.7***	0	***0.0***	0	***0.0***	288	***0.1***	***0.2***	9	***0.9***
S: Other service activities	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	22	***2.1***
T: Activities of Households as Employers; Undifferentiated Goods and Services Producing Activities of Households for Own Use	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	–	–
U: Activities of Extraterritorial Organisations and Bodies	0	***0.0***	0	***0.0***	0	***0.0***	0	***0.0***	***0.0***	–	–
Missing	4,394	***10.4***	2,332	***2.0***	72,338	***100***	79,065	***34.1***	–	–	–
Total	42,149	***100***	117,544	***100***	72,338	***100***	232,032	***100***	***100***	1,047	***100***

Source: CEDEFOP data and JVS data from EUROSTAT database.

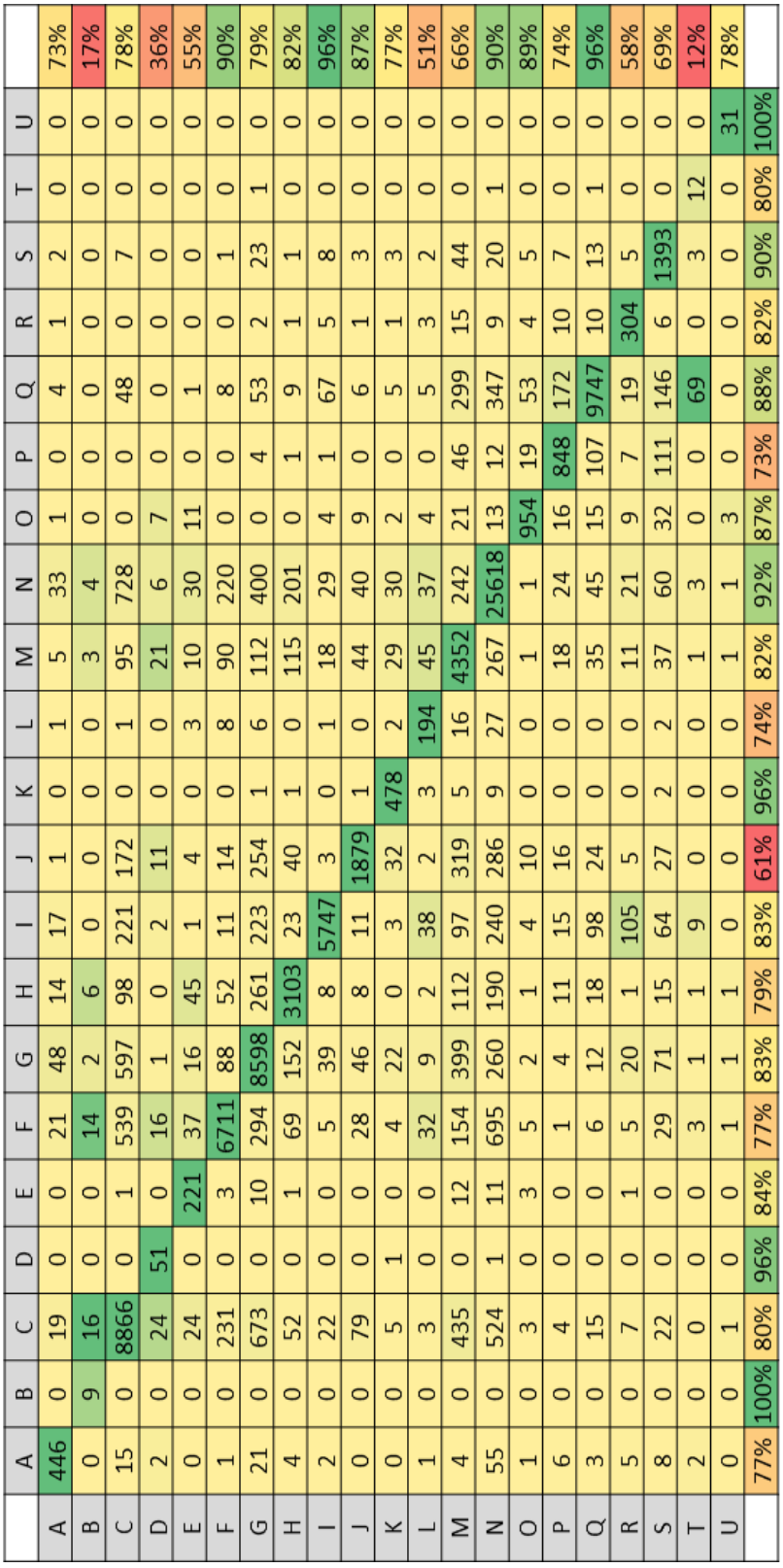
Table 16: CEDEFOP-OJV versus JVS: job ads and jobs to be filled immediately by ISCO (2015q4)

ISCO-08 First digit	CEDEFOP 2015q4 Stichtag									JVS jobs to be filled immediately 2015q4		
	Gigajob		Meine Stadt		Onlinestellen-markt		Total					
	freq	%	freq	%	freq	%	freq	%	% w/o @	freq in K	%	% w/o @
1 Managers	1,108	**2.6**	2,522	**2.1**	1,438	**2.0**	5,068	**2.2**	**2.3**	9	**1.1**	**1.3**
2 Professionals	2,975	**7.1**	15,856	**13.5**	11,190	**15.5**	30,020	**12.9**	**13.3**	124	**15.1**	**17.2**
3 Technicians and associate professionals	2,904	**6.9**	17,236	**14.7**	10,549	**14.6**	30,689	**13.2**	**13.6**	141	**17.1**	**19.5**
4 Clerical support workers	8,856	**21.0**	13,779	**11.7**	8,904	**12.3**	31,539	**13.6**	**14.0**	54	**6.6**	**7.5**
5 Service and sales workers	2,189	**5.2**	11,860	**10.1**	5,953	**8.2**	20,001	**8.6**	**8.9**	112	**13.6**	**15.5**
6 Skilled agricultural, forestry and fishery workers	147	**0.3**	385	**0.3**	194	**0.3**	725	**0.3**	**0.3**	9	**1.0**	**1.2**
7 Craft and related trades workers	8,150	**19.3**	28,625	**24.4**	17,214	**23.8**	53,990	**23.3**	**24.0**	158	**19.2**	**21.8**
8 Plant and machine operators and assemblers	7,523	**17.8**	9,350	**8.0**	4,711	**6.5**	21,584	**9.3**	**9.6**	66	**8.0**	**9.1**
9 Elementary occupations	7,357	**17.5**	14,775	**12.6**	9,348	**12.9**	31,479	**13.6**	**14.0**	51	**6.1**	**7.0**
10 Armed forces occupations	0	**0.0**	0	**0.0**	0	**0.0**	0	**0.0**	**0.0**	–	–	–
Missing	941	**2.2**	3,157	**2.7**	2,839	**3.9**	6,937	**3.0**	–	99	**12.1**	–
Total	42,149	**100**	117,544	**100**	72,339	**100**	232,032	**100**	**100**	822	**100**	**100**

* Note: The major group 10 (Armed forces occupations) is not acquired within the scope of the JVS

Source: Own calculations on CEDEFOP data and special analysis on JVS data by IAB

Figure 9: Confusion Matrix of ML on FEA data classifying NACE



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
A	446	0	19	0	0	21	48	14	17	1	0	1	5	33	1	0	4	1	2	0	0	73%
B	0	9	16	0	0	14	2	6	0	0	0	0	3	4	0	0	0	0	0	0	0	17%
C	15	0	8866	0	1	539	597	98	221	172	0	1	95	728	0	0	48	0	7	0	0	78%
D	2	0	24	51	0	16	1	0	2	11	0	0	21	6	7	0	0	0	0	0	0	36%
E	0	0	24	0	221	37	16	45	1	4	0	3	10	30	11	0	1	0	0	0	0	55%
F	1	0	231	0	3	6711	88	52	11	14	0	8	90	220	0	0	8	0	1	0	0	90%
G	21	0	673	0	10	294	8598	261	223	254	1	6	112	400	0	4	53	2	23	1	0	79%
H	4	0	52	0	1	69	152	3103	23	40	1	0	115	201	0	1	9	1	1	0	0	82%
I	2	0	22	0	0	5	39	8	5747	3	0	1	18	29	4	1	67	5	8	0	0	96%
J	0	0	79	0	0	28	46	8	11	1879	1	0	44	40	9	0	6	1	3	0	0	87%
K	0	0	5	1	0	4	22	0	3	32	478	2	29	30	2	0	5	1	3	0	0	77%
L	1	0	3	0	0	32	9	2	38	2	3	194	45	37	4	0	5	3	2	0	0	51%
M	4	0	435	0	12	154	399	112	97	319	5	16	4352	242	21	46	299	15	44	0	0	66%
N	55	0	524	1	11	695	260	190	240	286	9	27	267	25618	13	12	347	9	20	1	0	90%
O	1	0	3	0	3	5	2	1	4	10	0	0	1	1	954	19	53	4	5	0	0	89%
P	6	0	4	0	0	1	4	11	15	16	0	0	18	24	16	848	172	10	7	0	0	74%
Q	3	0	15	0	0	6	12	18	98	24	0	0	35	45	15	107	9747	10	13	1	0	96%
R	5	0	7	0	1	5	20	1	105	5	0	0	11	21	9	7	19	304	5	0	0	58%
S	8	0	22	0	0	29	71	15	64	27	2	2	37	60	32	111	146	6	1393	0	0	69%
T	2	0	0	0	0	3	1	1	9	0	0	0	1	3	0	0	69	0	3	12	0	12%
U	0	0	1	0	0	1	1	1	0	0	0	0	1	1	3	0	0	0	0	0	31	78%
	77%	100%	80%	96%	84%	77%	83%	79%	83%	61%	96%	74%	82%	92%	87%	73%	88%	82%	90%	80%	100%	

Rows: True category, Columns: Categorized category

Source: Own presentation.

Annex D: Greece

7.1 Introduction

ELSTAT, as member of the WP1 of Big Data ESSnet project, investigates the feasibility of using scraped advertisement data from on-line job portals to improve job vacancies statistics. Additionally, under the SGA-II, ELSTAT chose to explore the text mining approaches to automatic classify jobs to ISCO-08 taxonomy. With this test ELSTAT intends to try a first approach to machine learning techniques.

The "landscape" of on-line job vacancy varies significant between EU member states, in terms of the number of job portals as well as level of coverage of job vacancies. A good knowledge of the on line job portals is essential in order to provide a good basis for drawing conclusions on the level of coverage, structure and/or trend of job vacancies in the local level.

The Greek labour market was characterized by a relatively high share of permanent jobs to the total dependent employment on the one hand, and a high share of self-employment that dominated overall employment, on the other hand. However, the depth and duration of the recession have resulted in a severe deterioration of the national labour market and social conditions. One in four jobs that existed before the crisis has been lost. The severity of national labour market distress is, also, reflected in the duration of unemployment. More than 70 per cent of the unemployed have been without a job for more than one year (ILO, 2014).

7.2 Data Access

The purpose of our web scraping experiment was to scrape very specific structured information selected from the job portal. The data were scraped directly from job portals. An internet investigation of on-line Greek job market was carried out first, in order to understand our landscape and to select the job portals.

Briefly describing the followed process, the job portals were sorted based on the following criteria: a) the number of advertisements (size); b) monthly visitors (June 2016) and c) the Alexa popularity ranking. The selected job portals were Kariera.gr and Skywalker.gr. The analysis for the selection of the job portals is presented in the Deliverable 1.1. (2016).

7.3 Data Handling

ELSTAT has focused on scraping ads directly from job portals. The purpose of our web scraping experiment was to scrape very specific structured information selected from the job portal and for this purpose tools for general scraping purposes (such as import.io and content grabber) were used (see Figure 1).

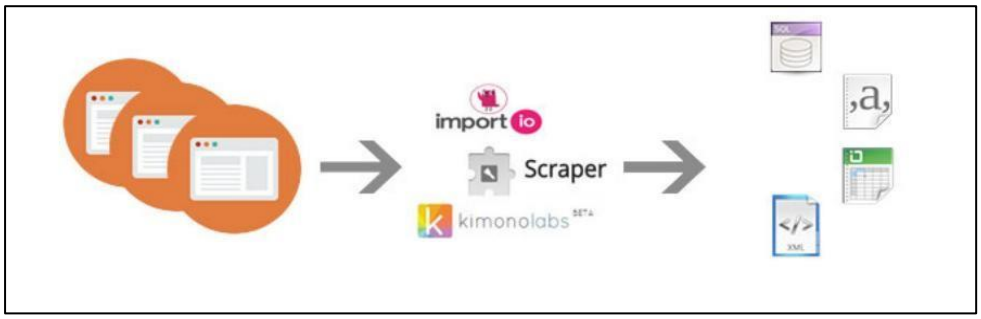


Figure 1: Web scraping process

The scraped dataset includes the following fields:

- job title
- company name
- job description (a "snippet" of the job description between 40-60 words)
- salary and job type (full time/temporary)
- posted date and
- location

The selected job portal has links from the job offers to a second level of standardized information, which consists of the full-text of the job advertisements plus further semi-structured information. The first results reveal that there are problems with missing data, especially as regards the field salary and contract type. In addition, there are problems with taxonomy in the Job Category and Location fields.

7.3.1 Automatic text classification

The main idea of the experiment was to use a phrase based classification (PBC) approach as proposed by Bekkerman and Gavish (2011) to explore the automatic text classification of vacancies to the ISCO-08 taxonomy (4th digit level). Obviously, there are many text classification scenarios. In our case, the ad, which is the text that we want to be classified, is a short commercial document, written in an informal way and following the trend of the local market. In Figure 2, some examples of ads from different Greek job portals are presented.

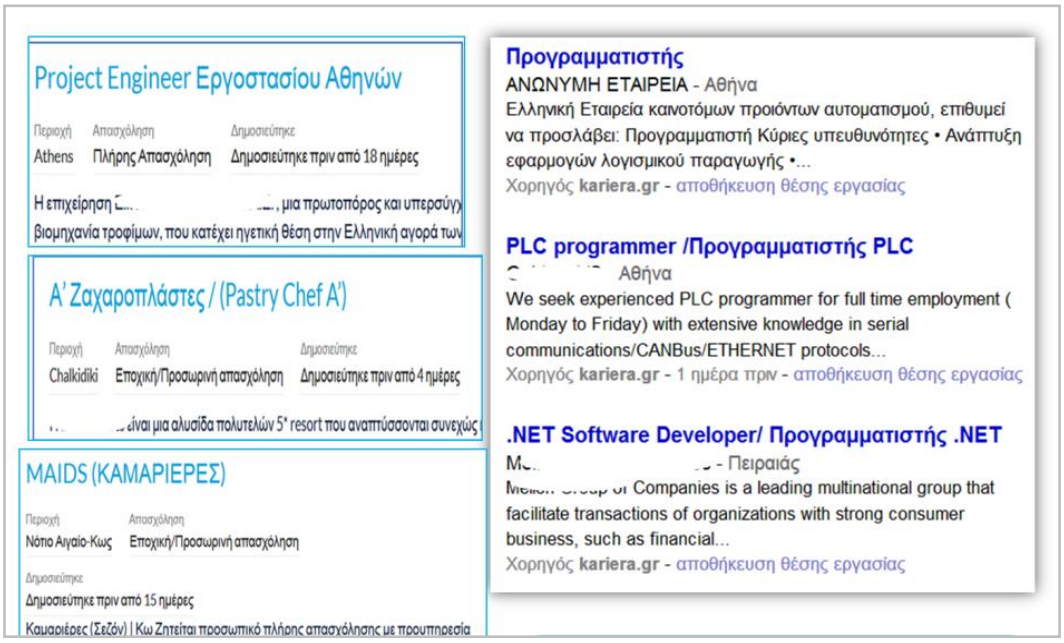


Figure 2: Examples of ads from different Greek job portals

A great number of on-line ads of local market are written not using only Greek language, but in a mixed way using Greek and English vocabulary to describe the vacancy i.e. Project Engineer, Programmer, Software Developer and even for common jobs like Maids or Pastry chef.

Information needed for mapping vacancies to a class of ISCO-08 is mostly concentrated in a few key-phrases appearing in the description of ads. This makes it a suitable scenario to apply a PBC approach. This involves lazy learning from a labelled features setup, constructing a group of decision-stump-like classifiers, each of which is triggered if a document contains a certain feature, such as a word or key-phrase.

This approach is very natural for multilabel text classification (when each document may belong to several classes), as a number of decision stumps may be triggered for classification. An advantage of PBC is the explainability of classification results, which is an important feature of consistency. An ad categorized into a class can be easily explained by the fact that this vacancy contains a key-phrase used to categorize it into this specific class [Bekkerman and Gavish (2011].

The Text Classification is used to assign the category labels to the new documents at the training stage which are based on the knowledge gained in a classification system. In the training phase, a classification system is built using a learning method and a set of documents which are given, attached with class labels.

7.4 Methodology

7.4.1 First Experiment

In the context of the second sprint, the first experiment undertaken by ELSTAT aimed at exploring to what extent the job portal data covers what is measured by the job vacancy survey. The experiment lasted two months, from June to August 2016. The raw data obtained from job portals required

preprocessing first, before it can be analyzed. The removal of the duplicate ads is an important step of this process. Once the data fields have been cleaned, the scraped dataset contained 3060 single advertisements. Only for 59% of the ads (1817) the company names were identified. While the majority of ads without company name started like that: “Leading Company…” or “Well Known Firm...”. In total, 262 company names could be matched. The 49% of these companies were matched with the companies from the Statistical Business Register.

Classification of companies by economic activity reveals that one out of three companies (34%) refers to head offices [management consultancy (20%) and employment activities (14%)]. Additionally, job description classification of the ads by major groups (ISCO-08) reveals that almost 50% of the ads are about Services and Sales Workers. However, there are low percentages of on line ads for several major groups.

7.4.2 Second experiment

In the 2nd experiment, ELSTAT decided to focus on IT domain that is most likely for the jobs to be advertised on line. An API was set up, from November 2016 to January 2017. 925 ads for IT domain were collected. The ads referred to non- IT jobs (28%) and the ads for working abroad (7%) removed. So, the “clean” dataset contains 592 single advertisements. The 89% of company names were identified from the description, which corresponded to 162 company names. 75% (121) of these were matched with the Statistical Business Register (SBR) and 28% of companies were matched to the JVS sample. However, the response rate of these companies in the JVS was low (38%).

The classification results of the companies by economic activity (NACE rev.2) reveal that 31% refer to computer programming, consultancy and related activity companies. However, there is a demand for IT-job from companies in all other Economic Activities in smaller percentages.

Given that the 2nd experiment concerned only the IT domain, we explored a little bit more the job description of the on line ads, since we have noticed that job title are not used in a consistent way (for example, an employee called a “Data Scientist” at one company has a different job title in another one). This is not peculiar due to the fact that there isn’t a clear definition for these new IT jobs. More information about the 1st and the 2nd experiment are presented in the Deliverable 1.3. (2017).

7.4.3 Third Experiment: Automatic text classification

Building upon the 2nd experiment, a classifier has been trained using the dataset of 592 ads from job portal’s IT domain (Training Data Set). The ads were classified in fifteen classes of ISCO-08 taxonomy (4-digits level), in collaboration with Labour Force Survey Section. In Figure 3, the IT tools used in the various processes are presented.

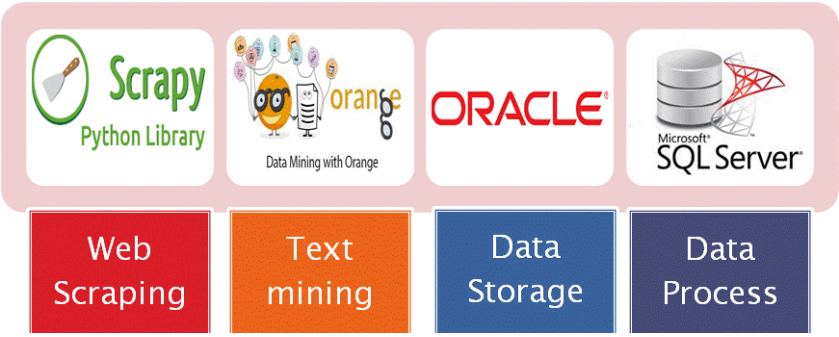


Figure 3: IT tools used in the various processes

7.4.3.1 Taxonomy of Classes

In order to apply a kind of "supervised" learning, we need the categories to be known beforehand and determined in advance for each training document. Indeed, using ISCO-08 taxonomy (4th digit level) for the classification of ads, this precondition is satisfied.

7.4.3.2 Create a Controlled Vocabulary

The next step in the process is to create a controlled vocabulary of words that are supposed to characterize classes in the taxonomy. A semi-manual feature selection process was used for the creation of **controlled vocabulary**.

More specific, first the words of the training dataset were sorted in descending order of their frequency. We scanned through the sorted list in order to decide whether a *word* (Greek or English) should be included in the controlled vocabulary. It is worth mentioning that the training dataset contains mainly four (4) types of words:

(a) job names, such as engineer, programmer etc;

(b) seniority words, e.g. senior, junior etc;

(c) function words, such as research, financial etc;

(d) irrelevant words, most of them with high frequency

A **noisy word table** was built, containing all the function and irrelevant words, used to clean the training dataset. Then, a **translation look-up table** was created, where abbreviations and common misspellings were translated into controlled vocabulary words for standardization reasons. Next, key-phrases were built using **controlled vocabulary words** and an index was developed to assign an ISCO-08 4-digit code to each **key-phrase.**

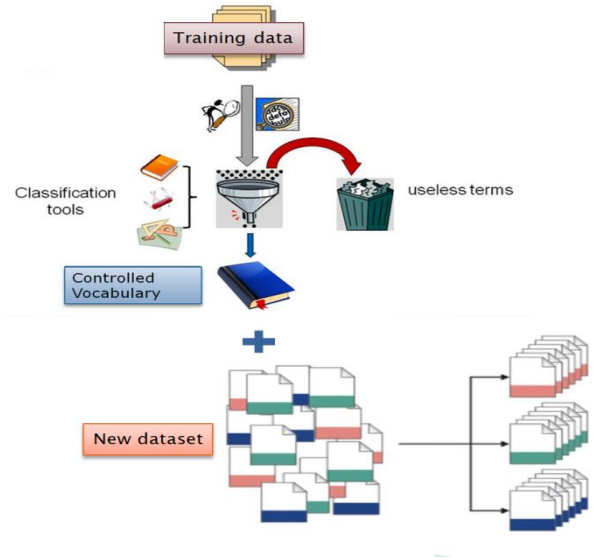


Figure 4: Creation of Controlled Vocabulary

In this way we constructed the decision-stump-like classifiers, each of which is triggered if the ad contains the key-phrase.

7.5 Statistical Output - Validation of classification consistency

We deployed the classification process to the training dataset and it worked well. As it was mentioned above, the training dataset had been classified to ISCO-08 taxonomy (4-digits level) manually in collaboration of the Labour Force Survey Section.

For the validation of the classification consistency, the initial classification was compared to the automatic one. The results are presented in the Table 1. For each class where an adequate number of ads existed in the training dataset, the precision was very high.

ISCO-08	Title EN	Initial classification		Successfully classified	
		Number	%	Number	%
2512	Software developers	180	30%	175	97%
2513	Web and multimedia developers	83	14%	82	99%
2514	Applications programmers	70	12%	60	86%
2511	Systems analysts	59	10%	57	97%
3512	ICT user support technicians	40	7%	29	73%
2519	Software and applications developers and analysts not elsewhere classified	39	7%	36	92%
2166	Graphic and multimedia designers	35	6%	34	97%
2522	Systems administrators	35	6%	34	97%
3511	ICT operations technicians	14	2%	0	0%

Table 1: Validation of classification consistency of the training dataset

The key-phrases could be improved by adding or changing words and indexes, as this is a dynamic process. A new scraped dataset from IT domain was created. It contained 737 ads. The dataset was first pre-processed, cleaned and standardized automatically, as described above. Then, the classification process was deployed and the 64% of ads were automatically classified.

Checking, manually, the classification consistency, we concluded that the 78.6% of automatic classified ads was successful. The validation of classification consistency of the new dataset is presented.

		Automatic classification														Successfully Classified
	Nace rev2	2166	2511	2512	2513	2514	2519	2521	2522	2523	2529	3512	3522	No_ Code	Total	%
Manual Classification	2166	22													22	100
	2511		47												47	100
	2512		1	100	5	3	1						1		111	90.1
	2513			1	48										49	98.0
	2514			4	2	56									62	90.3
	2519		1	1			25					1			28	89.3
	2521							1							1	100.0
	2522			3	1				29						33	87.9
	2523									5					5	100.0
	2529										3				3	100.0
	3512											31			31	100.0
	3522												0		5	0.0
	More than one job/Non-IT jobs /missing info		65	4	3				2		1			0	70	0.0
	Total	22	114	113	59	59	26	1	31	5	4	32	1		467	

Table 2: Validation of classification consistency of the new dataset

7.6 Conclusions

- The process of on-line web scraping job vacancies even if it is easily achievable using various IT solutions, inserts a certain degree of risk due to dynamic changes of job portal web sites market.
- The phrase classification method worked efficiently for the job classification of bilingual on-line scraped ads.
- The vocabulary used in the ads, in both languages was manageable for the size of this experiment.
- Automatic classification was successful taking into account that its consistency depends on the size of the training dataset and the number of ads for each class.
- More work is needed to deal with automatic classification of ads that describe more than one different vacancy, ads with slightly different Greek and English job descriptions or ads with less or misleading information. Expanding this approach to cover all occupation groups would also require further development

7.7 References

1. Greece: Productive jobs for Greece. International Labour Office, Research Department. – Geneva: ILO, 2014.
2. Deliverable 1.1. "Inventory and quality assessment of job portals", 2016 https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_final.docx, retrieved on 8th May 2018.
3. Deliverable 1.3. Final Technical Report (SGA-1), 2017. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/20/Deliverable_1_3_main_report_final_1.0.pdf, retrieved on 8th May 2018.
4. Sebastiani, F. (2002) "Machine learning in automated text categorization." *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47.
5. Ron Bekkerman and Matan Gavish (2011) "High-Precision Phrase-Based Document Classification on a Modern Scale".
6. Kapila Rani et al, (2016) "Text Categorization on Multiple Languages Based On Classification Technique", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3), 2016, 1578-1581.
7. Xiaojin Zhu (2007), "Semi-Supervised Learning Literature Survey", Computer Sciences, University of Wisconsin – Madison. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.9681&rep=rep1&type=pdf, retrieved on 8th May 2018.

Annex E: Portugal

7.1 Objective

Under Big Data ESSnet, WorkPackage1, Portugal decided to test automatic coding for occupation on data extracted from job vacancies websites. The data extraction was done through web scraping with scripts specially developed for this purpose.

For automatic coding, two models were tested: one based on exact string matching and already used in occupation coding of the household surveys and another, developed for this purpose, based on machine learning.

With this test we intend to:

- evaluate the use of web scraping for extracting data from the web;
- evaluate the quality of the web data extracted;
- evaluate the adaptability of the existing coding model to a new context;
- try a first approach to machine learning techniques.

For all the test steps we used free and open source tools.

7.2 Data Access:

There are many job vacancies websites in Portugal, aimed at the public or private sector, general or specialized, with national or international offers; in addition to these job vacancies sites there are also the sites of the agencies specialized in employment, public and private (see attached list).

Some sites are aggregators of offers from other job vacancies sites, i.e., they display repeat job vacancies and refer back to the original sites.

Given our objective (occupation automatic coding) we chose a site that was not an aggregator, in an attempt to avoid duplicates and in order to have somewhat more structured data (however, we do not ignore the challenge associated with identifying duplicate job vacancies).

The site chosen was Sapo Emprego[63] which is one of the biggest and oldest job vacancies site in Portugal. All job vacancies available in Portugal were collected in one single day (about 28,000 offers). The data for each offer was spread over two pages in the website:

- job vacancies listing page - job title, company name, offers date and geographic location; with 20 offers per page, totaling about 1400 pages (accessible by page navigation);
- job vacancies detail page - company description, job description (mix of profile and tasks), offer category, geographic location and offer id.

[63] http://emprego.sapo.pt/, retrieved on 8th May 2018.

Not knowing beforehand, the quality of the data collected that would enable us to do the coding we chose to extract all available data, in a total of 10 variables (job title, company name, company description, job description, category, location, id, url, date).

However, the company description, job description and offer category fields were not used in any of the models because they had noise (diverse data, unstructured) that only confused the models and did not add any value: however, this data, with further work could be usable and useful.

The data extraction was done through web scraping using the Python language and the Scrapy framework (there was no need to use any database since it was a single extraction and the dataset was accessed on 2 local desktops).

7.3 Data Handling:

7.3.1 Dictionary for automatic coding of occupation:

After the data was scrapped, our first approach was to use the automatic coding procedure for occupation. For this purpose we used a classification dictionary for occupation.

Automatic coding for occupation is fairly recent in Statistics Portugal (2017). At this moment it has a production rate of 56.29% and an agreement rate of 91.6% at 2 digit level. Automatic coding is done by exact string matching of human supervised dictionaries to the data to be coded. Portugal's National Classification of Occupations (CPP2010) dates from 2010 and is equivalent to ISCO2008. There are 48 codes at 2 digit level and 135 codes at 3 digit level. The current version of the dictionary has 42628 unique occupation text strings coded at a 2 digit level and 41084 at a 3 digit level.

The dictionary is based on two distinct sources:

- Lists with pairs of text strings and codes and
- Manual coding of text strings from 8 household surveys for a 7 year period (2011-2017)

A third source is derived from the two previous ones: text strings from the household surveys with an Optimal String Alignment[64] metric distance equal or lesser than 3 where added to the dictionary.

The lists are:

- National Classification of Occupations (CPP2010)
- Census 2011: list partly used to classify data collected from CAWI and PAPI (after Optical Character Recognition)
- GPIE: list built-in the software that interviewers use to collect data, basically is an extension our CPP2010.

Data collected from household surveys (2011-2017) that was manually coded amounts for a total of 907505 cases, i.e., occupation descriptions (text strings) classified with a two or three digit code, depending of the type of surveys. These surveys are:

- Labour Force Survey
- Household Finance and Consumption Survey

[64] The Optimal String Alignment distance (method='osa') is similar to the Levenshtein distance but allows transposition of adjacent characters. Each substring may be edited only once.

- Income and Living Conditions Survey
- Adult Education Survey
- Travel Survey of Residents
- Information and Communication Technology Survey
- Consumer Survey Data
- Fertility Survey
- Household Budget Survey

Manual coding from these surveys (text string and corresponding code) where added to the dictionary when the following frequency/consistency rules were observed:

1. String frequency n ≥ 10 and coding consistence of 90%. Some studies refer to inter-coder reliability or coding consistency to vary between 70% when office coders are compared to interviewer field coding and 84% for expert coders (Malte Schierholz, 2014). This frequency rule aims to deal with that potential inconsistency. This first rule precedes the second.
2. String frequency n ≤ 5 and coding consistency of 100%. This enables adding to the dictionary "rare" occupations.

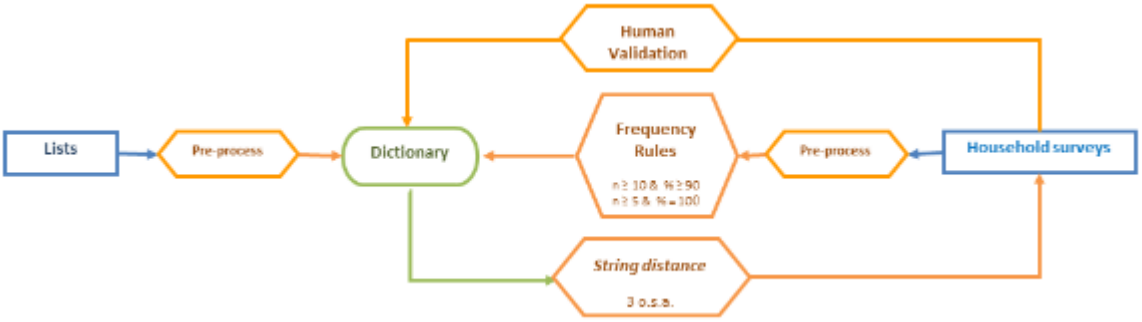
The combined result from these 2 sources (lists and household survey data) was then expanded with strings found in the household survey data that were within a string distance of ≤ 3 o.s.a. The results were human supervised to ensure that strings within that range were actually the same. For example, the Portuguese word for soldier – *soldado* - has a metric distance of only 1 o.s.a from welder – *soldador*. Therefore this very approximate matching had to be rejected for dictionary purposes.

This approach was particularly prolific because there is an extremely high diversity of textual descriptions in the data collected in the household surveys for the each exact occupation. The same word can be written in a multitude of variations due to spelling errors, (mis)use of abbreviation, caps, accentuation or hyphenation, just to name a few. This is quite understandable since interviewers input this data "on-the-fly".

In order to reduce noise and content diversity, all the data was preprocessed:

- Convert to lowercase
- Force encoding: UTF-8 + ASCII (to strip accentuation and other non-alphabetic characters)
- Punctuation removal
- Stopwords removal
- Trim and remove repeated whitespace

Flowchart: dictionary for occupation automatic coding



R packages used:

- stringdist by M.P.J. van der Loo
- tidyverse by Hadley Wickham

7.3.2 Quality Control:

The Occupation Dictionary was reviewed by teams of coders and the automatic classification simulations for 3 household surveys were also thoroughly reviewed by them. These procedures resulted in corrections such as:

- Removal of text strings that were clearly insufficient to code and
- Resolving inconsistencies between very approximate text strings with the same meaning but different codes.

This is still an ongoing process.

7.3.3 Classification by Exact String Matching:

The first step was to preprocess all job offer titles available with the same algorithm as the dictionaries in order to get cleaner text strings.

As we analyzed the raw data, it became evident that further preprocess was due in order reduce verbatim. The mean character length of the dictionary strings is 26 while the job vacancies title to be coded is 34. This is not unexpected as the data wasn't originally produced for classification purposes. We found that job offer titles often contain regular references to:

- Gender (ex.: male / female)
- Amount of working hours (ex.: 3h30, 7h00)
- Part of the day (half time , full time, mornings, afternoons)
- Usual expressions such as "hiring...", "job offer for ...", "urgent need for...", "wanted ..."

These were removed with use of REGEX or a new list of specific stopwords.

Another regularity detected in the job offer titles was the reference to locations. Since none existed in the dictionary, we opted to remove them. It was used a list of localities for the whole country for the 3 existing levels of administrative regions. This list was preprocessed, tokenized, duplicates removed and used as an additional stopwords list.

Nevertheless this list had to include some exceptions since some words used in local names have a relevant meaning to job classification. Words such as "vendas" and "saude" which are common in names for locations are the same exact words for "sales" and "health" (direct translation). In order to only remove those words when not referring to locations would require a specific algorithm witch wasn't available at this time. So the option was to not remove these words from the job offer title even if they could refer to a location rather than a specific area of job offer.

7.3.4 Results and final thoughts

Data was then classified by exact string matching with poor results as expected: 18.9% of job vacancies were classified by this method (n = 5226) a much lower value than the average 56.29% production rate when using this method with data from household surveys. This clearly demonstrates that data collected with a purpose other than occupation classification poses a serious challenge to an exact string matching approach.

Further data cleaning either of the job vacancies title or job description would improve the production rate but surely not with satisfying results.

7.4 Methodology

For the coding test with supervised machine learning we used Python and the SciKit-Learn library in a Jupyter Notebook environment.

7.4.1 Model selection:

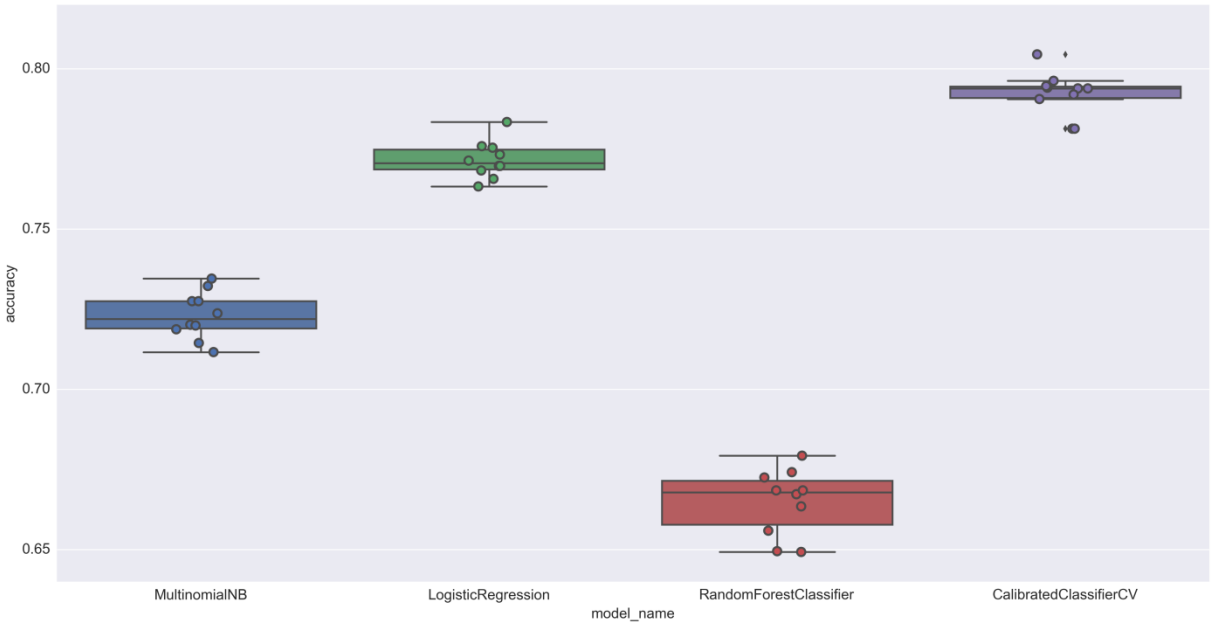
For model selection we used the dictionary from B1 as training set; 459 coded strings related to armed forces occupations were excluded because are not relevant for this study. We ended with 42.169 coded strings.

For feature extraction we used the Bag of Words model with unigrams and bigrams; all features were used (42.169 x 50.177 sparse matrix).

The algorithms Logistic Regression, MultinomialNB, Random Forest and SVM (Linear Kernel) were tested.

A 10-fold cross validation procedure was used to evaluate each algorithm, configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm was evaluated in the same way.

Figure 1 - Model selection



The algorithm chosen was the **SVM with Linear Kernel (CalibratedClassifierCV)** because performs better than the other tree classifiers, with a median accuracy of around 80% and 0,7% std. On the other hand and as we will explain later, this algorithm presents a higher agreement rate with the results obtained in B1.

7.4.2 Model evaluation (SVM with Linear Kernel)

The dataset was split into training/test data with a test_size=0.2.

The test produced an accuracy = 78%, a precision = 0.81, a recall = 0.79 and a f1-score=79%.

In the 3 figures below we can observe the confusion matrix (CM) that shows the discrepancies between predicted and real labels.

The Figure 2 show the CM without normalization; the majority of predictions end up on the diagonal (predicted label = real label).

There are however a number of misclassifications. To have a better representation of which class is being misclassified we normalized the CM by class support size (number of elements in each class); the CM in Figure 3 was normalized by row (the diagonal is the recall) and in Figure 4 by column (the diagonal is the precision). There are classes that stand out (81, 33, 12, 25 and 94) and we must pay attention to them when we will do the job vacancies classification. The normalization is also important in class imbalance representations, which is the case.

Future working concerning model selection and evaluation:

- Analyze and tweak precision and recall for some classes;
- Find a model to balance the classes;
- Find more algorithms in Sklearn to implement and optimize;
- Explore different algorithms for feature selection, and try out combinations between feature selection parameters and models;
- Explore GridSearchCV to optimize parameters.

Figure 2: Confusion matrix without normalization

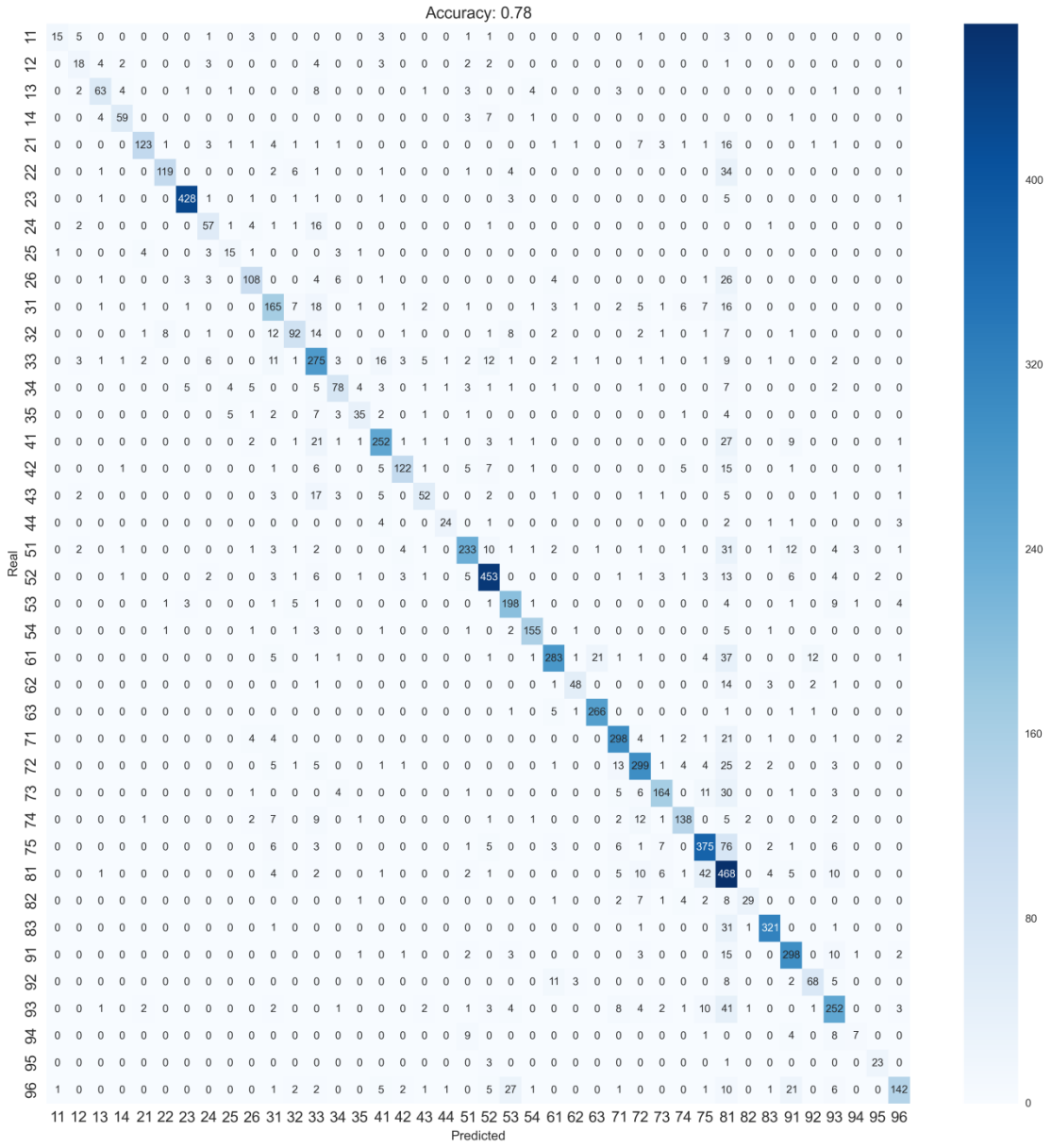


Figure 3: Confusion matrix with normalization (diagonal = recall)

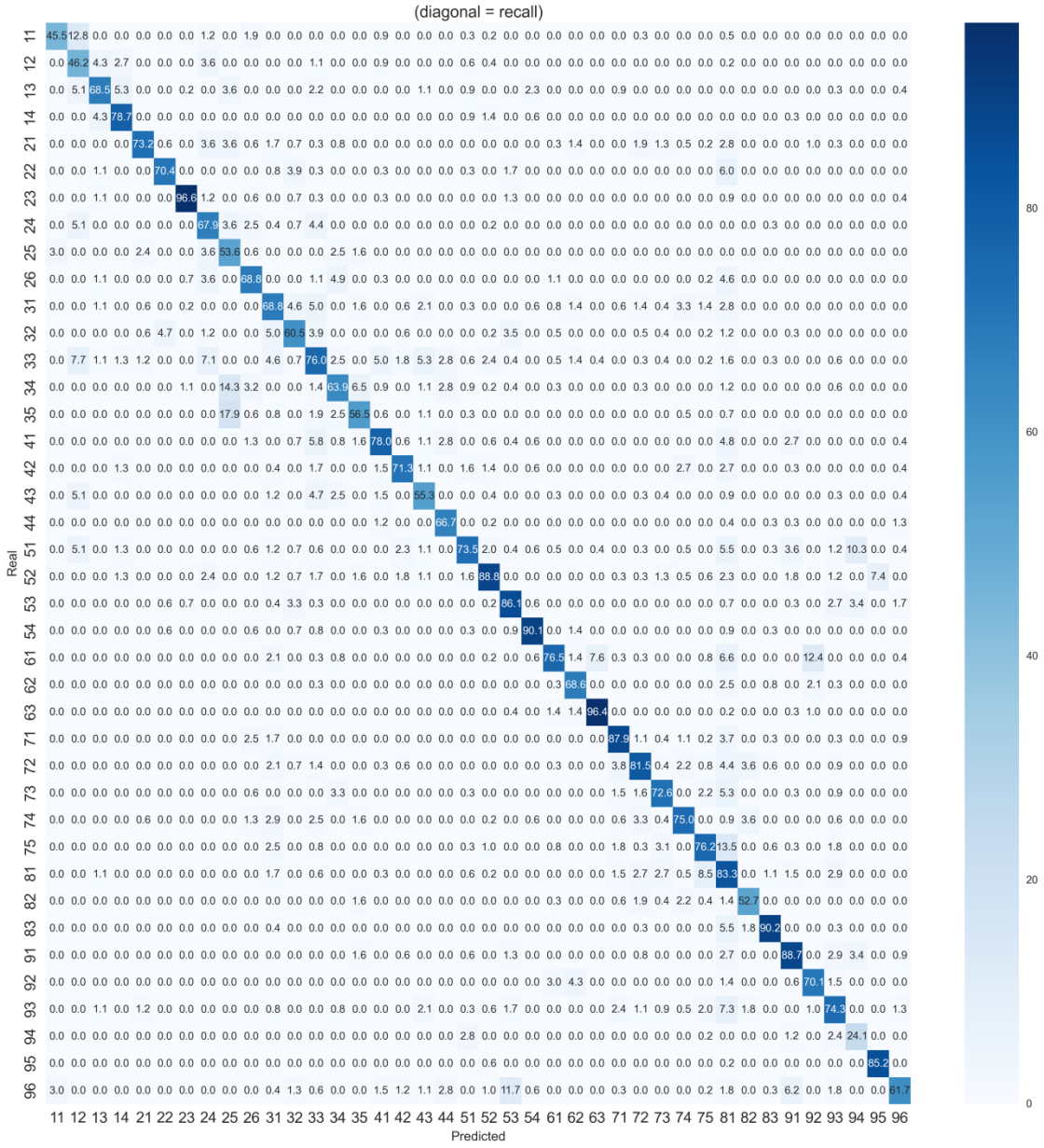
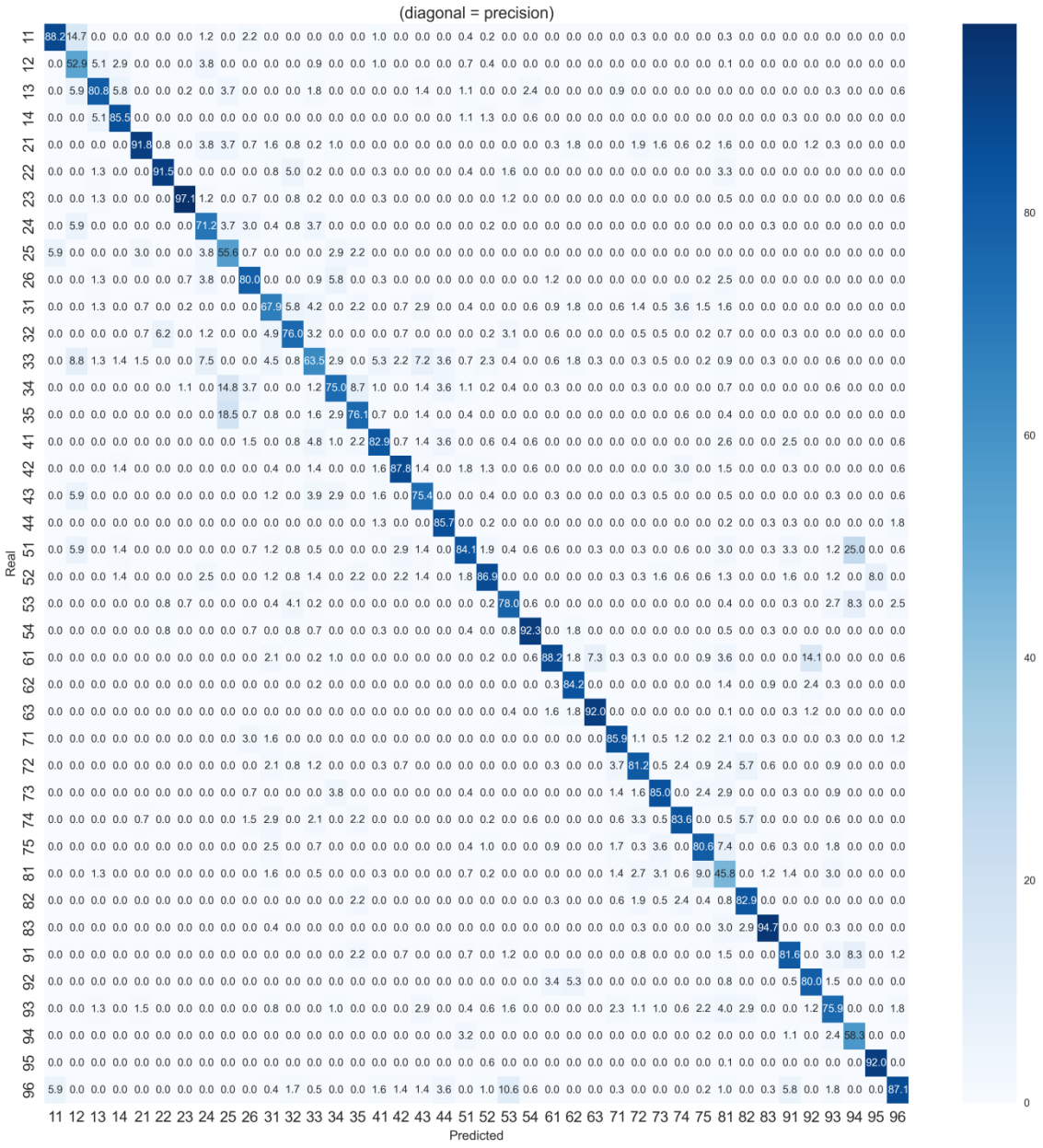


Figure 4: Confusion matrix with normalization (diagonal = precision)



7.4.3 Job vacancies classification

For job vacancies classification only the title job field was used, since the other job vacancies (company description, job description and offer category) greatly shuffled the algorithm.

7.4.3.1 Preprocessing tasks

These are the preprocessing tasks that job title was subject:

- Convert to lower case;
- Remove accents;
- Remove punctuation;
- Remove Portuguese StopWords;
- Remove specific words that are not relevant for the occupation classification but which shuffle the classifier, because they have double meanings (ex: shifts, place names like “saúde” – health or “vendas”- sales, etc.);
- Remove whitespaces and tabs.

7.4.3.2 Classification and evaluation

About 97% of job titles coded with the B3 method are classified in the same way with this classifier, which gives us some assurance about the quality of the remaining jobs codes.

It was not possible so far to assess in depth the quality of the classification for all classes but a preliminary analysis proved to be very promising.

However all classes should be subject to a deeper analysis. One of the relevant issues was related to class 81 that accumulated a great part of the job titles impossible to code (class with less precision in the training).

Despite this incipient analysis it is already possible to identify some job titles description particularities (both dictionary and job vacancies) with impact to the classification and that should be considered in future analysis.

Job Vacancies:

- insufficient job title description for classification (ex: “worker”, “employee”, “trainee", “looking for talents”, etc.);
- job titles with “new” occupations that do not exist in the dictionary (ex: IT jobs);

- job titles in foreign language (in the future we could apply language detection algorithm and do the translation);

Dictionary:

- occupations that have only the plural or singular form;
- occupations that only have feminine or masculine form;
- orthographic errors - this prevent the application of lemmatization to solve inflection and gender forms;
- inconsistencies - same meaning, different code - evaluate withdraw some sources;
- missing text strings for emergent occupations (IT, tourism, etc.).

It is our understanding that inflections and gender normalization and inconsistencies correction will allow a substantial increase in classification.

The existing stemming tools for Portuguese are inefficient (porter, snowball). Concerning lemmatization there is at least 1 Portuguese dictionary for Hunspell (natura project) but the orthographic errors in the text strings of the dictionary are a deterrent of its use.

7.5 Conclusions and Lessons Learnt

Web scraping has proven to be very efficient to extract data from the web and is a good alternative when there are no other sources; once the scripts have been developed, the information can be extracted at any time.

Huge amounts of data are crucial for training machine learning models. The web provides this scale but also brings noise due to unstructured nature. And that's why the preprocessing tasks are so important. The dictionary from B1 was developed in a household surveys context so is not so sensitive to the kind of verbatim used in job vacancies web site. The lesson here is: one dictionary for each purpose.

To combine the two models (exact string matching and machine learning) can bring great changes in occupation coding of household surveys not only for what they can do *per se* but also for the benefits they give each other. This was one of the most valuable unexpected benefits from this test.

When comparing with the official job vacancies statistics (produced by Portuguese public employee agency – Instituto do Emprego e Formação Profissional), a first analysis reveals a very different size and structure. There are some occupations found on the web job sites with no expression in official statistics and vice versa.

The existing information is still not enough to draw definite conclusions but it raises a lot of questions for which answers should be sought:

- How to measure the credibility of the web sites? By the number of job vacancies?
- Are the web job vacancies all real?
- Job vacancies no longer available are removed from web sites?
- Which sectors of activity are covered by these sources?

Appendix – Job vacancies web sites

Job vacancies sites

Sapo emprego - emprego.sapo.pt

Net Empregos - net-empregos.com

Expresso Emprego - expressoemprego.pt

Empregos Online - empregosonline.pt

Carga de Trabalhos - cargadetrabalhos.net

Bons empregos - bonsempregos.com

Ofertas de Emprego - ofertasdeemprego.pt

Alerta Emprego - alertaemprego.pt

Trabalho Certo - trabalhocerto.pt

IT Jobs - itjobs.pt

Turijobs - turijobs.pt

Fashionjobs - fashionjobs.com

Empregos IT - empregosit.pt

Landing.Jobs - landing.jobs

Emprego Saúde - empregosaude.pt

OLX - Empregos - olx.pt/emprego

LinkedIn Jobs - linkedin.com/jobs

Employment agencies (public and private)

BEP - Bolsa de Emprego Público - bep.gov.pt

IEFP - Instituto do Emprego e Formação Profissional - iefp.pt

Garantia Jovem - garantiajovem.pt

Netemprego - Ofertas de emprego do IEFP e Estado - netemprego.gov.pt

Hays - hays.pt

Michael Page - michaelpage.pt

Egor - egor.pt

ManpowerGroup - manpowergroup.pt

Talenter - talenter.com

Adecco - adecco.pt

Randstad - randstad.pt

Job vacancies aggregators

indeed - indeed.pt

Jobtide - jobtide.pt

Careerjet - careerjet.pt

Trovit - trovit.pt

Annex F: Slovenia

7.1 Data Access

7.1.1 Background

The Slovenian law requires employers to publicly advertise job vacancies, but there is no obligation to send such information to any agency. The advertising can be done online, in a newspaper, on billboards, or even just on the bulletin boards inside the enterprise building if it is publicly accessible. The advertisements must be accessible for the minimum amount of three workdays. The exceptions to these laws are state owned companies and public administration, which have by law the obligation of sending job vacancies information to the Employment Service of Slovenia. These vacancies are available to us as an administrative source.

On the other hand, collecting full information about private companies' vacancies without direct questioning is therefore quite hard and is usually carried out with a job vacancy survey on a representative sample of enterprises. The survey is an additional burden for companies although the questionnaire is extremely rationalized. The data required by regulation, regarding the occupied post are obtained from the Statistical Register of Employment. With the questionnaire we collect only the number of vacancies on the reference day. To cover all fields of activity and all companies with at least one person in employment, almost 9,000 business entities are included in the sample (which is 19.1% of such companies registered in the Business Register of Slovenia).

With improvements as the goal (such as abolishing or shortening the survey), the ESSnet Work Package 1 (WP1) group has tried to implement a scraping process that would be able to effectively evaluate the number of job vacancies using the number of vacancies found online. SURS also makes part of the ESSnet WP1group and we have been exploring the situation in Slovenia.

So far our analysis of the situation shows that there are two main Slovenian job portals (Moje Delo - mojedelo.com, Moja Zaposlitev - mojazaposlitev.net) that advertise the biggest share of published job vacancies. The Slovenian team decided to collect (scrape) the data from those two job portals. For the purpose of scraping, Agenty (Agenty.com; former name of the application was Data Scraping Studio) was used. Scrapers were run manually every Monday morning from mid-May 2017 onwards until the end of the year, when the scraping application was changed. Afterwards custom-made programs were employed for scraping of the two websites. When survey on job vacancies collects data (on reference day), we scrape data on the reference day and the next day. Job vacancies that were valid on the reference day are those, which were scraped on the first day and were published during that day. Additionally, if a job vacancy is being published more than a month, we treat it as invalid job vacancy. On the day of scraping, each job portal is being scraped two ways: first we scrape a list of all job vacancies on the domain, and second we scrape a content of certain job vacancy. Following data are being scraped:

- Job vacancy title
- URL address of job ad page
- Company name
- Place of work
- Date of published of the job ad

- Description of job vacancy
- Description on how and until when to apply for the job (only on Moja Zaposlitev),
- Date of scraping.

Weekly about 2.000 – 2.400 queries are run (depends how many job vacancies are published).

7.1.2 Accessing Job Portals' data

We have secured a contract with Moja Zaposlitev, which gives us permission to scrape their site for job vacancies once a week. With Moje Delo we are still in the phase of making an agreement, but nonetheless we scrape their pages as well. We do so while adhering to the 'netiquette': we follow the Robots.txt exclusion protocol of the pages and the scraping is executed in the night/early morning, as not to interfere with server traffic.

7.1.3 Enterprise sites scraping

We also scrape enterprise websites on a sample of all existing domains. The sample is around 6,000 domains big and includes all known domains of big (250+ employees), medium (50-249 employees) and small (10-49 employees) enterprise and around 2,000 micro enterprises. Due to the big amount of pages, scraping takes place over several days at the end of the second month of a quarter (a reference day for the job vacancy survey).

7.2 Data Handling

Presently, data is obtained with Python programmes (Job Portals) using Selenium and BeautifulSoup and Python Scrapy programmes (enterprise data). Previously we used the freeware crawling and scraping Data Scraping Studio (presently known as Agenty); however it was reorganized into a browser application that ceased to suit our needs.

Scraped portal data are arranged into a table with information about internet address of the job vacancies, the title of the vacancy, dates of publishing, name of the company (this being the actual company advertising the job vacancy or an agency that lends its workforce to the companies), the code of the portal where the advertisement was found, request and response URLs and date-time variables, the full text in the unstructured part and, where available, the maximum application dates. These data are then processed in a SAS, where a deduplication-linkage process takes place.

The data is saved in comma separated value files on our disks and in the case of enterprise data into .html files to be used as check throughout the process.

Enterprise pages data is analysed and processed in the data mining and machine learning toolkit Orange where a table of information about the internet address, date of scraping and full website text is created. Additional columns, representing probabilities for numbers of advertisements present are added.

All sources are combined in the end to calculate the number of online job vacancies in SAS, which are then saved to SAS table files.

Since Slovenia is small, the size of these files does not represent complications. The usual size of a job portal scraping is barely 4,000 job advertisements (before deduplication) and represents only 6 MB of data. On the other hand enterprise pages scraping results in less potential job vacancies subpages (only 1,500-2,000) while still being more demanding in terms of size (around 8.5 MB). If we include auxiliary .html files this goes up to around 400 MB per session. Keeping in mind that job portals are scraped once a week, while enterprise pages are scraped only once a quarter, we accumulate around 0.5 GB of data every quarter.

7.3 Methodology:

7.3.1 General issues with scraping Job Vacancies at SURS

Our observed population are all business entities as a whole (LU), registered on the territory of the Republic of Slovenia which had at least one employed person when the sample was prepared. Natural persons who have no employees besides themselves are not the target population. Included are business entities with registered main activity from B to S. The reference day is the last working day in the middle month of every quarter.

We only consider job vacancies that are not going to be filled by unpaid trainees, contract workers (who are not on the payroll), employed persons returning from paid or unpaid leave, or persons who are already employed in the firm and who will occupy a post as a result of the reorganisation of the firm.

As an additional and reliable source beside survey data we also use an administrative source: the Employment Service of Slovenia (ESS). This is because the employers in the public sector and state-owned companies have to report vacancies to the ESS. With respect to the official estimations (quarterly data published by the Job Vacancy Survey), the share of all job advertisements from administrative sources represents 25 – 35% of all job vacancies. We also manage to collect an additional 10 – 15 pp. from scraped data. So the total coverage of job vacancies is around 45 % according to survey.

We expect the same characteristics as described above in the scraped vacancies. However we need to be mindful of over-coverage as quite a share of published vacancy ads are for student work or work outside of Slovenia. Furthermore there is the issue of under-coverage as the probabilities of advertising on the internet in each activity are not likely to be the same. Therefore there is surely some bias in the representativeness of advertisements for some activities.

Also a big issue to consider are duplicates of the same job advertisement in either multiple periods and on both portals as well. These represent around 8% of the share of all present job advertisements on these two portals.

Other issues include linkage of companies on portal advertisements to their actual register numbers and the numeric and linguistic mismatch between the number of job advertisements and the number of job vacancies. Namely, a job advertisement could represent more than one job vacancy for the same post even if this is not represented in the language of the advertisement description. In some events a job advertisement could also represent zero job vacancies; these are the so-called “ghost vacancies”. An enterprise would choose to publish a ghost vacancy in the event when they have no vacancy posts but are left with some advertisements on a job portal, so they dump them or use them for collection of CVs of people they might consider in the future.

These considerations do not fully extend to enterprise web pages. Ghost vacancies can be dismissed as the language differs when an authentic job vacancy is present on the page. The wording for an authentic job vacancy usually includes an explicit call for new employees with specific skills, while a CV collection page is usually less demanding, inviting every prospective employee to join the enterprise if they want by sending their CV.

Unfortunately, under-coverage is more severe as aside to the problems of representation above, we cannot be sure we included every existent internet page in the set of pages to be crawled and scraped. Over-coverage is also still present, with the same issues as with job portals.

7.3.2 Methodology of Job Portals data

Presently, we scrape job portals data once a week. Additional scraping is executed on reference days of the job vacancy surveys for the purpose of comparison between the two.

Presently we use two methodologies on Job Portals data. However the some of the steps of deduplication are shared with both methodologies. The following steps are a deduplication-linkage process that takes place for every week of scraped data:

1. We search for duplicates within periods of each scraped dataset according to the internet address: these duplicates are actually the same advertisement that has been present on the internet site over the last weeks. Since the advertisement does not change, the only difference between duplicates is the scraping date. We clean off every row in the datasets containing such duplicates except for the most recent ones.
2. Some of the publishing date fields have inadmissible characters like “Published:” or “Yesterday”, “Today” etc. We change that to the actual date of publishing matching the format of other advertisements.
3. All dates are separated into columns containing years, months, weeks and days.
4. A linkage process takes place, linking the companies in advertisement to their Business Register ID number (register number) and cleaning off the duplicates present in both datasets. The register holds the companies’ short and full names and their headquarters (HQ). It also holds information about their local kind-of-activity units (LKAUs) with the same variables that will have to be considered at linkage. The process is as follows:
    a. cleaning the data: all vacancies outside Slovenia and student work vacancies are removed as well as some characters (“&”, “+”, “/” and so on),
    b. linkage of company names in the vacancies based on unique short names in the register (around 98% of all companies have a known unique short name),
    c. linkage of the remainder of company names in the vacancies bases on the unique abbreviated names in the register,
    d. linkage of the remainder of company names in the vacancies based on the unique full names in the register; these steps link around 90% of all found companies to their register entries,
    e. there are some companies that contain their HQ location in their registered name, so to link those, we need to eliminate the HQ location from there; afterwards we repeat the previous steps; this results in additional 2 % of companies linked,
    f. the rest are still not linked, and that is because their names in the register are not unique; therefore we try to link them with the name and location in the vacancies

compared to the short names and registered HQ location of the companies. Unfortunately this is only effective with the short names,

g. for companies that do not use the exact name given to the register in their advertisements, an asymmetric spelling distance function is used; a list of companies that have the distance between the official short name and the name used in the vacancies under a chosen threshold is created, and we have to manually choose the right company,
h. the above is done with the full name as well; both steps give us another 2.5 – 3% of company links,
i. non-linked companies have to be handled manually; some 3% of all names are usually the same ones that were present in the previous periods, so we already have links prepared; the other remaining 1.5 – 2% are handled in this period.

5. After these steps, we have a union of both datasets with unique vacancies and register numbers for companies; however we must search for advertisements published on both portals. So we execute the so-called "soft" deduplication between portals, based on the register number of the company, occupation and location in the vacancies.

An example of the distribution of publications on portals of a deduplicated dataset is presented in the table below (see Table 17):

Table 17: Frequencies and shares of ads on job portals on the reference day 30th November 2017

Published	frequency	share (in %)	cumulative frequency	cumulative share (in %)
only on mojedelo.com	2,471	55.00	2,990	55.00
only on mojazaposlitev.net	1,736	38.64	4,207	93.64
on both or unable to determine duplication	286	6.36	4,493	100

7.3.3 Distribution fitting

Connecting the advertisements from the Employment Service of Slovenia with results of a Job Vacancy Survey and M-forms (forms for compulsory health insurance application for new employees), we have obtained a list of time periods between the first occurrences of a job vacancy we scraped and the moment when the vacancies are filled. It must be noted that these data are extracted from a source that only collects information about state owned and public administration employers and might therefore be biased.

For the purpose of understanding our data, we have tried to fit the empirical intervals to a probability distribution. In the process we have tested eight distributions, some of them in an analytical way and others, with more complicated distribution formulas, in a numerical way.

Since logic dictates that our data should be comprised of periods of days and therefore are non-negative integers, we have primarily focused on distributions on non-negative numbers. However our data was not perfect and we have some negative values in the sets. We have also tested all the

distributions described below on a subset of all non-negative values of our data and a transformation of all negative values to 0.

We used the Maximum Likelihood Estimation methods to find the best parameters for the following distributions:

- geometric $Geom(p)$,
- exponential $Exp(\lambda)$,
- Poisson $Poiss(\lambda)$,
- Lévy $Lévy(\mu, c)$ with $\mu = \min(\text{data})$.

The obtained parameters' estimations were those were the log-likelihood of their distributions reached their minimums.

Additionally we have used the moment method to find parameters for four additional distributions:

- gamma $\Gamma(\alpha, \beta)$,
- chi square $\chi^2(k)$,
- log-Cauchy $LC(\mu, \sigma)$,
- beta $\mathrm{B}(\alpha, \beta)$, with normalisation of data onto the interval $[0, 1]$.

These parameters were found by solving moment equations for the first moment $\mathrm{E}[\mathrm{X}]$ and second moment $\mathrm{E}[\mathrm{X}^2]$. The estimator for mean was the average $\bar{\mathrm{x}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_\mathrm{i}$ and the estimator for variance was the s^2estimator: $\mathrm{s}^2 = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2$.

Lastly we have found an estimate for the parameter of a Weibull $\mathrm{W}(\alpha, \lambda)$ distribution with an iteration to find the solution of the moment method equations for the parameter λ:

$$\ln\left[\frac{\bar{\mathrm{x}} + \mathrm{s}^2}{\bar{\mathrm{x}}}\right] = \ln\left[\frac{\Gamma\left(1 + \frac{2}{\lambda}\right)}{\Gamma\left(1 + \frac{1}{\lambda}\right)^2}\right].$$

We have decided that the acceptable error in the above equation should be lower than 0.00005 and have been incrementing λ by 0.00001 in each step of the iteration. Afterwards the second parameter was easily calculated from the first moment equation using the $\hat{\lambda}$.

Results

Our datasets have the following designations:

Set V: list of numbers of days between the first occurrence and the filling of a job vacancy – a list of 24,268 periods ranging from -3580 days to 894 day in a period with high spikes of occurrences for the periods of -1 to 10 days; the maximum is at 7 with 522 occurrences. The median is a 31 day long period.

Set X: subset of V, where the possible values of days start at 0 and the negative periods are skipped – a list of 23,526 periods ranging from 0 to 894. A high spike is present right at the start of the list with the period of 7 days again being the most prevalent with 522 occurrences. The median is a 33 days long period.

Set Y: all negative valued periods have been transformed into the period 0. This way the period of 0 days is the most prevalent period with 1165 occurrences. The length of the list is again 24,268 and the median stays a 31 days long period.

The graphs of these period occurrences are present in the Figure 10 (note - the dataset V extends to the left until reaching -3580):

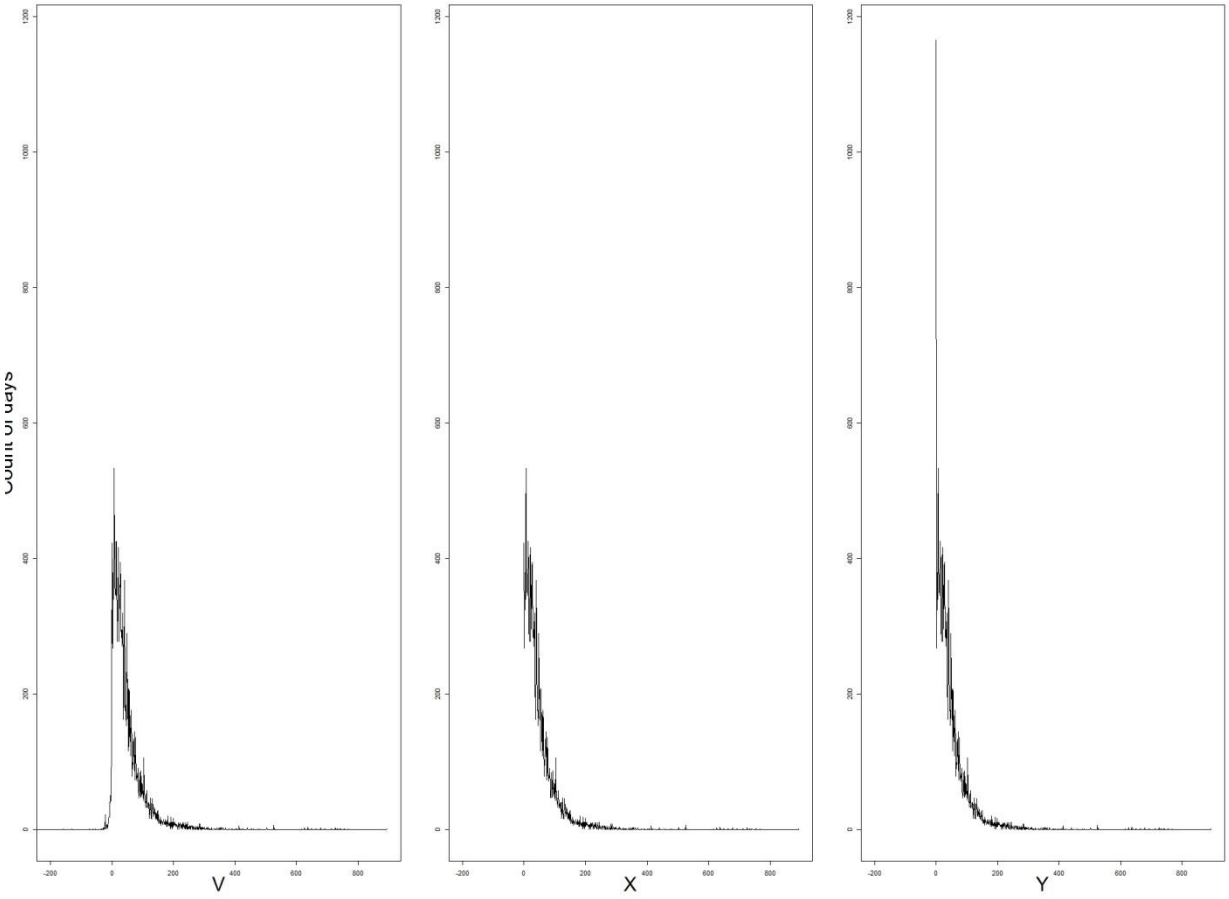


Figure 10: Graphs of occurrences of periods from date of publishing JV to employment

Our findings indicate that the best distributions with optimal parameters are the $\mathrm{Geom}_{\mathrm{V}}(0.0214)$, $\mathrm{Exp}_{\mathrm{V}}(0.0219)$, $\mathrm{W}_{\mathrm{V}}(0.866, 42.47)$ and the $\Gamma_{\mathrm{V}}(0.59, 0.013)$ for the dataset V. As you can see from the graph below, none of these distributions account for the negative periods. In reality we could not find a decent fit for those values with any distribution.

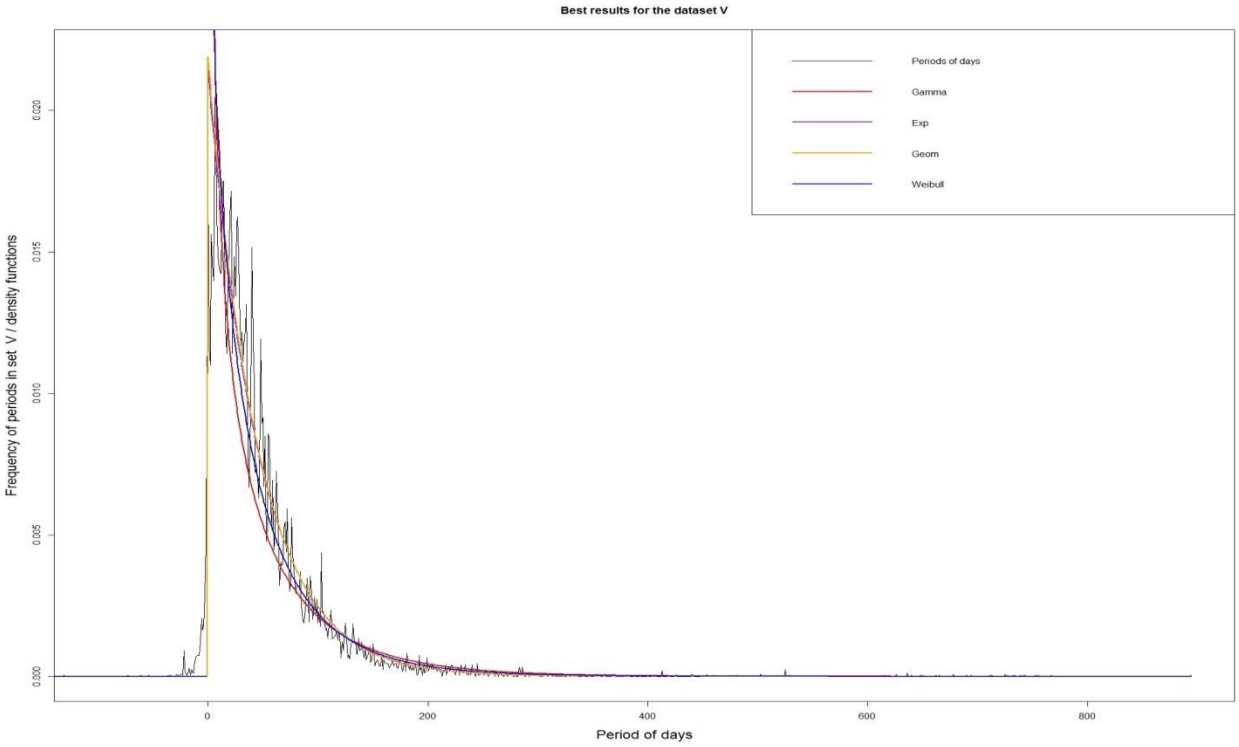


Figure 11: Comparison of frequencies of periods with the best optimal distributions for set V

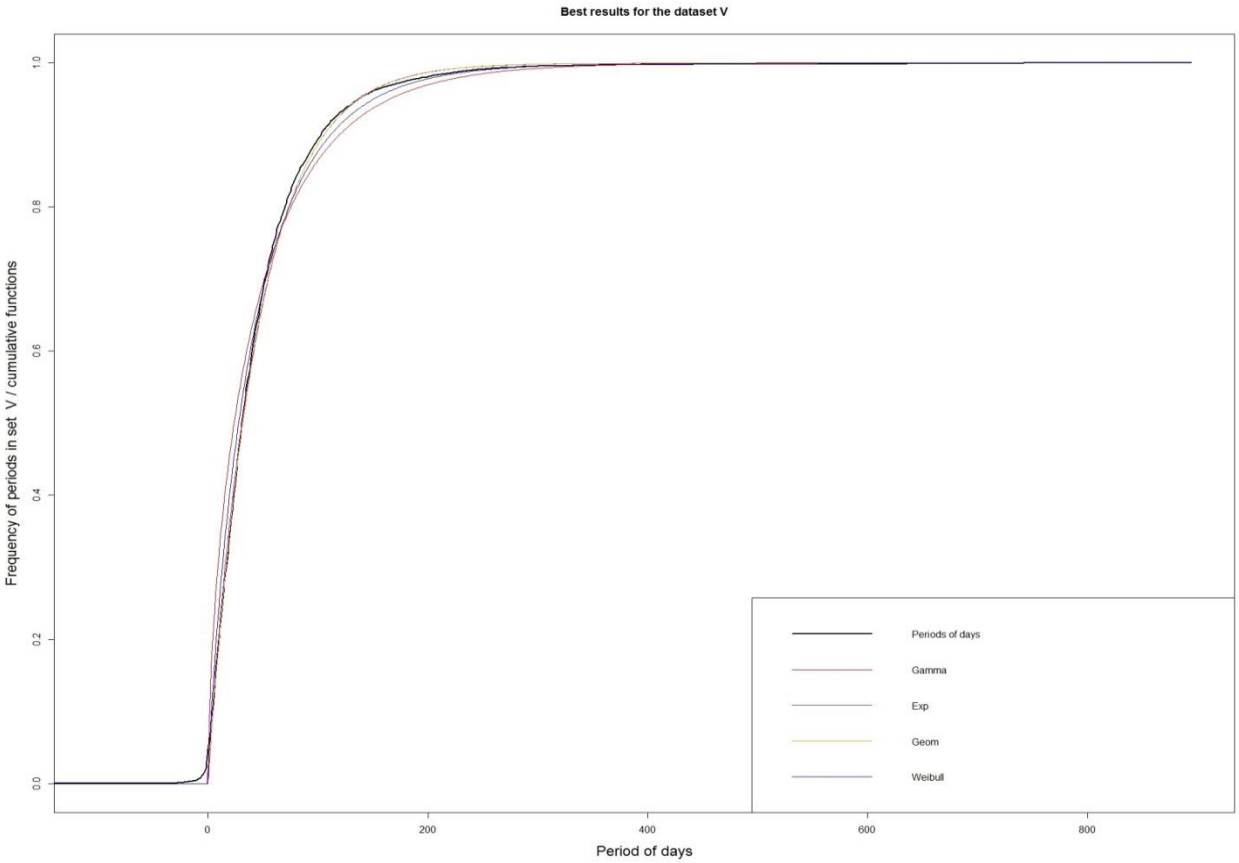


Figure 12: Comparison of frequencies of periods with the best optimal distributions for set V

For the dataset X and Y the best distributions seem to be similar to the above ones. For the set X we have $\mathrm{Geom}_X(0.0206)$, $\mathrm{Exp}_X(0.021)$, $\mathrm{W}_X(0.937, 46.27)$ and $\Gamma_X(0.615,\ 0.017)$.

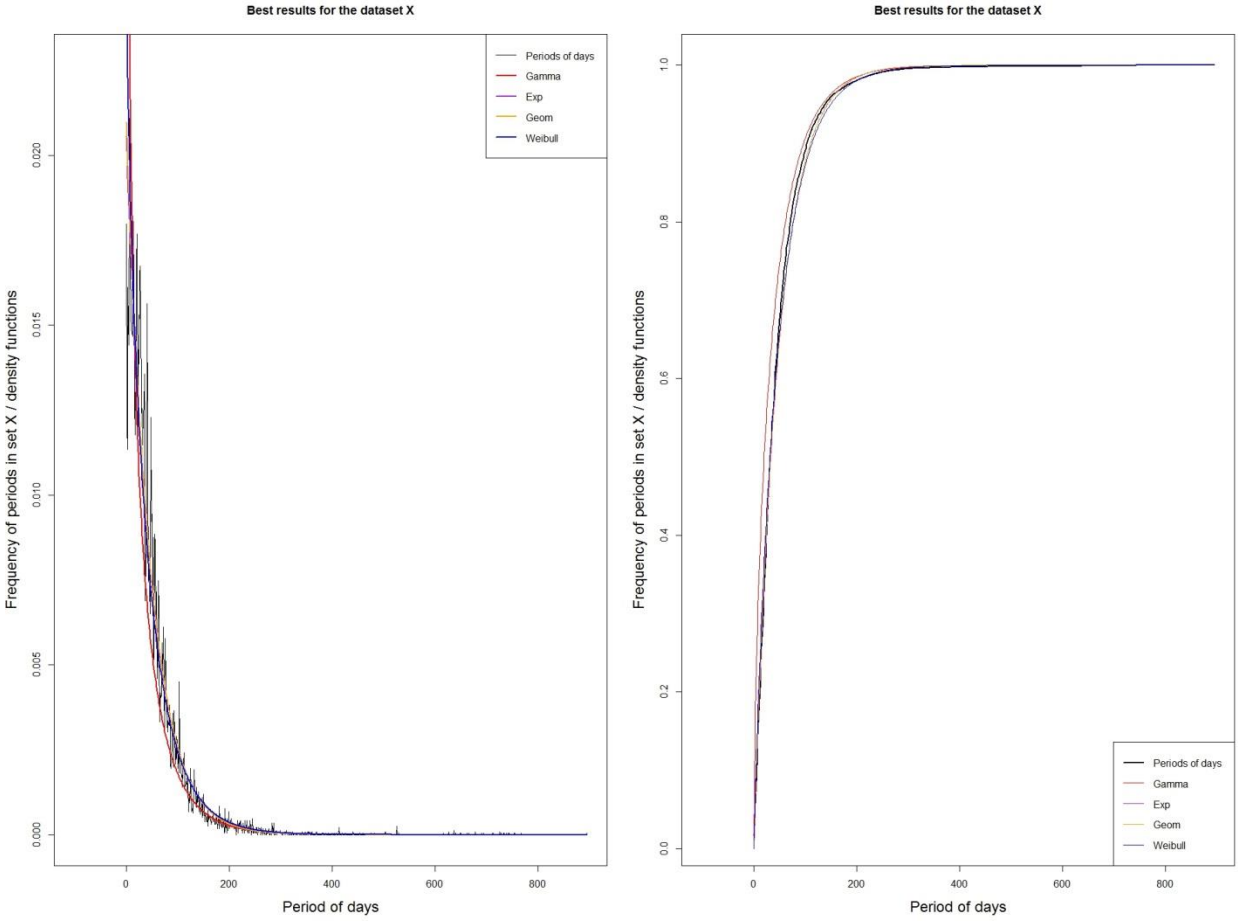


Figure 13: Frequencies and density/cumulative functions of set X

And for set Y we have $\mathrm{Geom_Y}(0.0212)$, $\mathrm{Exp_Y}(0.0216)$, $\mathrm{W_Y}(0.936, 46.27)$ and $\Gamma_\mathrm{Y}(0.745, 0.016)$.

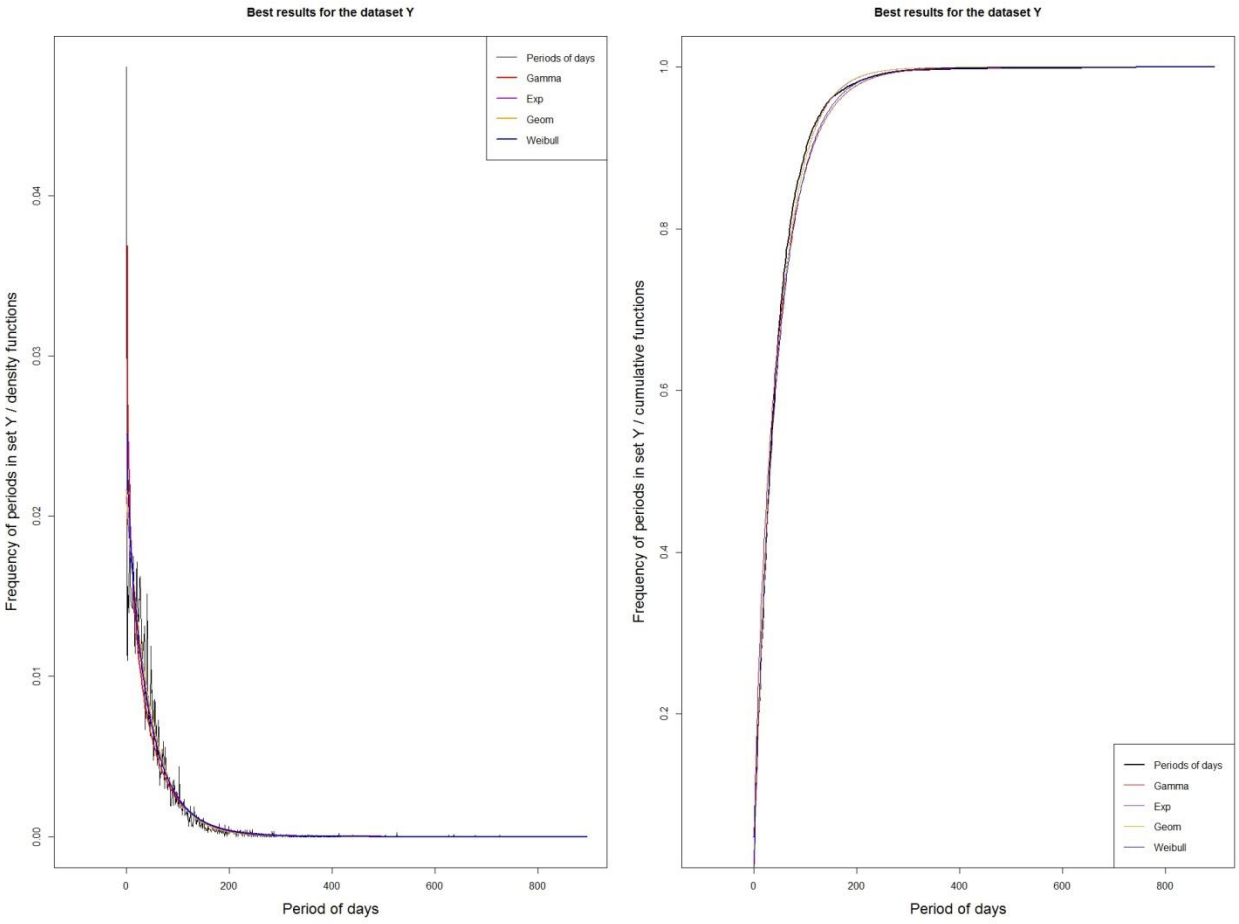


Figure 14: Frequencies and density/cumulative functions of set Y

For the sake of choosing the best distribution among the best from above a computation of root mean square errors comparing theoretical and actual densities and probabilities was done. Based on the comparison of root mean square errors of the 4 most fitting optimal distributions for each dataset the following decision results:

Dataset V: deviations from density values for $\mathrm{Geom_V}$, $\mathrm{Exp_Y}$, $\mathrm{W_V}$ and Γ_V are 28.72679, 28.72265, 28.68128 and 28.63782 respectively. Deviations from the cumulative probability values are 0.1598573, 0.1770967, 0.3170949 and 0.6566548 respectively. With the consideration that our data is discrete, we choose $\mathrm{Geom_V}$ as the distribution of this dataset.

Dataset X: deviations from density values for $Geom_X$, Exp_X, W_X and Γ_X for our data are 28.67523, 28.67109, 28.65170 and 28.87623 respectively. Deviations from the cumulative probability values are 0.2258340, 0.2112228, 0.2624345 and 1.0093686 respectively. We choose Exp_X as the distribution of this dataset.

Dataset Y: deviations from density values for $\mathrm{Geom_V}$, $\mathrm{Exp_Y}$, $\mathrm{W_V}$ and Γ_V are 28.71307, 28.70895, 28.65173 and 28.66138 respectively. Deviations from the cumulative probability values are

0.1721282, 0.1959139, 0.2645674 and 0.3741074 respectively. With the consideration that our data is discrete, we choose Geom_V as the distribution of this dataset.

Since the differences in the errors are so small and keeping in regard the stochastic nature of variables, we chose the geometric distributions for our data. We also decided that, knowing our data may be faulty in the negative values, the dataset is distributed as $\mathrm{Geom}(0.0206)$.

This means that the mean, median and standard deviation estimators of our data amount to 47.67309 days, 33 days and 48.17049 days respectively. Since we are talking about discrete variables these values represent 48, 33 and 48 days, respectively.

From here on the methodologies differ:

- we observe the number of all vacancy advertisements present on a given day or in a given month,
- we observe the number of new vacancy advertisements published on a given day or in a given month.

7.3.4 Methodology of enterprise webpages data

There are currently close to 200,000 enterprises present in Slovenia, but of course not all of them have their own domain. Some do not use the internet at all, some are present only on social media sites, and some advertise their job vacancies only through the Employment Service of Slovenia or job portals. In 2017 we have conducted an ad-hoc evaluation survey to understand what channels enterprises use to advertise their job vacancies. Together with information from an online enterprise register specialized in Slovenian enterprises we have collected a set of domain names from 27,000 enterprises. Every year we select a new sample for the Job Vacancies Survey (JVS). For the scraping projects we select those enterprises that have their domain addresses known and create a scraping sample. The results that we will present in this paper are calculated from the sample in 2017 which consisted of around 2,500 enterprise domains. In the year 2018 we changed the sample, which is now consisted of the full number of small, middle and big enterprises (accounting for around 6,000 domains) and 2,000-2,500 micro-enterprises present in the JVS sample. This gives us a sample of 6500 enterprise domains to be crawled and scraped. In both samples the names and register numbers of companies are known, and therefore we do not have any problems with linkage for these data.

Scraping of enterprise webpages is executed in such a way that the least amount of webpages is actually scraped as to shorten the executing time and also lessen the traffic burden on internet domains. The first step is a crawl of the domain webpages for subpages that include keywords related to job vacancies in their addresses, such as ‘new employees’, ‘career’, ‘co-worker’ etc. The selected subpages are then scraped as a set of possible job vacancy sites. Their contents are saved to a corpus on which estimation with machine learning and statistical learning methods takes place.

First some cleaning of the corpus is carried out: each site gets its textual content vectorised and is then subjected to a logistic regression classification that classifies the sites as containing or missing job advertisements. Afterwards positive subpages (those containing job advertisements) go through a combination of a linear regression and AdaBoost method where the number of job advertisements present is estimated.

The logistic regression method was trained on a corpus of 1,000 documents, which were pre-processed beforehand. The documents were transformed into the lower case and contained URLs were removed. Afterwards the texts were tokenized with separation on every whitespace, and then cleaned of special characters (punctuation marks, dashes, brackets, ampersand, dollar signs...). The resulting documents had their text filtered according to the number of occurrences of each word. The top and bottom 10% of occurrences were eliminated, as they were either stopwords, or did not carry enough information being too specific. In the end the 10,000 most used tokens were selected to be used in the training set of the logistic regression.

The same pre-process took place again on all positive subpages to train the AdaBoost and linear regression methods. Both methods are trained to only consider content in Slovenian and job advertisements that do not include student work or work in a foreign country. It must be said that this training set contains only 175 observations and can admittedly be improved. The main issue to be enhanced is the representation of subpages with a high number of job advertisements. Before we do this, a more comprehensive study of representativeness of job vacancy pages has to be carried out.

After job vacancy estimation a deduplication process takes place. Deduplication is carried out on 3 auxiliary datasets. All positive texts in the recent corpus are compared for duplicates by comparing distances of vectorised corpus rows, first for identical rows and then for 'more than similar' rows (distance threshold is 4). The positive texts are also compared to the positive texts of the previous scraping period. The identical double distance calculation is executed to find duplicates from the previous period.

7.3.5 Joining the sources

After the deduplication – linkage process the data sets are joined together. Another deduplication step takes place, this time filtering out the jobs that are linked to the same register number. The actual number of job vacancies is selected with a priority method: when multiple sources include job vacancies with the same register number, only one is taken into account according to the following priorities:

1. if one of the sources is the ESS, we select the number of vacancies detected from this source,
2. if the ESS source is absent, the job portal data takes precedence before the enterprise websites data,
3. finally, if only enterprise website data is present for a register number, we consider this source as the right one.

We end the process with a table of register numbers of enterprises in the survey sample and their detected job vacancies. Also attached to the data are the results of the correspondent job vacancy survey.

7.4 Statistical Outputs

After the deduplication – linkage process the data sets are joined together. Another deduplication step takes place, this time filtering out the jobs that are linked to the same register number. The

actual number of job vacancies is selected with a priority method: when multiple sources include job vacancies with the same register number, only one is taken into account according to the following priorities:

4. if one of the sources is the ESS, we select the number of vacancies detected from this source,
5. if the ESS source is absent, the job portal data takes precedence before the enterprise websites data,
6. finally, if only enterprise website data is present for a register number, we consider this source as the right one.

Depending on the methodology used we can output 4 different statistics. With the combination of administrative sources, job portal data and enterprise webpages data and using the above results for distribution fitting, we can calculate an estimate of **available** online job vacancy advertisements present on the reference day or in the reference month. Furthermore, our deduplication process can be used to estimate the number of **newly posted** online job vacancies in these periods.

Due to the difficulty of extracting data from unstructured text, the issue of the number of vacancies in a job vacancy advertisement was not yet solved for job portals data. In the following definitions we use the assumption of perfect representation: one advertisement represents exactly one job vacancy.

Definitions

Online job vacancies are defined as job advertisements that were published on job portals, on internet pages of private companies or on the state employment agency job portal.

Available online job vacancies on the reference day are defined as posts which were on the reference day still available on the internet. The advertisements were published online on the reference day or before.

Available online job vacancies in a reference month are defined as posts which were available on the internet in the reference month. The advertisements were published online in the reference month or before.

Newly posted online job vacancies on the reference day are defined as posts which were published on the internet on the reference day. This indicator represents the number of daily posted advertisements.

Newly posted online job vacancies in a reference month are defined as posts which were published on the internet in the reference month. It includes all posted advertisements in the month no matter if some posts are still open at the end of the month.

Online job vacancies followed by JV regulation are defined as posts which have been newly created, are unoccupied, are posts for which the employer is actively seeking suitable candidates outside the enterprise and which will be filled immediately or in the near future (according to the above assessment of the time gap between the date of publication of the advertisement and the employment). The number of online job vacancies followed by regulation is calculated as online job vacancies on the reference day, summed to the number of published job vacancies in the past but no longer present on the internet that are still in the range corresponding to the average time gap between published advertisement and employment.

Alternatively these statistics can be calculated taking into account the job vacancy advertisement regulations – the vacancy can be filled three days after being published in the earliest.

Results

Our ongoing work right now is to understand the nature of advertising of Slovenian enterprises. Our sample is composed of the enterprises featured in the ad-hoc evaluation survey. Linking the results of the survey with our administrative sources and scraped data shaped the below results. In total 8,885 enterprises' answers and collected data were compared in August and November. In addition both portals and a sample of 2500 enterprise domains were scraped for job vacancies and joined to this data.

Stated in the Table 18 are the numbers of advertisements/vacancies received from administrative sources, scraped from job portals before deduplication and scraped directly from in the last two scraping periods (3nd and 4rd quarter of 2017, namely on 31st August and 30th November). Note that in the case of job portal data we assume that one advertisement corresponds to one job vacancy. Keep in mind that the data is not yet deduplicated. A lot of vacancies are present for multiple weeks and/or on both portals and many of the vacancies are for either student work or work abroad. Also note that estimation of the number of job vacancies on enterprise websites is done using an averaged ensemble of machine learning methods.

Table 18: Number of detected job vacancy advertisements in two periods (with cross-source duplicates): 28th August and 30th November 2017

	Number of advertisements detected		Number of enterprises detected	
	28th August	30th November	28th August	30th November
ESS	5,416	5,622	1,118	1,162
Job portals	13,861	15,405	1,185	1,394
Enterprise websites	325.5	576	82	252

After deduplication within each source, between periods and between all three sources and including only those enterprises that are present in the evaluation survey sample the numbers look like (see Table 19):

Table 19: Number of detected job vacancy advertisements in two periods (deduplicated)

	Number of detected advertisements		Number of enterprises detected	
	28th August	30th November	28th August	30th November
ESS	5,416	5,622	1,118	1,162
Job portals	1,210 (and 325 on agency sites)	1219 (and 107 on agency sites)	308 (and 6 agencies)	292 (and 4 agencies)
Enterprise websites	142	349	43	29
Reported JV/Reporting enterprises	6,849	6,327	1,720	1,641

We can see that the amount of advertisements with duplicates/ for student work/for work abroad is quite big across all three sources. Especially the number of job vacancy advertisements from job portals is reduced for a factor of four and more. The drop of the number of job vacancies from job

portal, from 13,861 to 1,210 is quite big, but only those data can be used for more detailed analysis (by activity, by location of work…). Comparing these results to the results of the job vacancy surveys from the same time periods we see that we were not quite able to capture every job vacancy reported. We can see in the last row of the second table that we lack some enterprises, and therefore more advertisements.

Some other experimental statistics follow calculated from job portal data:

Table 20: Available online job vacancies on the reference days

	31st August	30th November
Number of advertisements (no duplicates)	1,265	1,203
Number of same advertisements present on both portals	103	82
Total number of detected advertisements	1,368	1,285

Table 21: Available online job vacancies in the reference months

	31st August	30th November
Number of advertisements (no duplicates)	3,319	4,207
Number of same advertisements present on both portals	223	286
Total number of detected advertisements	3,542	4,493

Table 22: Newly posted online job vacancies on the reference days

	31st August	30th November
Number of advertisements (no duplicates)	116	74
Number of same advertisements present on both portals	7	2
Total number of detected advertisements	123	76

Table 23: Newly posted online job vacancies in the reference months

	31st August	30th November
Number of advertisements (with no duplicates)	1,854	1,991
Number of same advertisements present on both portals	130	124
Total number of detected advertisements	1,984	2,115

We can see that new vacancies represent only about a half of all monthly available vacancies, which gives some credence to the above calculation of the time gap between the publishing date and employment. We can also observe how less than a tenth of all jobs available on the last day of either reference month are constituted by new vacancies.

Annex G: Sweden

Summary of recent work

Since the summer 2017, Statistics Sweden has continued to work on the job portal data from the Swedish Employment Agency and comparing with the data from the Job Vacancy Survey. The work includes:

- Extending the data collection period from the Employment Agency from five to ten years. The data now cover the period 1 January 2017 to 31 May 2017;
- Cleaning of the data;
- Local unit identification;
- Time series analysis of data from the Employment Agency and the Job Vacancy Survey;
- Continued downloading of data from the private job portals Jobbsafari and Metrojobb.

7.1 Introduction

The legal basis for Statistics Sweden to do web scraping is not clear and we have only carried out tests on a small scale on websites of the public sector. No data has yet been captured using web scraping on private company web sites. Statistics Sweden thus at an early stage of the ESSnet decided to concentrate on job portals, and in particular data from the portal of the Swedish Employment Agency (SEA) and Platsbanken (PB). SEA is a main player on the labor market in Sweden and they provide us with data directly from their data warehouse. The approach taken is to explore the PB data as a starting point to understand job portal data in general, and to evaluate the quality of the PB data.

An inventory and assessment of job portals in participating countries was delivered by work package 1 in 2016. Following the inventory, Statistics Sweden selected four job portals out of twelve, based on the size (the number of advertisements as of June 12, 2016) and the type, i.e. they are general job portals as opposed to specific job portals concentrating on certain professions. In the end, three of them provided us with data.

Platsbanken (PB) is the largest job portal in Sweden and contains a rich amount of data.

Metrojobb is the second largest job portal in our assessment. It is a private hybrid job portal combining original job offers with scraping of company web sites and other job portals.

Jobbsafari is a private hybrid job portal and the third largest in our assessment. Jobbsafari focuses on quickly identifying new advertisements, but data contain fewer variables compared to PB and Metrojobb.

The preliminary results show that the quality of the data varies a lot between the different portals; their data could be useful for different purposes. With the PB data, we tested algorithms for identifying professions (using text mining). We can link data to the Business Register since unique organization identification and the visiting address are available in the data, and we can link to the sample used in the Job Vacancy Survey, at the local unit level.

Jobbsafari has the ambition to detect as many advertisements on Internet in real time as possible. Our hypothesis is that their data has the highest coverage of job openings, but the quality of the information, in the aspect of available variables, is lower. So far they provided cleaned data that make the matching to other datasets difficult to interpret.

Metrojobb has a large portion of advertisements from outsourcing companies and many advertisements stay online for quite a long period, so-called ghost advertisement[65]. They generate noise in the online data.

Our approach has been to consider PB as the main source and use the other two sources as a complement for various purposes.

7.2 Data access

An xml file covering the additional period from 1 January 2007 to 31 December 2011 was extracted from the data warehouse of the SEA. New variables (visiting address and municipality) were included in the dataset. The data types were checked and data were added to the MS SQL server.

We use API for downloading data from the private job portals, Metrojobb and Jobbsafari. The data quality and the number of variables has not changed during the data collection period. Data are downloaded daily from an open network and saved in xml file format, and added to the MS SQL server when needed.

An overview of the amount of collected data is shown in Table 1.

Table 1. Number of collected advertisements

Portal	Period	Number of collected advertisements
Swedish Employment Agency	1 Jan 2007 - 31 May 2017	4 115 875
Jobbsafari	1 Apr 2017 - 31 Aug 2017	534 991
Metrojobb	1 Mar2017 – 9 Jun 2017	86 501

The main variables collected and a preliminary quality review is shown in Table 2.

Table 2. Main variables and rate of valid values

Variabels	Rate of valid values (without obvious errors), %			
	PB	Metrojobb	Jobbsafari	Comments
Title of advertisement	100	100	100	
Text of advertisement	~100	~99	~74 (snippet)	Jobbsafari only presents part of the advertisement

[65] Advertisements that do not indicate any current open job opening and have stayed online without proper management. They are not within the target population

Posting date	100	100	100	
Municipality	99	~98	~87	Partially comparable due to different standards
Code for profession	100	~100	100	Not comparable due to different standards
Ending date	100	100		
Scraping date	-	-	100	
Employer name	100	100	-	Not comparable with Business Register
Number of job openings in advertisement	100	-	-	
Visiting address	87	-	-	
Organization id	~100	-	-	
Data source	-	100	100	Url for web page where data was scraped

Judging from the preliminary quality review of the portal variables, the SEA provides the best data in terms of useful variables and the basic quality of the variables. The data are structured and contain 24 variables, e.g. date posted on the web, date for final application, organization identification number, name of employer, visiting address, municipality, job category, occupation, and the complete text of the advertisement. The important variables, e.g. posting date, employer name and municipality, have very few missing values.

The biggest problem with the Metrojobb and Jobbsafari data is that their data cannot be directly linked to the employer since they do not collect the identification of the organization. Additional important variables such as the name of the employer and the address are also missing, making it almost impossible to identify the organization.

In the following, we concentrate on the PB data.

The posting date for the advertisement is given by the user, or added by the system if the user does not give it. This guarantees good quality for the variable. The variable Ending date is also given by the user, but the system will add an invalid date if the user does not give it. We calculate the number of days an advertisement is on line by subtracting the ending date from the posting date. The questionable quality of the ending date will cause extreme values that we discard in the analysis of the advertisement days online. 74% of the advertisements are online between 14 and 90 days. The median is 28 days.

All the 290 municipalities in Sweden are represented in PB. The analysis of the municipality shows that advertisements are mostly from the large cities. Stockholm, Gothenburg, Malmö, and Uppsala

are the most featured municipalities. The largest share of the advertisements pertains to the Stockholm region, which is reasonable.

The number of job openings in an advertisement is a unique variable in the PB data. 21% of the advertisements include more than one job opening. The median is 29 and the mean is 91. However, it is not clear how reliable this figure is, since we find extreme big numbers of job openings in both the public and the private sector. During one quarter, some employers post more than 50 positions in the public sector, and more than 500 positions in the private sector. We are not sure how to treat these values. It could be an indication of recruiting companies that post ghost advertisements.

The organization id is very useful for linking to the Business Register (BR). It allows us to add the sector, the size of the organisation, and the local unit identification number.

Using the organization id, 99% of the advertisements find matches in the BR. Of the less than 1% that do not find a match, it is either because the organization identification is lacking in the PB, or there are no matches in the BR.

160 544 employers were identified in the data. Of the top ten employers, six are outsourcing companies, and their share of the advertisements is 7.3%. The remaining four enterprises are from the public sector. Their share of the advertisements is 3.2%.

PB data shows good coverage regarding sectors, regions and occupations. Both big and small companies use PB, and enterprises with different economical activities. The high skilled jobs are well covered in the PB data base, as showed by the occupation id and when comparing with main occupation categories. In general, low skill jobs seem to be less advertised on PB, compared to high skill jobs.

Organizations of different sizes are covered in the PB data. Based on the frequencies of the organization appearance, the yearly average share of the big companies (more than 199 employees) is 73%. Companies with 1-9 employees are 2% on average, 10-49 employees 8%, 50-99 employees 9% and 100-199 employees 8%.

7.3 Data Handling

De-duplication of the PB data was carried out at SEA before the data was delivered, but we still identified 3% duplicates, which is equal to 6% of the total job openings. We used the variables employer name, posting date, municipality, title of advertisement and text of advertisement for de-duplication.

The municipality of the employers can only be compared partially, since they do not always follow standard. Some are unofficial names of different kinds and some include postal code, and some are empty but the name is included in the address. In order to clean and standardize the municipality, a Python program was used. This corrected 15% of the incomparable municipality names, 75% remained unchanged.

Linking at work place level is necessary in order to infer variables such as sector, local unit identification and NACE code. These variables are necessary when comparing PB to the Job Vacancy Survey by subgroups.

The variable local unit identification is derived by linking to the BR using a combination of the organization id, the employer name, the visiting address and the municipality. Two string comparison

algorithms were applied, jaro-winkler and levenshtein distance. Jaro-winkler measures the minimum number of single-character transpositions required to change one word into the other. Levenshtein distance is the minimum number of single character edits (including insertions, deletions or substitutions). Since both algorithms consider the string orders (i.e. the order in which the variables are compared), we use two orders of the strings of the combination of variables. For example, a string from SEA in the origin order is “sanna stenhuggeri anderssons dödsbo leif erik göte stamvägen 81 43497 kungabacka” (string 1), which is a combination of the variables employer name and visiting address. When we order the string, a new string is created as ”43491 81 anderssons dödsbo erik göte kunsbacka leif sanna stenhuggeri stamvägen“, called string-ordered (string 2). This pair of strings is compared with the strings fetched from BR of the same organization id. The official legal name of the organization, the official name of the local unit, and the address of the local unit are combined in two orders, one in the variable order and the other in the string order. The example above has only one local unit, the correspondent strings are “anderssons dödsbo leif erik göte stamvägen 81 kungsbacka” (string 3), and the string-order “81 anderssons dödsbo erik göte kungsbacka leif stamvägen”. The string 1 is compared to the string 3 and the string 2 with string 4. Four scores are calculated and compared. In the end, the highest score of the comparison is taken as the correct matching. If there are one to many comparisons, i.e. the organization has many local units, the highest score is taken. If it is a one to one comparison, i.e. the organization has only one local unit, the highest score is recorded. The procedure is presented in Figure 1.

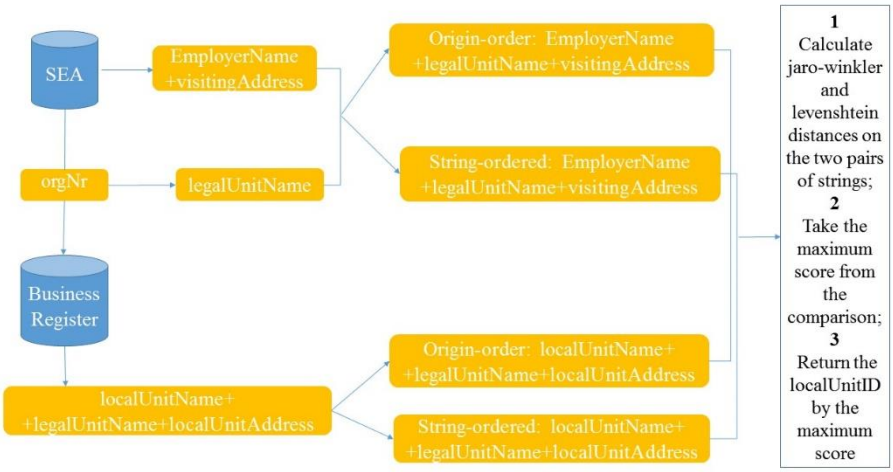


Figure 1. String comparison for matching PB with BR

The evaluation of the result is shown in Table 3. The average scores are high, although in the cases of minimum values the matched scores can be very low. This is because we did not use a threshold score; if the organization id from the PB is matched to the BR and it is one to one comparison, we will take the match as the correct local unit. In the future, we can set a reasonable threshold by further examining the semantic meaning of these scores. Low scores can be due to bad quality of the names used in PB data.

Table 3. Result of string comparison

Method	Max-value	Min-value	Average
Levenshtein	1	0,44	0,91
Levenshtein sorted	1	0,45	0,88
Jaro winkler	1	0,41	0,90
Jaro winkler sorted	1	0,41	0,85

Since Jobbsafari makes great efforts on detecting new job advertisements online almost in real time, we can use the sources in order to get an idea of the coverage of PB. The sources are the urls from which the advertisements were scraped. In total, 265 unique sources are identified in the Jobbsafari data. Figure 2 shows the url shares from Jobbsafari advertisements during April and August; 70% of the advertisements comes from PB. In the source, we also find recruiting companies such as academicwork and uniflex, and other sources like Facebook.

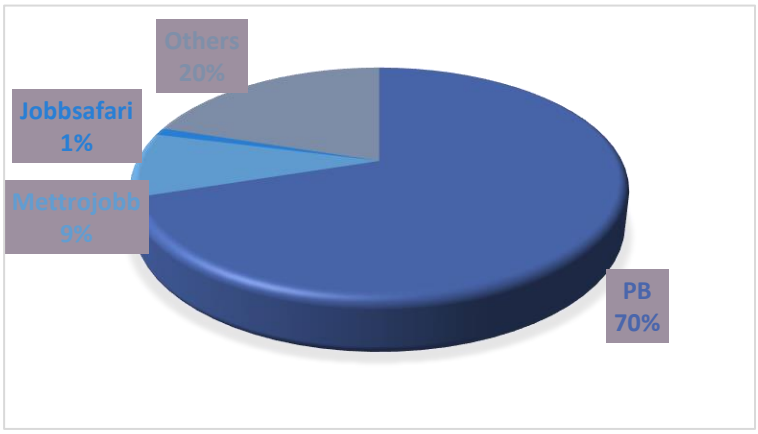


Figure 2. Origin of scraped advertisement in the Jobbsafari data.

7.4 Methodology: Time series analysis

In this section we illustrate a few major findings based on some (preliminary) empirical comparisons between compatible time series from PB and the Job Vacancy (JV) Survey. Our main hypothesis is that the PB data should represent the unknown population of the total number of job openings in the Swedish employment market, to a certain reliable level. Since the JV data represent the official statistics summarized by the estimates from the monthly survey, it would be rational to expect that the continuously assembled PB data have some additional information about the number of job openings. If this hypothesis holds, this extra information could be utilized in a modelling framework for different purposes.

Depending on the quality, we can theoretically think of several potential approaches to use the PB data. The most ambitious approach would be to use this alternative data source as a substitute for the survey. In order to do this we should have very precise information about the quality of PB data, which is practically unfeasible. Another option, which is technically achievable, is to use the PB data as a complement to the JV data in order to improve quality of the estimates. This option would imply some kind of correction of JV estimates. A combination of the two mentioned options (still achievable) might be to use the historical JV data and the latest PB data to generate forecasts for some periods when the estimates from the survey are not yet available. For these periods the survey would be omitted and the forecasts might then be evaluated, after the most recent survey data become available. This approach would still be feasible in a practical setting since the PB data are collected at almost daily frequency meaning that these data are available long before the survey is carried out. This fact implies a possibility to generate the estimates more quickly compared to an ordinary survey (flash estimates). The relevance of the last approach would strongly depend on the level of similarity with respect to the main properties of time series from the two respective sources.

7.4.1 Comparison of data

For the purpose of this analysis, both the PB and the JV data are structured as quarterly time series, starting with the first quarter 2007 and ending with the first quarter 2017. The JV time series are actually aggregates from the monthly estimates and the aggregation method is defined (in the survey documentation) as some kind of weighted average approach.

Here, we focus on the properties over time in order to check whether the PB series are suitable for any kind of joint modeling, together with the corresponding JV series. There is a large discrepancy between corresponding levels as shown in Figure 3a, although the properties of the two top time series are likely to have a lot in common. Both time series seem to have significant seasonal effects and the trends are likely to move together in the same direction, although not with the same rate of acceleration. The growth rates display striking similarity, especially after year 2012, as shown in Figure 3b.

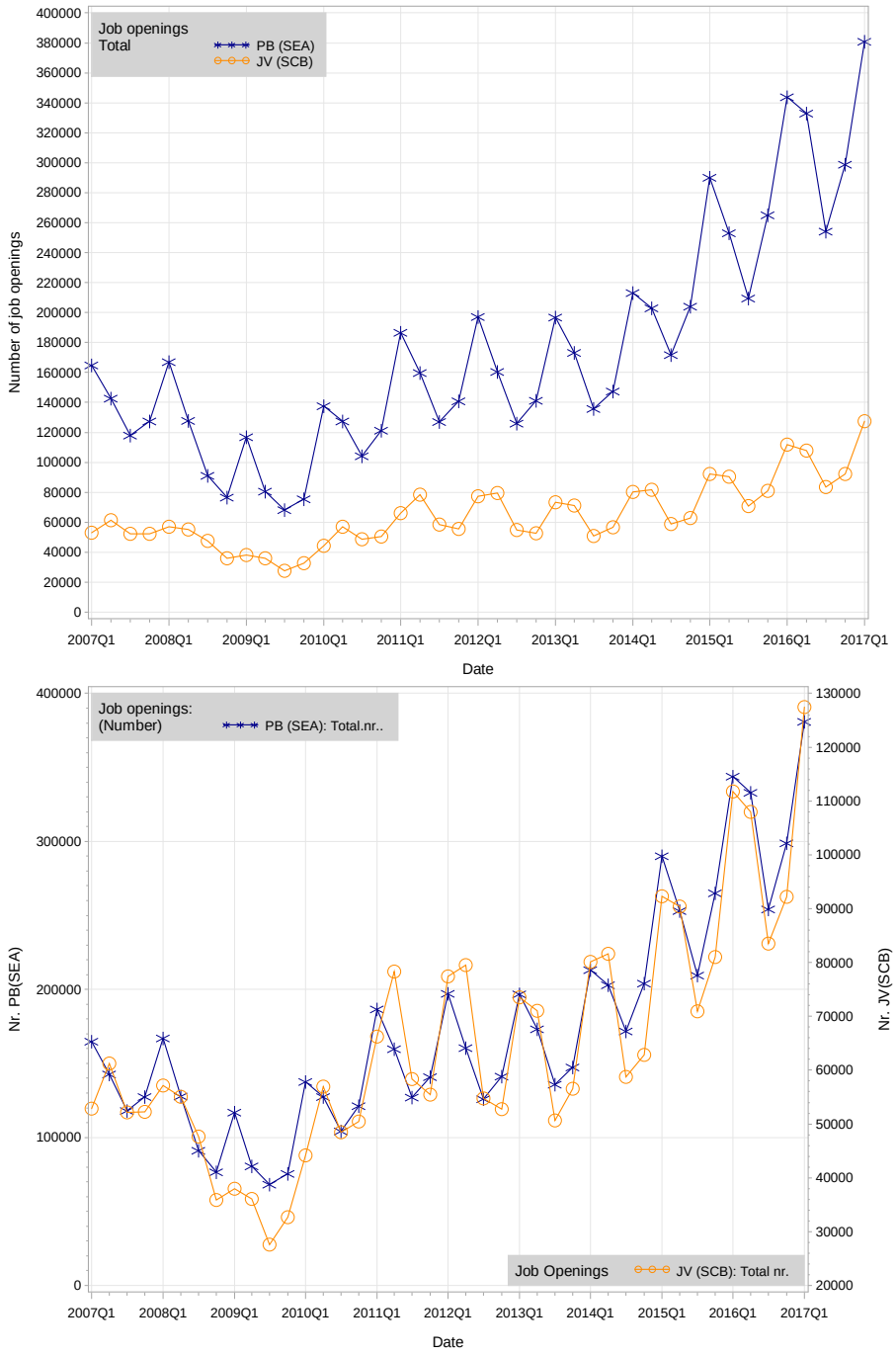


Figure 3a. PB (SEA) vs JV (SCB) – total number of job openings (Private sector and Public sector, jointly); quarterly values from 2007q1 to 2017q1; Upper diagram: Ordinary scale with values for both variables to left axis; Bottom diagram: left axis assigned to PB and right axis to JV.

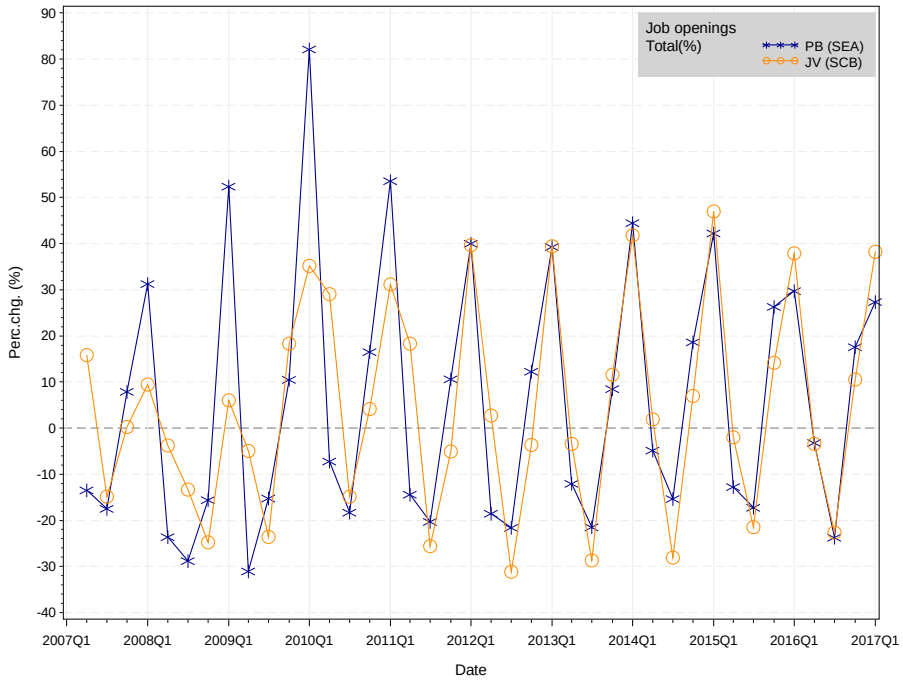


Figure 3b. PB (SEA) vs JV (SCB) – total number of job openings (Private sector and Public sector, jointly); Growth rates as percentage changes from the previous quarter; quarterly values from 2007q1 to 2017q1.

According to the official statistics (JV), the total number of job openings at quarter 1, 2017 was about 127 000 which was far less than the corresponding number of job openings aggregated from the PB data, which was approximately 381 000. Obviously, the differences in levels are persistent over time.

When the data is divided into the two main sectors, private and public, the dissimilarity between time series from the two sources is far more obvious in the private sector than in the public sector (see Figure 4). It seems that a large part of the discrepancy between the totals originates from the discrepancy between the corresponding private sectors. Also, the private sector is more interesting for analysis since it is much larger than the public sector. About 70 % of the total number of job openings originates from the private sector.

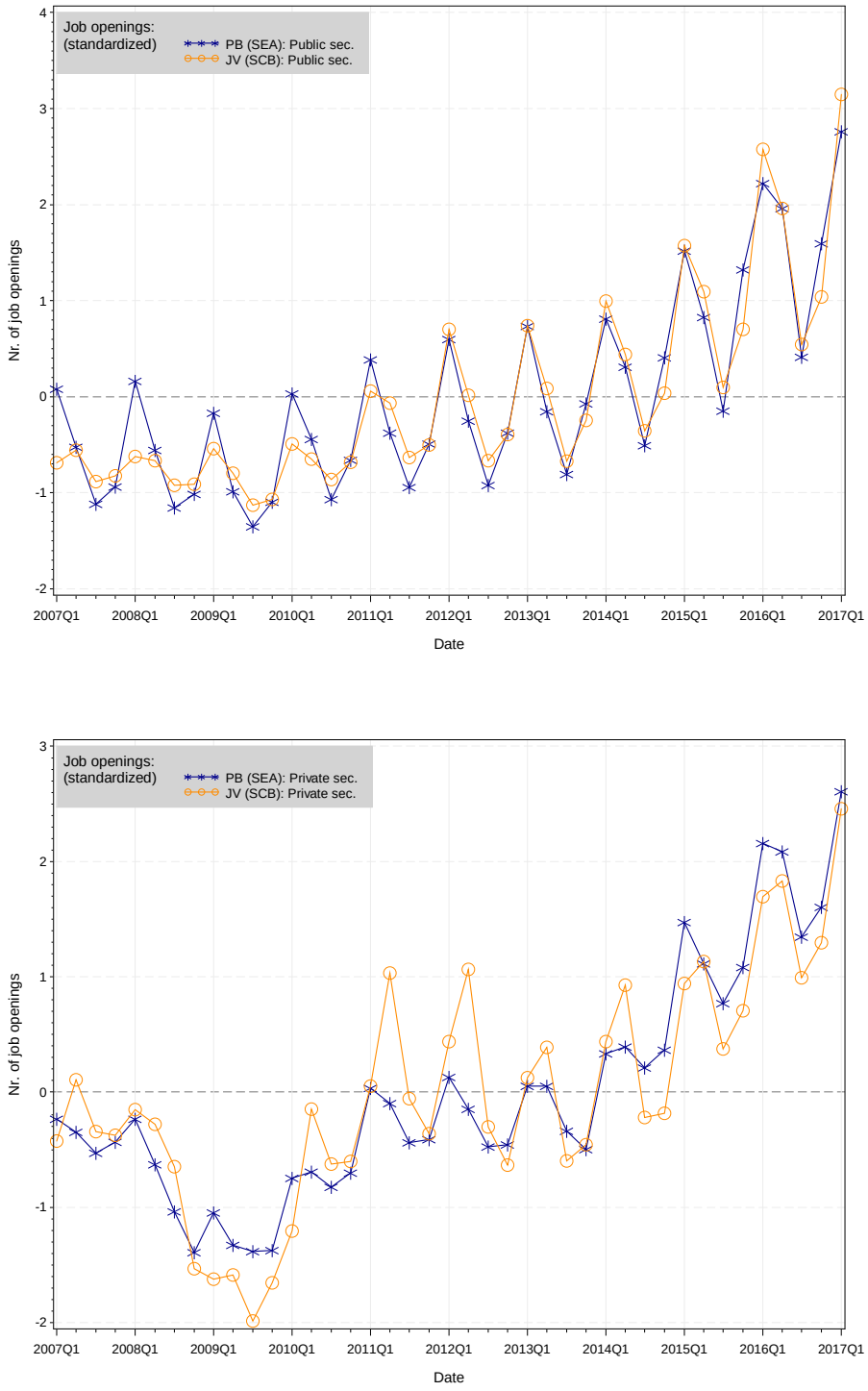


Figure 4 PB(SEA) vs JV(SCB) –number of job openings, standardized values (normal variates with mean 0 and std. dev. 1); Upper diagram: Public sector; Bottom diagram: Private sector; quarterly values from 2007q1 to 2017q1.

The data from private sector may also be studied at two more detailed dimensions, regional grouping and grouping according to NACE sections, since it is possible to perform matching between the two sources at these levels. The preliminary results indicate that the estimated numbers of job openings from the two competing data sources, for some NACE categories, are very close. The published values (JV) and the PB estimates from quarter 1 2017 are presented in Table 4. The confidence

interval for the JV point estimates covers the corresponding PB estimates for three categories. Since PB data are not collected according to any scientific method, the estimates have no confidence intervals.

Table 4. PB(SEA) vs JV(SCB) – number of job openings per NACE group; 2017 quarter 1; including 95% confidence limits for JV estimates.

Domain	NACE group	PB	JV	JV (95% LCI)	JV (95% UCI)
A	Agriculture, forestry and fishing	1295	0	0	0
B+C*	Mining, quarrying and manufacturing	12487	11916	11177	12655
D+E*	Electricity, gas, water supply, steam and air conditioning	996	1129	900	1358
F	Construction	5578	4298	3275	5321
G	Wholesale and retail trade; repair of motor wehicles and motorcycles	20767	13423	11071	15775
H	Transport and storage	12204	3918	3039	4797
I	Accomodation and food service activities communication	17221	5070	3000	7140
J*	Information and communication	8311	8337	7089	9585
K+L	Real estate, financial and insurance activities	4967	3878	3273	4483
M	Professional, scientific and technical activities	30516	9270	8176	10364
N	Administrative and support service activities	94214	12167	9464	15320
P+Q	Education, human health and social work activities	29716	6492	5280	7704
R+S	Art, entertainment, recreation and other service activities	20517	2874	2334	3414
NA	Non matching posts (not identifiable NACE)	494	0	0	0

* indicates domains in which PB- estimates are covered by the confidence interval for JV-estimates.

However, there are several categories where the PB estimates are quite close to the confidence limits and a few with very large discrepancies. For instance, the difference between PB and JV estimates in NACE section N (administrative and support activities, education, human health and social work) is about 80,000 job openings, which is approximately one fourth of the difference between the totals. Consequently, an in-depth investigation about the reasons behind the discrepancies in these sections is needed.

At the same time, the properties of the corresponding time series are likely to vary from group to group. Generally, when we have good matching between numbers of job openings from the two data sources there is also an acceptable resemblance in terms of properties of respective time series.

When we look at regional grouping, the corresponding levels at each subgroup are quite different in magnitude (PB numbers greater than JV numbers) which corresponds to what we find in the totals. Here we do not focus on regional grouping leaving this part for a future study.

7.4.2 Discussion about modeling

Here we briefly illustrate some ideas for potential use of alternative data by implementation of a plausible statistical model. The total numbers of job openings from the two sources may jointly be used as variables in each modeling framework.

We primarily think about a possibility to thin out the total number of surveys during a year in order to reduce costs and use forecasts instead. After the original surveys have been conducted and the up-to-date survey data become available, these data might then be used as benchmarks and revisions might be done afterwards.

- We could compare several time series models but the performance of the models can only be evaluated in relative terms. For instance, we might use the forecast errors as the main criteria and choose the best model accordingly. But what is good enough in terms of the forecast error?
- We could utilize information from different regions and NACE sections and implement a small area estimation model. However, this would require a certain level of reliability in each grouping category, which is not the case here.
- Since there is a good correspondence between the respective time series from some grouping categories, we could apply a bivariate time series model and make forecasts only for these categories.
    - Here we refer to the so called *cointegration* principles, meaning that the bivariate time series (from PB and JV, respectively) may share common trends which would suggest that the time series from PB might reflect the same phenomenon as the time series of interest (from PB). This aspect might be utilized as evaluation tool for different grouping categories of PB data.
    - One possibility is to perform seasonal adjustment and focus on trends.
- The subject matter specialists perform some kind of subjective correction for outliers in the JV data. We can think of doing something similar in PB data. This would probably reduce the discrepancies between the two data sources but the problem of finding relevant criteria arises here.

7.5 Lessons learned

Since the legal issues for web scraping for official statistics are unclear in Sweden, we were not able to take a general web scraping approach for data collection. We contacted job portals and reached agreements in the common interest of better official job vacancy statistics. Some lessons learned so far:

- In our experience, the contacts and discussion have been very open and there have been no difficulties to reach an agreement on data for testing purposes. Even so, it takes a lot of time and effort. Close communication is essential in order to confirm the interpretation of variables. It would for example be desirable to develop common methods for standardization of variables.
- Even with an agreement, it is essential to behave properly and follow ethical standards for the internet, and for example make it simple for the portal owners by using their available APIs.

- We find that our approach could be considered by NSIs as an alternative to investing in a web scraping system. Web scraping is complicated and expensive; however, an in house system may be a more stable data source. The private companies come and go, while the Internet remains as the communication channel and platform.
- From our preliminary empirical analysis we can conclude that the data from PB might be useful as a complement but not as a substitute to JV survey. Time series properties from both data sources appear to be very similar at higher aggregation levels and also for some NACE grouping categories. However, large discrepancies between levels from two data sources are still difficult to handle and deserve more attention in the future.

Annex H: United Kingdom

7.1 Introduction

In the SGA-2 phase, the UK pilot's main focus was on investigating the possibility of nowcasting JV estimates based on the online data. More specifically, the idea was to try implementing (e.g. machine learning) models, where the values from the JVS would serve as the target variable and the online data would feature as predictors.

Despite the relative simplicity of the nowcasting idea, it soon became apparent that its realisation will need to deal with several challenges. First of all, in order to build a nowcasting model, there was a need for high-quality datasets with a sufficient volume and time series of data. At the beginning of the SGA-2 we did not have access to any 3rd party data source[66] and it was not clear how much longer the ongoing negotiations will take. This led us to extend our web-scraping system with several new spiders, collecting JV counts from 4 more job portals. At the same time, we continued our efforts to create partnerships with 3rd party job portals and aggregators, such as Burning Glass, Adzuna and Indeed, which led to several collaborative arrangements and provisions of data that proved critical for the rest of the analysis. These collaborations, the acquired datasets and their properties are described in more detail in the **Data Access** section.

The second main challenge became apparent upon closer inspection of the data and the differences between the time series of JVS and the online sources. The data were first matched based on the company name to enable direct comparisons of the JV counts at the highest possible (company level) granularity. An interactive dashboard was created to provide comprehensive visualisations. This revealed wide differences and gaps not only between online sources and the JVS, but also between individual online sources. One source of the differences which was further investigated was a potential presence of a time-lag between online data and the JVS (e.g. stemming from definitional differences between a job ad and a job vacancy), or assumption of an average JV lifespan made when transforming a dataset representing flow of new vacancies into a stock of JV counts. These investigations, the methodology of matching and visualisations of the data are further described in the **Data Handling** section.

For the actual nowcasting, several ideas for a model were tried, implemented and compared with each other, as well as against the baseline persistence model (which simply nowcasts the previous JVS value). However, the noisiness of the data as well as the preliminary results indicated a need for filtering the data and working only within its subset where nowcasting was likely to be more accurate, albeit at the expense of a smaller sample size. Despite this, the improvements over the baseline model were only moderate. Similarly, classifying the current time series trend (into increasing/not-increasing) achieved accuracy of only close to 70%. The methodology of the nowcasting models, their evaluation and the reasons behind their mediocre performance are described in the **Methodology** section.

[66] We had access to Cedefop data from 2015-2016. However, this dataset did not contain a company name field, which was necessary for the subsequent analysis.

The nowcasted company-level estimates were subsequently scaled up to the level of total number of job vacancies, producing a nowcasted index. Similar indices were produced for individual industries with sufficient number of data points. Other than the index based on company nowcasts, the final **Statistical Outputs** section contains experimental plots of the counts by location as well as an index based on a time series model, a late addition to this report.

The structure of this report therefore follows the general approach set out for the ESSnet as a whole:

- Data Access
- Data Handling
- Methodology
- Statistical Outputs
- Future Perspectives

The main emphasis of this report is on the most recent developments since September 2017, although some aspects from the earlier investigations are included for completeness.

7.2 Data Access

As part of getting access to relevant datasets, we explored two principal avenues: web-scraping online job boards and arranging partnerships with 3rd party companies. In both cases, we were mainly interested in obtaining data representing a *stock* of job vacancy counts, as opposed to a flow of new vacancies, since the JVS itself is representing a stock of live vacancies at the collection date.

7.2.1 Web-scraping

We started to do web-scraping in early 2017, when we implemented spiders for scraping JV counts by company from 3 different job portals: Careerjet, CV-library and Universal Job Match. These spiders would scrape such JV counts for *all* companies found on the website, i.e. performing so called *full-size* scraping. In contrast, we also implemented several spiders for so called *sample-based* scraping which would scrape only the JV counts for a specifically built sample of companies[67]. Note that in both cases, we only scraped the JV *counts*, i.e. not the actual vacancies. This decision was made for two main reasons. First, scraping only the JV counts often greatly reduces the load imposed by the spiders on the website's servers (e.g. for Careerjet, only 26 requests needed to be made, one for each starting letter of the company name). Second, the UK's JVS only provides information about the JV counts, i.e. this was the only field where a direct comparison could be made.

The sample-based scraping was gradually extended to a sample of 150 companies, and spiders scraping the JV count information from these companies' websites were added. Finally, 4 more full-size scraping spiders were introduced at the end of September 2017.

[67] The sample-based spiders were extended one-by-one: with each new company added to the sample, an entry corresponding to the company would be manually identified at each portal. This meant that the retrieved data were already matched and JV counts corresponding to the same company could be directly compared, unlike in full-size scraping, where a matching algorithm needed to be run first to find matching company names.

All of our spiders have been running daily from within Google Cloud, with automatic triggers at midnight and storing the scraped data in a Mongo DB database. The scraping system was implemented with a mix of technologies including Python as the main language, Scrapy and Selenium for web-scraping and BeautifulSoup and regular expressions for extracting JV counts from the HTML.

7.2.2 Third party sources

Apart from the web-scraping, we pursued negotiations with several 3rd party companies to get access to their data, successfully forming a formal partnership with Burning Glass (analytics company with high quality job vacancy data) and Adzuna (one of the largest job vacancy aggregators and search engines in UK), as well as obtaining data from Cedefop (European Union agency working on pan-European web-scraping system) through an informal agreement. Finally, a small sample of (mainly aggregated) data was provided by a CEB (previously Wanted Analytics). It is worth to say that in all cases the partnership was arranged free of charge, focusing rather on collaboration or providing a service to the companies in return (e.g. validating their data against the JVS).

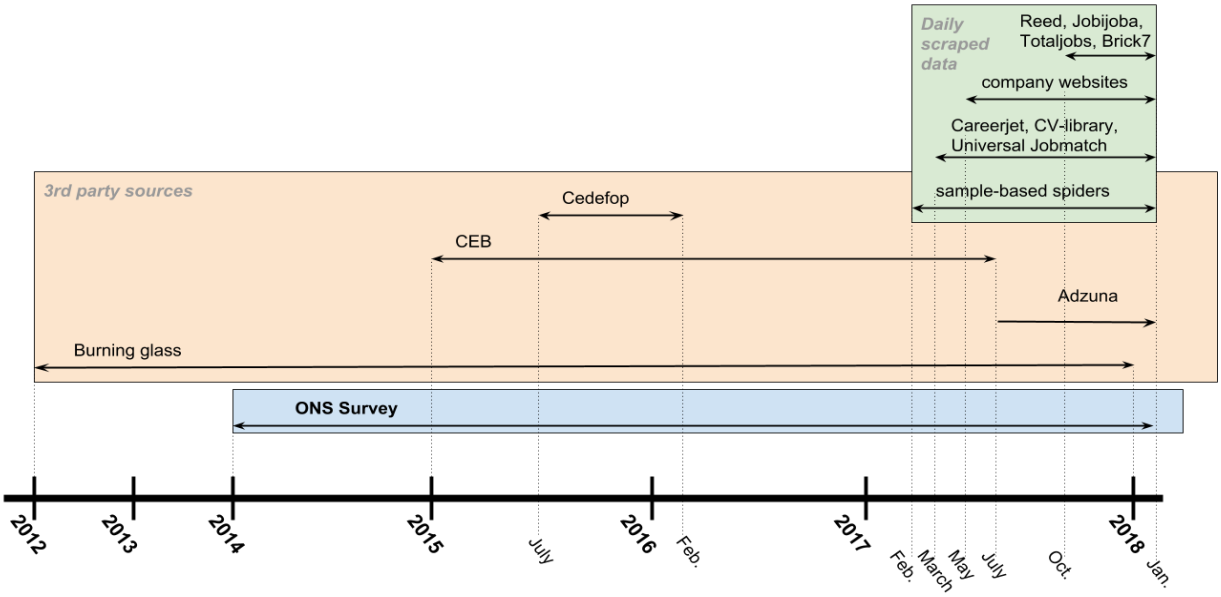


Figure 15: Timeline of our datasets

The main reason for our efforts to bring in data from 3rd party sources was a need for historic data, since the web-scraping can only usually get the current snapshot of the scraped website. Apart from that, there are also other benefits: the data from 3rd party sources are usually processed, clean and at a more granular level (individual job ads) then the web-scraped data, with presence of many more fields such as job title, industry or location, enabling richer analysis.

Figure 15 gives a timeline summary of the datasets we eventually worked with.

7.3 Data Handling

As mentioned already, in this project we were mainly working with JV counts. The desired format of the data was thus very simplistic and there was little need for extensive data processing (such as deduplication or text processing of job ad descriptions) and most of the processing were simple aggregations of job vacancy level data into JV counts. In this section, we will therefore focus on two main processing steps that are of more interest: flow to stock conversion and matching of the data by company name.

7.3.1 Flow to stock conversion

Not all the datasets were readily available to use. Burning Glass (BG), our largest and (subjectively) most important online dataset came in as a *flow* of new vacancies and first needed to be converted to represent stock measures. However, the BG data lacked information on the expiry date of the job ads, inevitably requiring some assumptions to be made about the lifespan of the individual job vacancies.

A simple way to approach this is to convert the data assuming the same lifespan for each job ad. This can be estimated from other datasets where this information is present, for example, in Cedefop data, we measured an average difference between expiry date and posting date to be ≈ 27 days. This assumption can be further split by industry (see Figure 16), although there is little way of measuring if this constitutes an improvement.

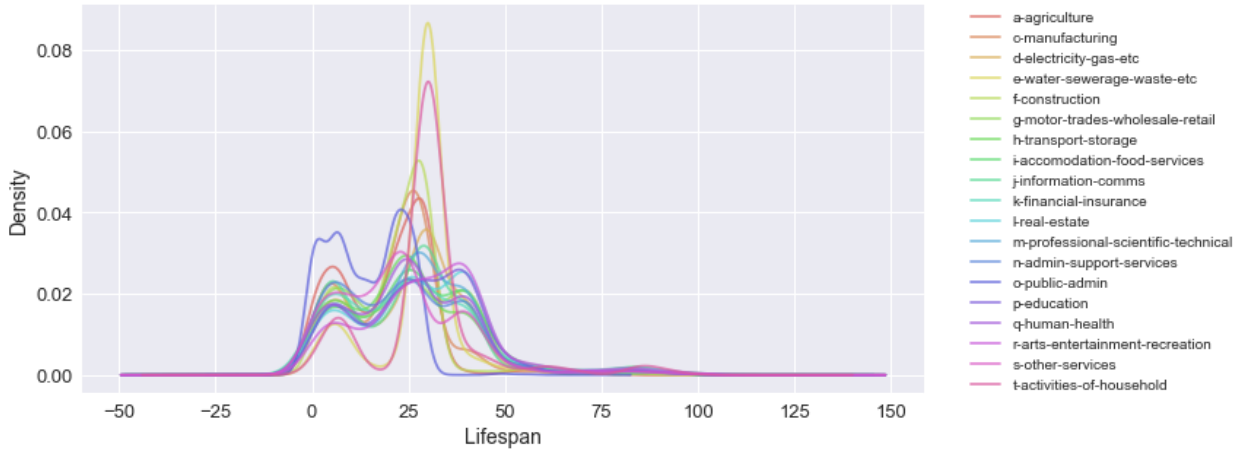


Figure 16: Densities of job ad lifespans by industry, Cedefop data

One way of validating this assumption was through an investigative experiment we carried out, looking at the trend alignment of time series for Burning Glass and JVS. The idea was to try different assumptions of job vacancy lifespan for the flow-stock conversion and see how the resulting time series aligns with JVS (on the total aggregation level). A quick interactive visualisation tool was built in Jupyter Notebook, with a slider for adjusting the assumed JV lifespan, enabling a visual inspection of the results. The time series were further compared by standardizing both series for zero mean and unit variance (as we were mainly interested in alignment of trends) and looking at the percentage of months with matching trends, as well as the Euclidean distance between the two curves, although. More sophisticated time series similarity measures may be better suited, e.g. to handle similar time series segments that are out of phase (Ding et al, 2008). Based on these metrics, an optimal vacancy

lifespan of 36 days was identified for Burning Glass data (Figure 17). Figure 18 shows the alignment after applying this assumption for the transformation.

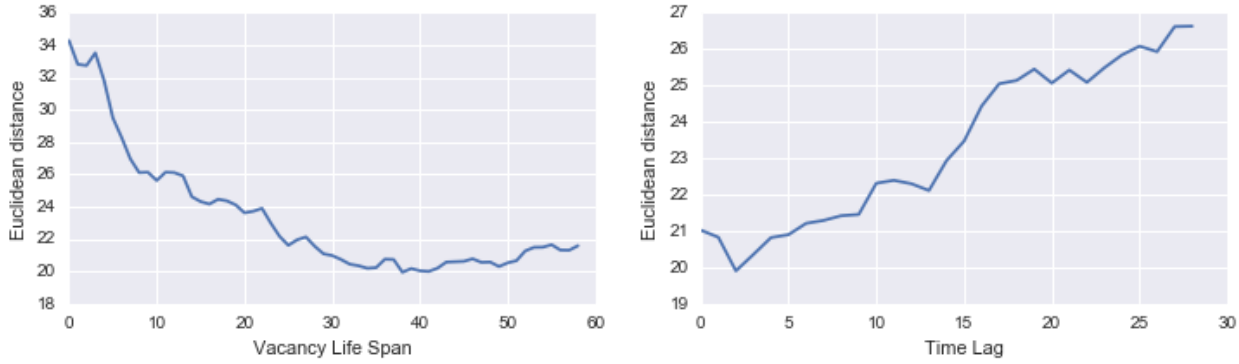


Figure 17: Euclidean distance between a standardized JVS and BG time series, after transforming the data with respective JV lifespan (left) or time-lag (right). Optimal values can be visually found at 36 days (lifespan) and 3 days (time-lag)

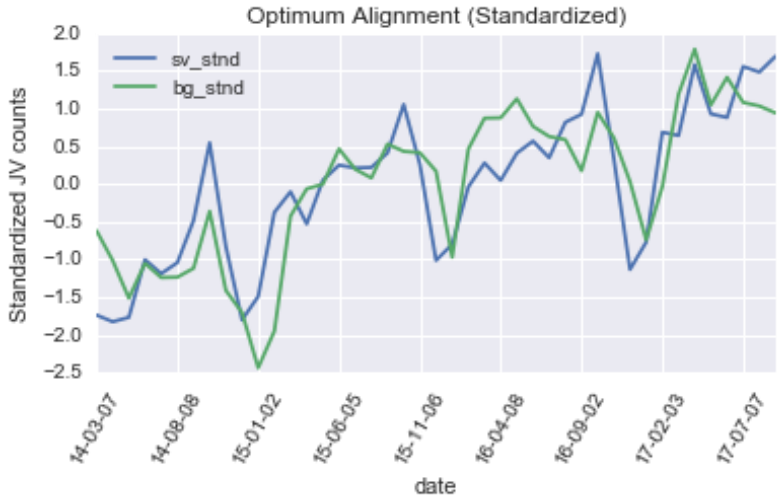


Figure 18: Comparison of the standardized time series for JVS and BG, after transforming BG with the assumed JV lifespan of 36 days and time-lag of 3 days

This analysis has further looked at finding a possible value for optimal *time-lag*, i.e. a shift of the online data in time to better match the JVS trend. There is a valid argument for this type of investigation due to the already present conceptual difference between a job vacancy and a job ad, with the latter likely to be published with a slight delay after the actual vacancy has been open[68]. Using a similar approach as described earlier, we identified an optimal time-lag of 3 days (Figure 17 and Figure 18).

Some concerns have been raised that applying multiple transformations to the data can lead to overfitting. Also, the fact that the reference dataset (JVS) is only produced at a monthly basis could

[68] It should be noted, though, that the JVS "measures the number of vacancies that are ***actively*** seeking recruits from outside organisations", making the time-difference gap between job vacancy (target concept) and job ad (measurable concept) narrower. See link below for more detail about the methodology of JVS https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/vacancysurveyqmi, retrieved on 8th May 2018.

introduce errors in the analysis, e.g. by misaligning the time series on peaks due to lack of granularity. For Burning Glass, we therefore only applied the necessary flow-to-stock transformation with a computed JV lifetime assumption and did not shift the data for the time-lag anymore.

7.3.2 Matching data on company name

The datasets representing stock measures in time were subsequently put in the same format and aggregated on 3 different levels: JV counts by company, by industry and the total JV counts. A programmatic data access API in Python was built to provide a standardized way of accessing the data, greatly facilitating and speeding up subsequent analysis and visualisations.

One of the first visualisations we did show simply the time series of the total JV counts by source (Figure 19) and already indicated several challenges to be dealt within the nowcasting model, especially the large differences between the online sources and the JVS.

Some of these differences are expected, as each source advertises vacancies only for a subset of companies, with the size of this subset greatly varying from source to source. Trends are more aligned, although even these often disagree between JVS and online data. Our next steps therefore aimed at comparing time series at a *company* level[69]. In order to do this, the data needed to be matched first.

A heuristical matching algorithm was implemented, looking at pairs of company name strings and computing a matching score. First, several transformations were applied on the strings (lower-casing, removing stop words…) to remove irrelevant terms and make them comparable. Next, a score was computed by adding up inverse document frequency (IDF, Salton and Buckley, 1988) values for words appearing in the intersection of the two word sets and subtracting IDF values for words in their difference. The score was normalized to a 0-1 range and few more tweaks were applied (e.g. penalizing too many stop words in the difference of the word sets). A pair was then declared to be a match if it passed a set threshold. This was empirically set to 0.9 after inspecting the matched data, with a preference to avoid false positives (incorrectly matched pairs), which could negatively impact the subsequent analysis. Figure 20 illustrates the described steps of the matching on an example.

[69] While the aggregated JVS time series have (relatively wide) confidence intervals, the company-level microdata should be more precise, as they represent a direct response of the surveyed businesses. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/vacancysurveyqmi#validation-and-quality-assurance

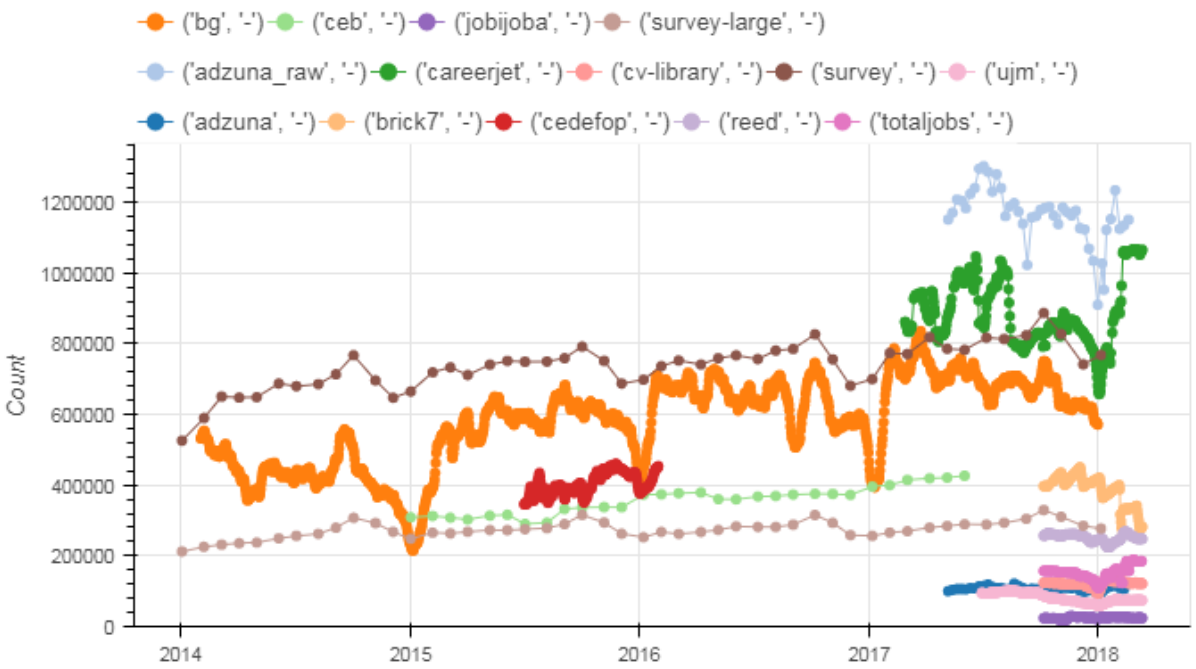


Figure 19: Time series of total JV counts by source

	Survey	Careerjet
Raw strings	FERRO UK SERVICES LTD INC FERRO EU ABCD UK BRANCH	Ferro Eu Abcd, Uk Branch
Lower-case	ferro uk services ltd inc ferro eu abcd uk branch	ferro eu abcd uk branch
Remove special chars	ferro uk services ltd inc ferro eu abcd uk branch	ferro eu abcd uk branch
Remove stop words	ferro ~~uk~~ services ~~ltd~~ ~~inc~~ ferro eu abcd ~~uk~~ branch	ferro eu abcd ~~uk~~ branch
Assign IDFs	0.9 ferro; 0.35 services; 0.9 ferro; 0.81 eu; 0.96 abcd; 0.8 branch	1.00 ferro; 0.90 eu; 0.96 abcd; 0.83 branch
Final score	**score = 0.934 = (green - red) / all** (also penalize stopwords)	
Is match?	✔	yes **[score > threshold (0.9) ?]** no ✔

Figure 20: Determining a match between two company name strings

About 51% of company names corresponding to the largest UK enterprises[70] were matched to some company name in Adzuna's data, 37% in case of Burning Glass and 32% for Careerjet. When including all entries found in the JVS in 2014-2017 period, about 7200 entries were matched to some company name in Adzuna's data, 6400 in Burning Glass data and 2700 in Careerjet data (with additional matches with the rest of the sources), providing for a solid sample to work with. For more details on the matching, refer also to the SGA-1 report (Swier et al, 2017), although the algorithm used here was slightly updated.

[70] Enterprises with more than 2500 employees, forming the first stratum of the stratified sample for JVS (usually ≈ 1300 entries)

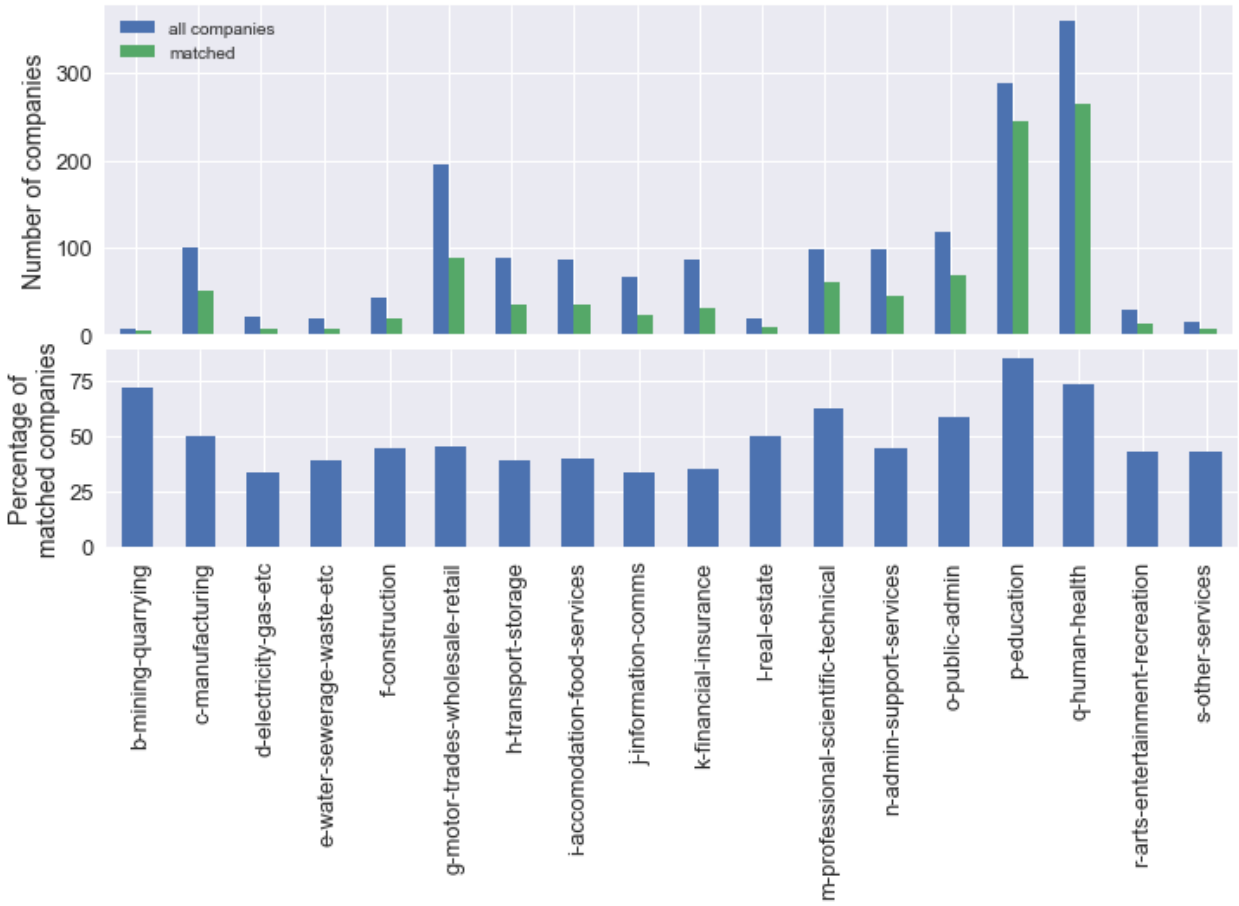


Figure 21: Industry distribution of the largest enterprises in the JVS and before and after matching, the matching further biasing towards the most populous sectors P and Q

The nature of matching (based solely on the company name string) caused also several qualitative issues. First, the likelihood of producing a successful match was favouring several industries, especially sectors P (education) and Q (human health, see Figure 21). This was of no surprise since the common entries in these sectors include universities or NHS trusts[71], which have a consistent naming and a transparent organisational structure. The least proportion of matches was produced for sectors J (information & communication technologies) and K (finance) where more complex enterprise structures can be expected.

The matched pairs also contained a few false positives, although manual inspection deemed the results satisfactory enough as an input for further analysis. The biggest issue was thus the largely reduced size of the matched dataset, especially its representativity of the original population and an introduced sampling error, biased towards the larger companies (only 15.2% of the companies with less than 2500 employees were matched, as opposed to 58.1% for those with more than 2500 employees).

Once the matching was determined, the data were joined on the company name to produce a *matched company-level data frame* (MCDF, see

[71] NHS Trust is "an organisation within the English NHS generally serving either a geographical area or a specialised function (such as an ambulance service)", see https://en.wikipedia.org/wiki/NHS_trust, retrieved on 8th May 2018.

Table 24), a single table with all JV counts for all sources, along with relevant extra fields (industry, number of employees), forming a base dataset for the nowcasting.

	Date	Survey	Careerjet	CV-	...	SICC	Other
...	**...**	...	...	...	...	...	...
TESCO STORES PLC LTD	**17-04-**	2345	1351	1525	...	G	56
TESCO STORES PLC LTD	**17-04-**	2357	N/A	1526	...	G	34
...	**...**	...	...	...	...	...	...
HSBC LIMITED INCL HSBC	**17-04-**	123	120	N/A	...	K	83
...	**...**	...	...	...	...	...	...

Table 24: Example of the format of the Matched company-level data frame (MCDF)

7.4 Methodology

The basic idea behind nowcasting is simple. Suppose we want to nowcast a JVS estimate for a company **C** at a time **T**. We choose a *window* of **M** months (we usually chose a window of 7 months). All the time series values for the online sources falling in this window (can) serve as predictors in the model. Additionally, the JVS estimates themselves for the company **C** can also be included as predictors, although only up to a certain time **T – D**, where **D** is a chosen *delay* representing the usual processing time of the JVS (e.g. in ONS, it takes up to 2 months to get the processed results from JVS since its collection date). Other fields can be used as predictors, including company attributes (number of employees, industry, …), date (e.g. month of the year) or even information from the company name (e.g. presence of certain keywords like “university”) etc. Finally, one may choose to nowcast only the trend of the JVS (increase or decrease against the previous value), transforming the problem from regression into a classification setting. Figure 22 summarizes the nowcasting idea.

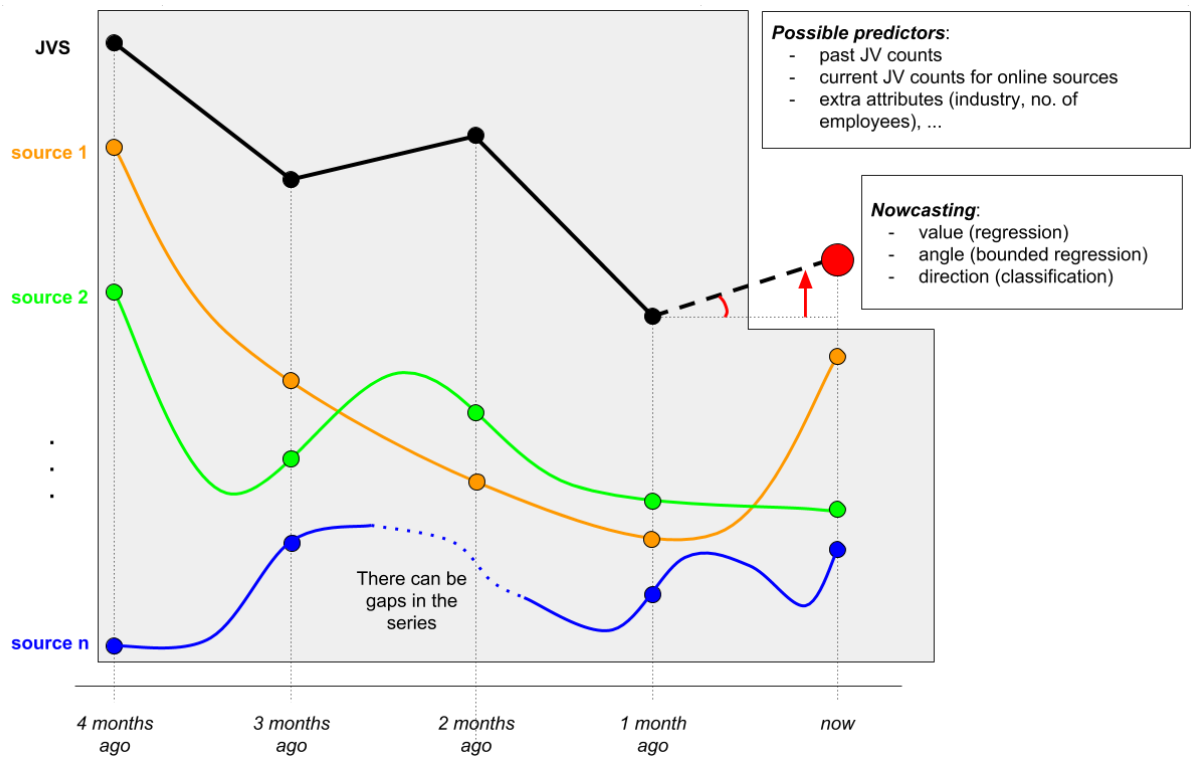


Figure 22: Nowcasting idea. A window of 4 months is chosen here. Note there can be missing values in the time series

7.4.1 Baseline model and metrics

As a baseline model to evaluate quality of the nowcasting models, we chose a *persistence* model. For a given company at a time **T**, the persistence would simply choose the JVS estimate at a time **T - D** (where **D** is the mentioned JVS processing delay).

To evaluate and compare performance of individual models, we were mainly looking at the differences between the true and the nowcasted values (nowcasting error). More specifically, we considered the following metrics:

- ***Mean absolute error*** (MAE), i.e. a mean of the mentioned differences
- ***Median absolute error*** (Median AE), i.e. a median of the differences. This provides an estimate of how well the nowcasting works in the usual case
- ***Root mean squared error*** (RMSE), i.e. a square root of the mean of squared differences, being more sensitive to larger nowcasting errors

In addition to the mentioned metrics, we considered the ***support*** of the model, i.e. the number of company/date combinations for which the model could output a nowcasted value.

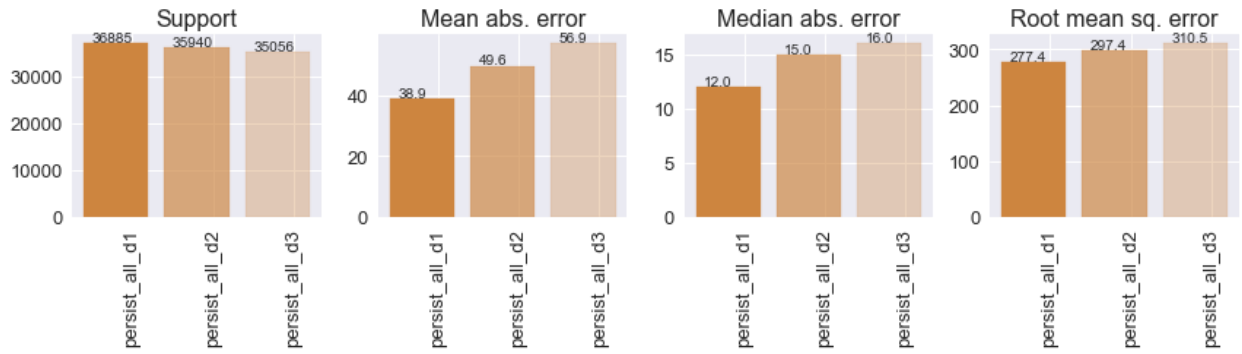


Figure 23: Metrics for the persistence models with survey processing delays of 1, 2 and 3 months (suffix d1, d2 and d3). As expected, as the delay increases, the persistence nowcasts get worse.

7.4.2 Simple linear regression idea

In mid-February 2018, we organized a hack-day within our Big Data team at ONS, with few more participating data-scientists from the rest of the office. The idea was to provide participants with the cleaned and processed data (such as the MCDF) and let them explore different ideas of nowcasting for duration of the day. At the end of the day, the teams presented their findings, with a variety of interesting approaches, including time series models, models based on neural networks and others.

One of the simplest ideas presented involved creating a linear regression (LR) model for every input entry (combination of company and date for which we want the nowcast). A single online source would be chosen whose values (during months prior to the nowcasted date) would feature as predictors. Corresponding JVS values would then represent the target variable. The linear regression model would be trained on these values and the nowcasted estimate obtained as a prediction of the model based on the current value of the online source. See Figure 24 for a demonstration of the idea.

The obvious question is which source should be picked to function as a predictor in the model. One way to answer this question is to choose the source which has the best correlation[72] with the JVS on the dates prior to the nowcasted point. A high correlation for a period of several months indicates that the source reliably follows the trend pattern of JVS and is thus a good candidate to base the current nowcast on. Furthermore, choosing the *best* source for given situation (rather than pre-selecting a specific source for all inputs) increases the resilience of the system which does not rely on a single data source but can substitute for the second-best option in case of unavailability.

[72] Other option is to choose the source with the highest number of month-to-month trends matching the JVS time series

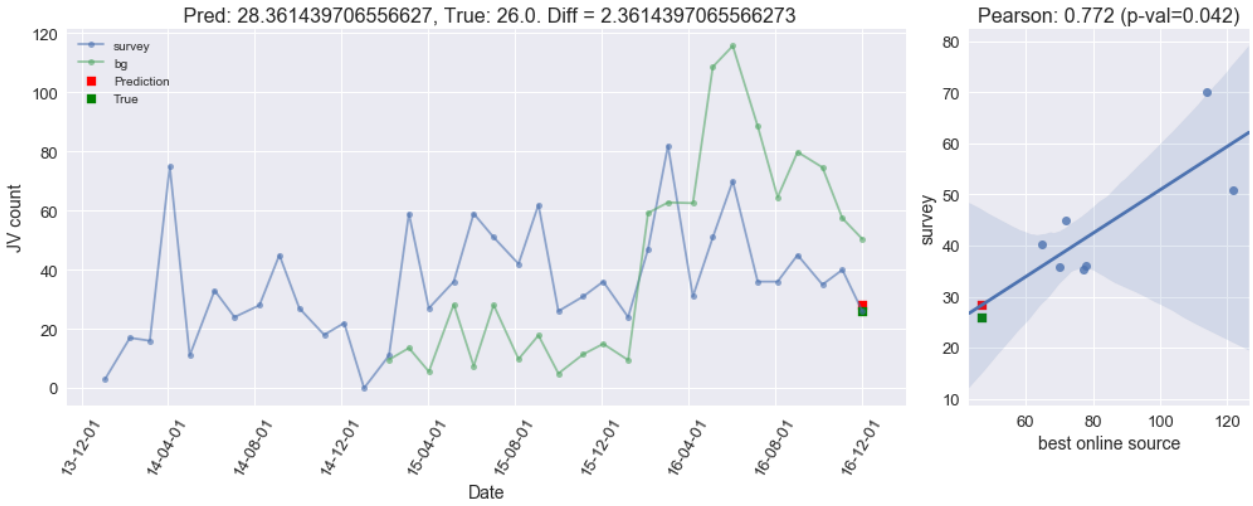


Figure 24: Nowcast based on the simple linear regression idea

Still, such model performed *worse* than the persistence while having much lower support (Figure 25). Upon closer inspection of the nowcasts, a need for pre-filtering data was identified, which we discuss in the next subsection.

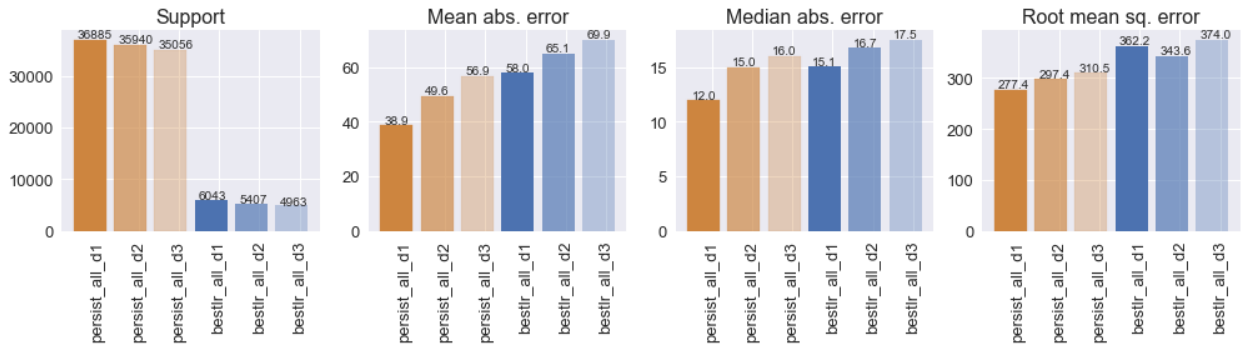


Figure 25: Linear regression model ("bestlr") compared to the baseline for 3 values of delay

7.4.3 Filtering

Figure 19 showed the large variance and differences between the JVS and the online sources on a total JV scale. Perhaps not surprisingly, the company-level time series often exhibit even greater oscillations, for which there may be various reasons (companies running huge one-off recruitment campaigns, change of a business structure…). Similarly, there are major scale differences between the online sources and the JVS at the company level. Again, there are many possible reasons for this, among others advertising through recruitment agencies, advertising job ads via offline channels or even incorrect matching from previous step. Figure 26 shows two cases where the differences seem prohibitively large to continue with nowcasting, showing the need for filtering the data.

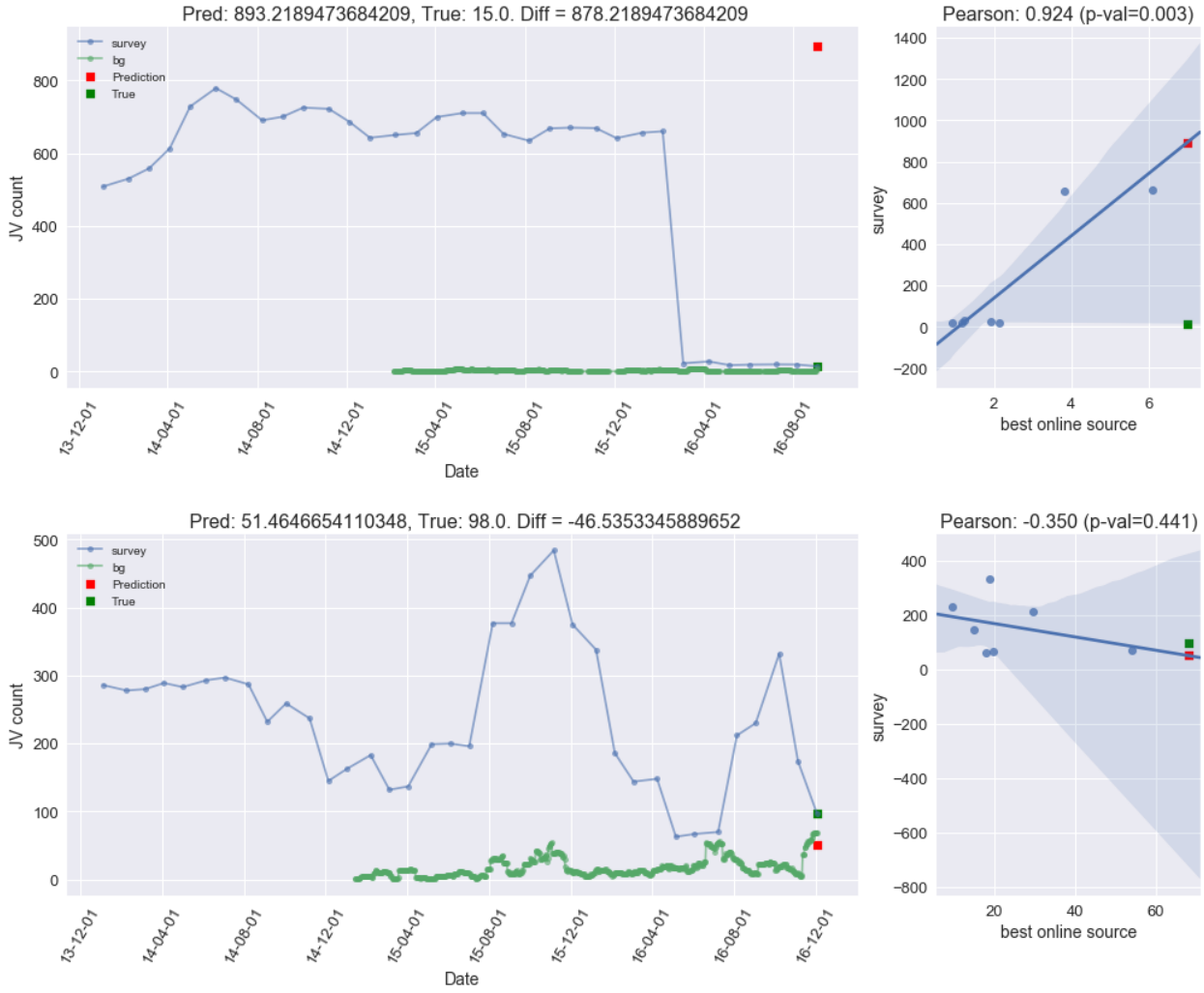


Figure 26: Cases that should be filtered. Top plot: due to the large differences between JVS and BG, even the smallest bump in BG causes the nowcasted value to spike in the linear regression model. Bottom plot: Despite the relatively accurate nowcast, the two sources display negative correlation on the months prior to the nowcasting date and thus nowcasting based on this source should be avoided.

A quick visualization dashboard was built in Jupyter Notebook to explore the benefits of the filtering. The key point here is to make a good compromise – while filtering data may improve the nowcasting accuracy, it also reduces the sample size and thus in turn the confidence in the nowcasted index. We ended up filtering the data in three ways, always only based on the data points prior to the nowcast date[73]:

- By **correlation** (Figure 27). Here a strange pattern was found, where the RMSE of the LR model actually increased with increasing threshold on correlation until the value of ≈0.7, after which the RMSE started sharply decreasing. To prevent filtering too much data, a minimum value of 0 was chosen as the threshold (effectively requiring a positive correlation)

[73] Or, in case of a survey processing delay D being greater than a month, on points up to time T – D (where T is the nowcast time)

- By **ratio of scales**. Here we essentially filter out cases where the ratio of scales between the JVS and the (best) online source is too large on the months prior to the nowcast date. A value of 3 was chosen based on visual inspection.
- By **ratio of variances**. Similarly, we imposed a restriction on the ratio between the variance of the JVS and the variance of the (best) online source. An upper threshold of 2 was empirically chosen here.

Figure 28 shows the improvement in nowcasting accuracy after with filtering. However, on the filtered data, the LR model still performs only comparably with the persistence, achieving very little improvement.

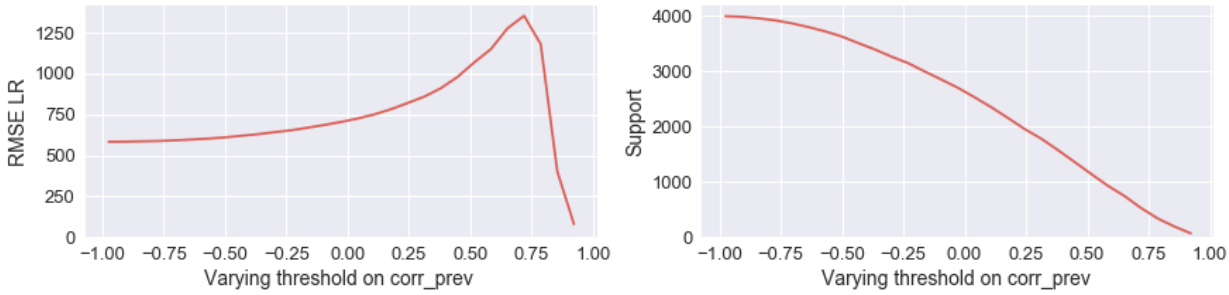


Figure 27: RMSE of the LR model when varying the lower threshold on correlation

Figure 28: Error metrics of the LR model after filtering. The "persist_all" refers to the Persistence models with all available data, with the rest of the bars showing the comparison against the Persistence model narrowed down to the sample after filtering.

7.4.4 Long short-term memory neural networks

Long short-term memory (LSTM) units are building blocks for recurrent neural networks layers, developed in 1997 (Hochreiter & Schmidhuber, 1997). Their applications are well suited for data with a temporal dimension and for problems where information from more distant past is useful when making a decision, in addition to the information from the current input and the immediate past.

In this experiment, we trained a neural network model with LSTM layers on our filtered dataset, with JVS estimate being the target variable. As for predictors, we took a window of 7 months (prior to the nowcast date) and the corresponding monthly JV counts for the best correlating source. We also included the previous JVS values in the predictors[74], as well as the month of the year[75]. A grid search with 5-fold cross validation was run to determine the optimal network architecture, early stopping used during training to prevent overfitting and the model checkpoints created to output the best seen model (as evaluated using a validation dataset).

Based on the outcome of the grid search, an architecture consisting of 2 LSTM layers with sigmoid activation function and 25 neurons each was chosen. An extra dense layer with no activation function and a single output neuron was then added on top of the LSTM layers to enable regression. The network was trained using 80% of the data for training with the rest used for validation, optimizing with respect to the mean absolute validation error. Note that unlike with the previous LR-based nowcasting where a new linear regression model was built for every input, here we build a single neural network model used for all inputs.

It should be noted, though, that the confidence intervals around the validation mean scores computed by the grid-search were very wide. Furthermore, a closer inspection of the loss over time (i.e. evolution of the training and validation errors during the training) revealed widely differing patterns even for different runs of the same network architecture (Figure 29). This was deemed to indicate an inability of the network to learn a stable pattern in the noisy data (despite the applied filtering).

In the models we presented so far, we have only been using the data points falling on the JVS collection dates (once a month). To see if the daily granularity of the online data can add value and improve the results, we also trained the neural network using the *daily* counts from the online data on a window of D + 2 months (where D is the delay)[76]. Despite the extra information, the results did not seem to constitute an improvement, which may have been due to a small window size used.

Figure 30 shows final comparisons of all models. The use of filtering is accounting for most of the improvement. Models with smaller delays performed better in general, which suggests previous JVS estimates are important predictors. The “LSTM-monthly” model performed best, with lowest mean and median absolute errors. The LR-based method seems to be a good compromise, with only slightly worse results than the “LSTM-monthly” while still being very simple to implement. Little, if any improvement was achieved for RMSE on the filtered data.

[74] We replaced the last D (where D is the JVS processing delay) JVS estimates with -1

[75] We experimented with other predictors such as number of employees - this was, however, only a static yearly estimate from business register and did not add to the accuracy of the model

[76] Note that using a larger window size would be desirable. However, due to the size of the input (≈30 x <size of the window> time steps), the training would take considerably more resources and time to train.

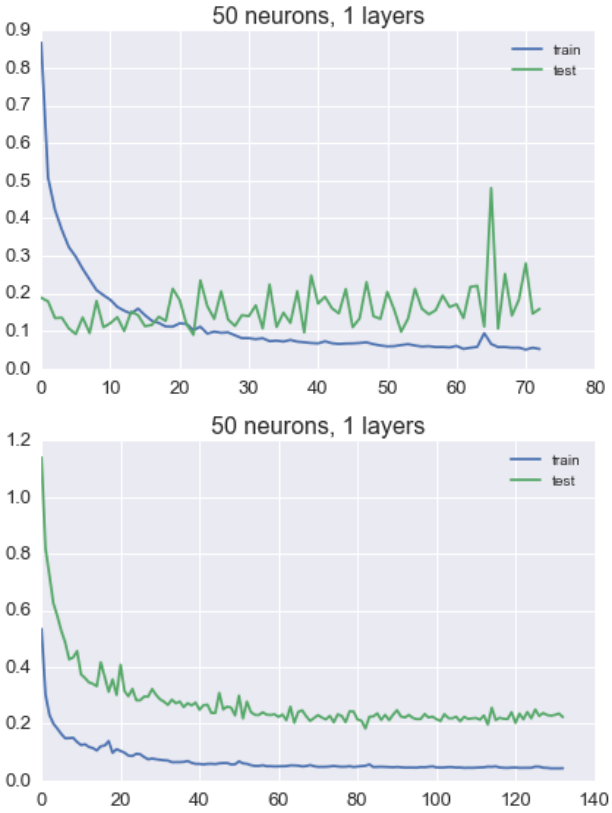


Figure 29: Loss over time for two runs of the network with identical architecture, with different train-validation splits.

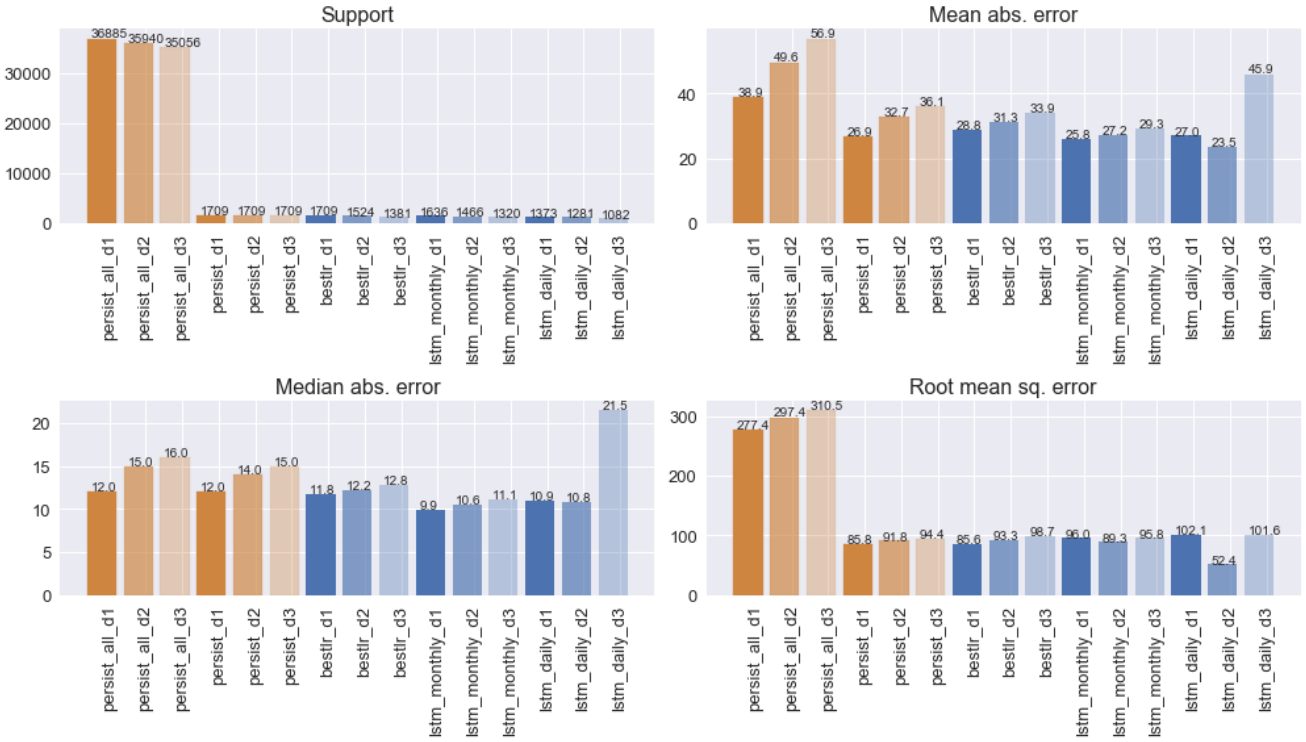


Figure 30: Final model comparison

7.4.5 Classifying trend

The models we introduced in the previous text were focused on regression, i.e. their output was a nowcasted JVS estimate value. It is then straightforward to turn these regressors into classifiers, nowcasting whether the time series has an increasing trend or not, based on a simple rule:

"time series has an increasing trend ⇔ nowcasted value > previous JVS value".

The results for these classifiers were, however, still mediocre, accurately nowcasting the trend roughly 2 out of 3 times and with the maximum accuracy staying below 70%.

7.4.6 Discussion

The conclusion from the nowcasting experiments above and its results indicate that it is very difficult to do nowcasting at a company level, especially to do it with high precision. There are several likely reasons for this. First of all, it seems that the monthly company-level JV time series has a highly oscillating nature, causing a small misalignment of the JVS and the online source to produce wrong nowcasts (see Figure 31 for an example). Smoothing the time series might help achieve better results, however, this might deny the purpose of producing real-time indicators. Second, the potential of the nowcasting based on other sources is greater for cases when the time series experiences a sudden significant spike (or drop), e.g. due to a start of a recruitment campaign or an important change in economic situation. In such cases, even a coarse indication of the upcoming trend may be useful.

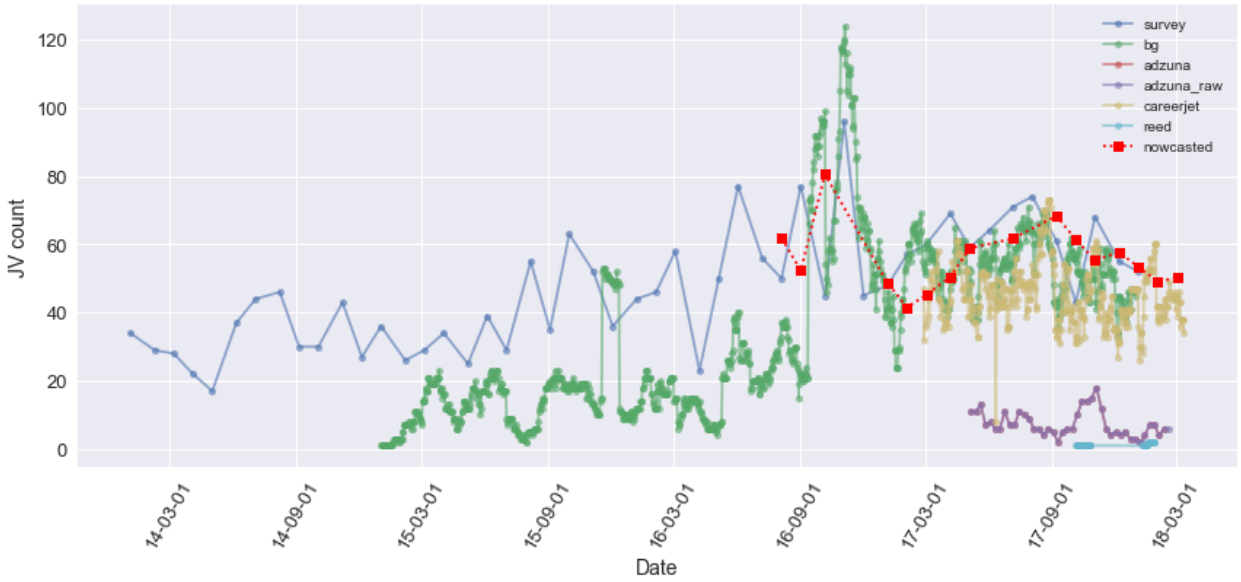


Figure 31: LSTM-monthly (delay 1) nowcasts for a specific company. The nowcasts are resembling a persistence shift due to the oscillating nature of the time series and a slight misalignment of the BG and JVS (e.g. around October 2016, BG shows an upward trend while JVS indicates decreasing trend)

Matching errors and errors present in the JVS itself[77] are some of the other reasons making the nowcasting difficult. However, the biggest factors preventing more precise nowcasts are the large differences in scales and trends between the JVS and the online sources, present for the vast majority of the companies.

Last, but not least, it is worth to say we had a bit of unfortunate timing, with most of the efforts in 2017 spent on getting access to reasonable datasets, and only short period of time at the start of 2018 available to work on the nowcasting models. A more thorough research, as well as in-depth research into existing methods would be needed to see whether company-level modelling can be considerably improved.

7.5 Statistical Outputs

In this section, we present a few statistical outputs of mainly experimental nature, which means there are usually little, if any, guarantees associated with the outputs. These are therefore not ready to be used as tools for decision making, and rather serve as proof-of-concepts illustrating some of the paths that are possible to follow in the future for creating statistical outputs based on online data sources.

7.5.1 Total JV counts in time

We start with simple visualisations of the total JV counts in time, as already showed in Figure 19. Figure 32 below shows the same information, where the online sources have been scaled to the JVS scale and averaged over each JVS collection date. We then averaged the online sources into a single time series. This average takes initially lot of influence from BG data (as it is one of the few online sources available at the start of the time period), gradually becoming smoother as other sources get involved. The relatively tight following of the JVS trend towards the end of the time series (where most sources are available) is encouraging, although longer time series (with multiple sources) would be required to see the viability of this simple approach.

A similar approach can be applied for individual industries, however, here we lack enough data sources with industry information (and long enough time series). Moreover, the confidence intervals around the industry-level time series of JVS are much wider, with a 10% coefficient of variation[78]. We thus did not pursue this direction further, although some industries displayed promising alignment of the time series (Figure 34).

[77] As an example, the JVS contained an entry with 0 vacancies for a company with 6000 employees, a very likely outlier.

[78] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/vacancysurveyqmi#validation-and-quality-assurance, retrieved 8th May 2018.

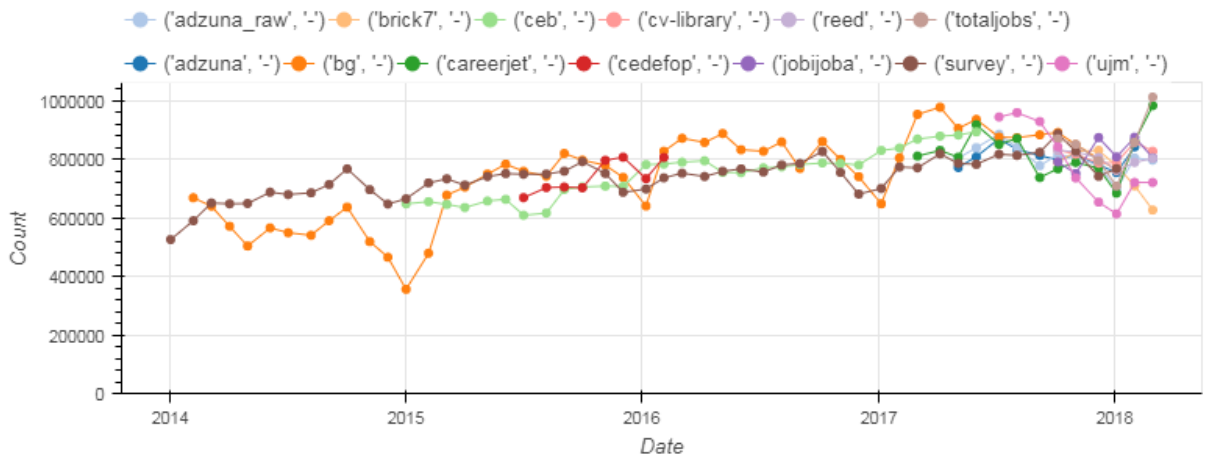


Figure 32: Time series of the total JV counts, averaged per month and scaled to the JVS scale. Zooming on the last few months of the time series reveals that the "end of year dip" happens one month later in each of the online sources, than it happens in the JVS.

Figure 33: Time series of the JVS and the average of the online sources (scaled to JVS scale)

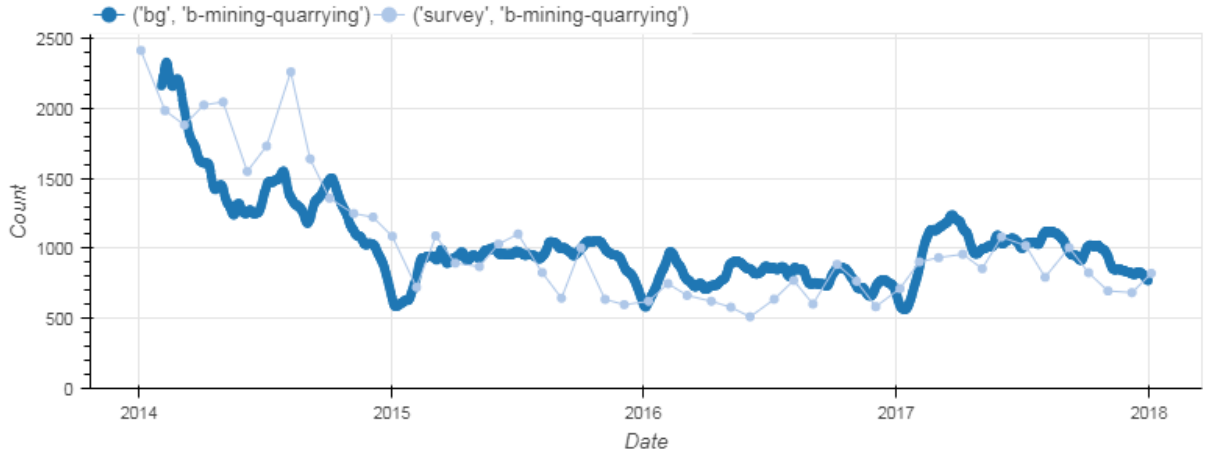


Figure 34: Time series of the JVS and BG for SIC class B (mining/quarrying)

7.5.2 Experimental location statistics

Some of the online sources contained information on location of the individual job vacancies. We therefore briefly looked at comparing the aggregated counts at the county and region levels for two sources: Adzuna and Burning Glass, our interest being in their similarities[79]. We also compared this to the JVS, using the address information from the business register, even though this address refers to the company's *headquarters*, rather than to the location of the job vacancy itself.

Figure 36 and Figure 35 show some of the created heatmaps and their comparisons, with discussion in the figure captions. Note that the London region has been taken out of the plots. This was done to make the colours for the rest of the counties span the full colour scale, as otherwise (due to the very high number of JVs in London) the out-of-London counties would only take colours from bottom end of the colour spectrum.

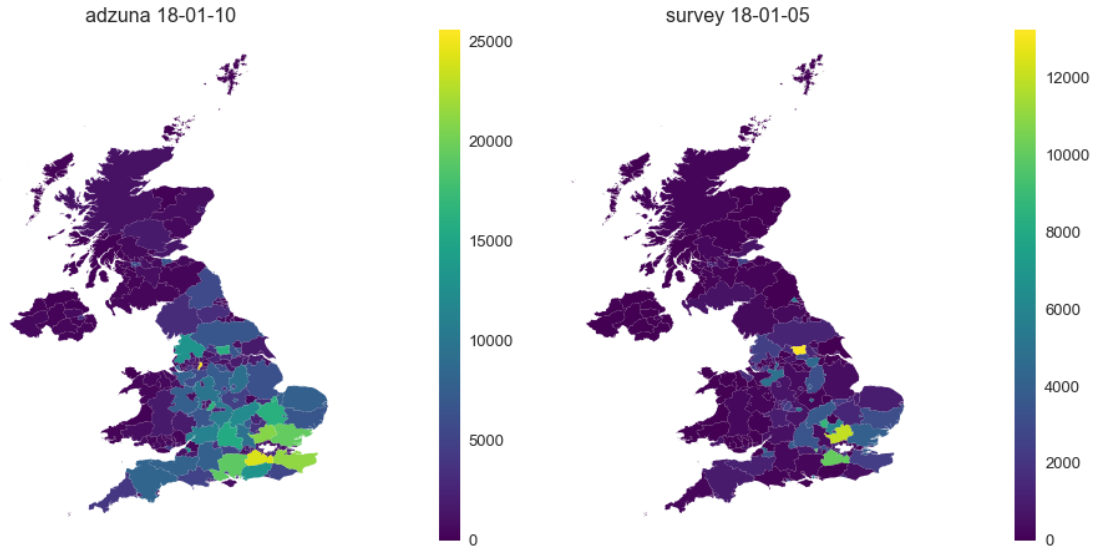


Figure 35: JV counts by county (with London region removed), Adzuna vs. JVS. This highlights the difference between the locations of the job vacancies and the locations of the *headquarters* of the companies advertising the job vacancies (the only location information contained in JVS)

[79] Note, however, that the compared snapshots of two datasets were from different dates, making them less comparable. This was because we only got Adzuna data with location information from 10th of January 2018 onwards, while the last batch of BG data (at the time of this writing) contained data till the end of 2017.

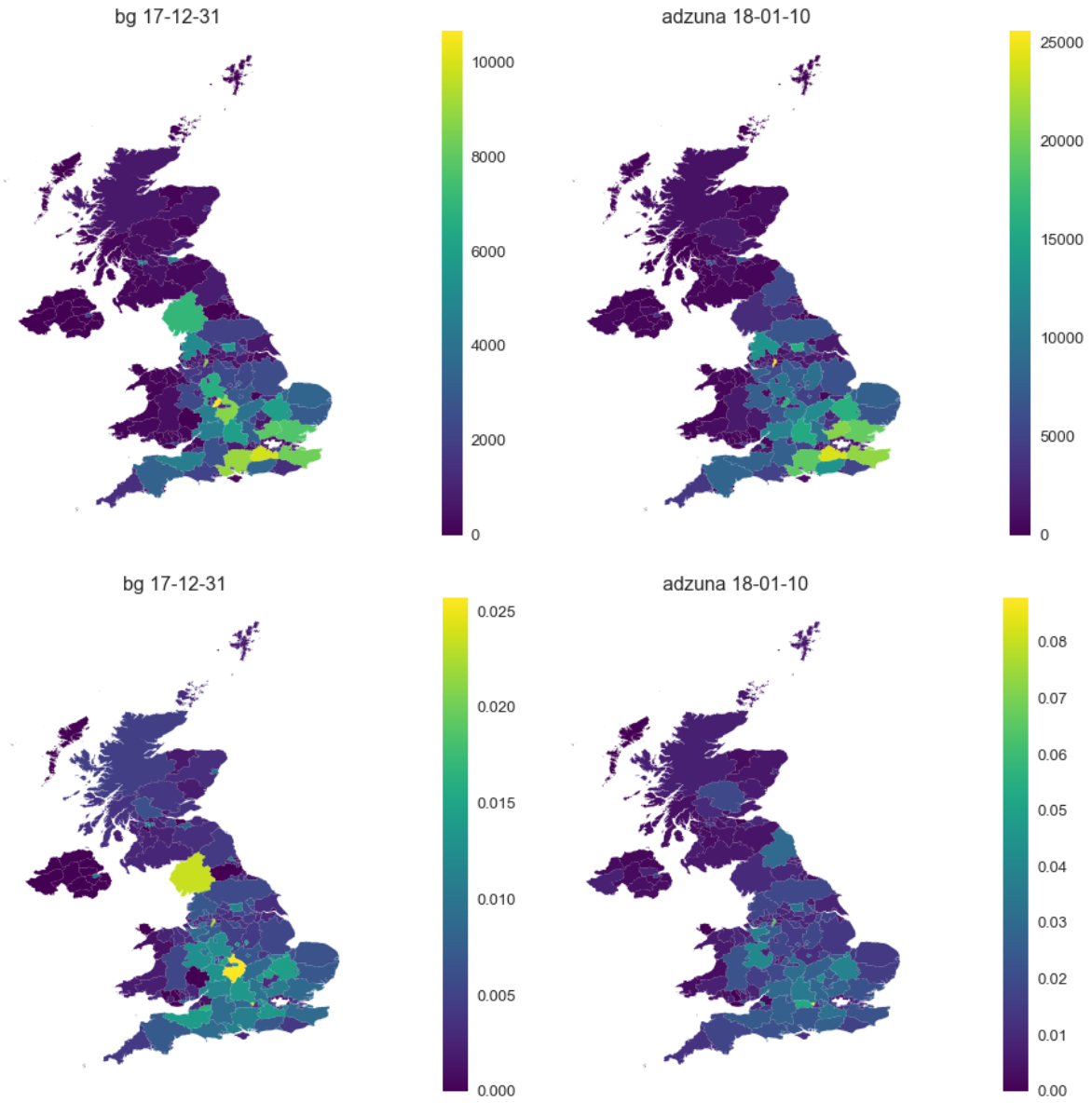


Figure 36: JV counts by county (with London region removed), BG vs Adzuna. The top plot shows the raw JV counts, with reasonable similarities in their distribution between the two sources. Some differences become more pronounced when one looks at the JV counts scaled by the working population of the county (bottom plot). In general, this (and other) experiment should be repeated with both datasets aligned (and possibly aggregated) over the same date period.

7.5.3 Nowcasts scaled to the total JV counts

Despite the relatively weak performance of the nowcasting models, we attempted to create an index of the total number of job vacancies by scaling up the company-level nowcasts. This was done in a simple way. First, we obtained all available nowcasts for the target date. Then, for the sample of companies for which we had the nowcasts, we computed the proportion of the total formed by the companies in the previous month. Assuming this proportion changes little from month to month, we then used it to scale the nowcasts up to the total JV counts level. Figure 37 displays the index based on the LSTM-monthly model with delay D = 1. The nowcasted curve only roughly follows the true JVS

totals and there is much room for improvement. This is not only down to the quality of the company-level nowcasts, but also due to a small sample size (with only ≈100 company nowcasts every month), skewed industry distribution and the simplistic way of scaling the company nowcasts up.

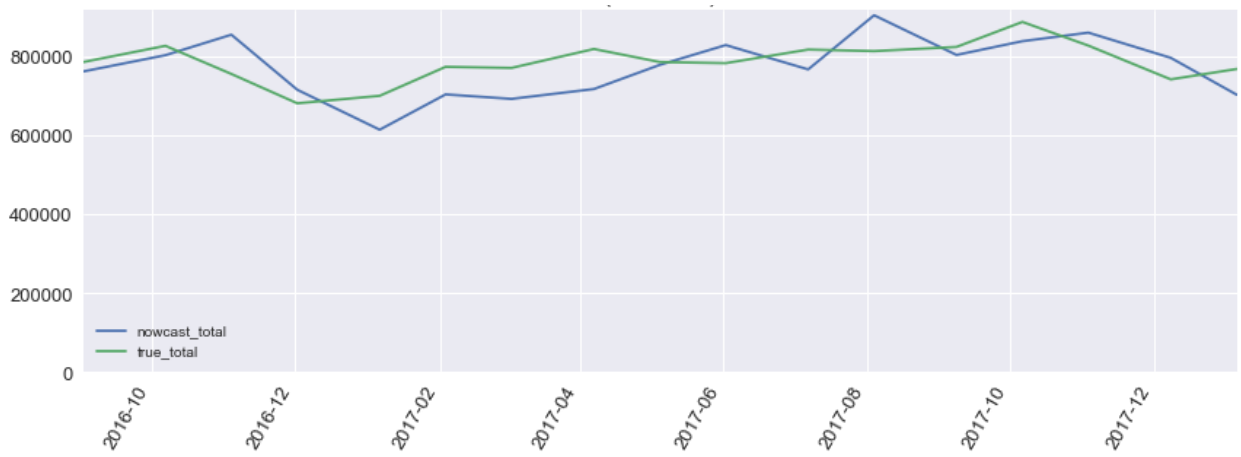


Figure 37: Nowcasted total JV counts vs. the true total

7.5.4 Index based on the time series model

Our final experiment was performed only a few days before this report was due. Here we looked at fitting a time series model S-ARIMA-X (seasonal ARIMA with exogenous regressor, see Durbin and Koopman, 2012) to the time series of total JV counts for JVS and Burning Glass. The idea was to see if a model based solely on JVS could be outperformed by adding additional information from Burning Glass.

The models were chosen by performing an exhaustive search through the possible S-ARIMA-X parameters and choosing the one with the lowest AIC values. Residual plots were examined for signs of remaining information or correlation. Only data till the end of 2016 were used in the fitting process, with the rest of the data used for evaluation.

The initial improvement of the model with added BG data was only very minor (5% improvement in RMSE). However, once we allowed time-varying coefficients in the model, the improvement was more visible, reducing the RMSE by roughly 50%.

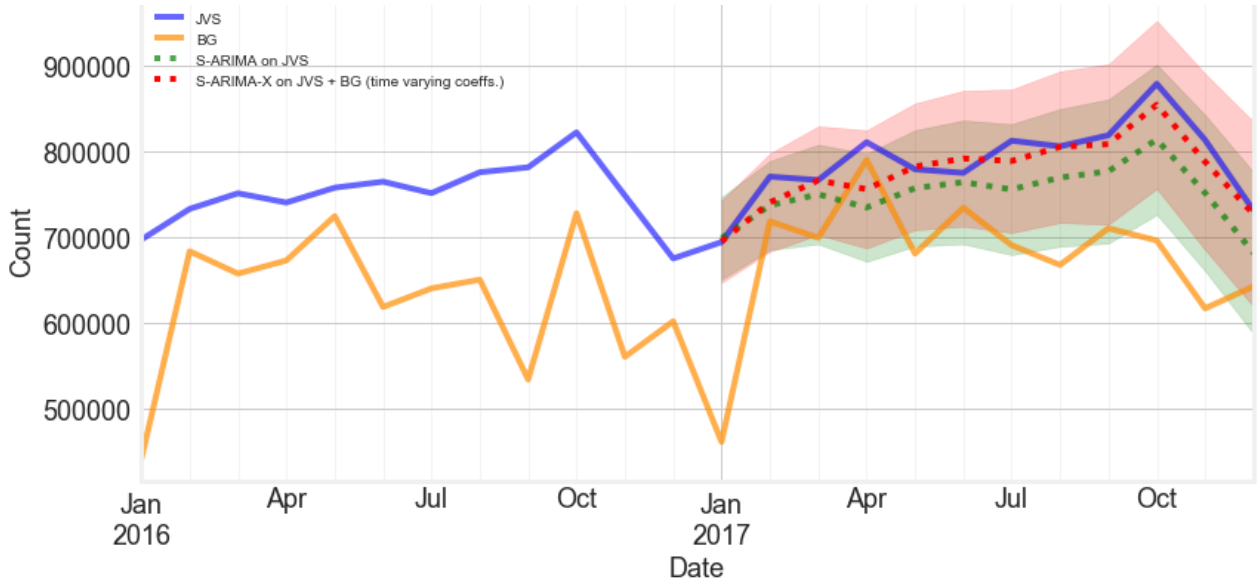


Figure 38: The nowcasts based on the S-ARIMA-X time series model. The green dotted line represents the nowcasts based only on JVS, while the red dotted line shows the (more precise) nowcasts that include the Burning Glass as exogenous variable. The shaded areas of respective colours indicate respective 95% confidence intervals.

Similarly, as with the previous experiments, this analysis too would require more data (especially longer time series) and investigation to determine and understand the reliability of the produced nowcasts. The improvement in RMSE however indicates a good potential of using this method. A possible direction in the future may involve including multiple sources as exogenous variables and looking at the feasibility of this approach at lower granularities (JV counts by industry or company).

7.6 Future Perspectives

This work has explored several paths utilising the online job vacancy data with the goal of producing or enhancing the current job vacancy statistics. The main focus was on exploring the possibility of doing nowcasting, potentially bringing more timely and frequent estimates of job vacancy counts.

A large portion of our efforts was spent on investigating nowcasting at the company level. This proved more challenging than it was initially expected, with the differences between the online sources and JVS being often prohibitively large and unstable, preventing applying effective machine learning. Therefore, only relatively moderate improvements were achieved over the baseline persistence model, with the best performing model (based on LSTM neural networks) achieving around 10-40% improvement (measured by mean/median absolute error of the nowcasts). Accuracy of classifying the current JVS trends (increasing/non-increasing) was reaching towards, but not surpassing 70%. In general, the nowcasting models seemed to work only at a coarse level, with precise company level nowcasts being much more difficult to achieve. Nowcasts scaled up to the total level could then roughly follow the true JVS trend, although still leaving a lot of room for improvement.

A more promising last-minute and briefly researched approach was using a time series (S-ARIMA-X) model which included the Burning Glass data source as an exogenous variable. Here a measurable

improvement over a JVS-only based S-ARIMA model was detected, with RMSE error reduced by 50%. It would be interesting to repeat and validate this analysis with other sources, as well as multiple data sources at the same time and to evaluate it on longer time series.

Although we managed to get access to (or web-scraped) a solid base of online data, most of the data was available only towards the end of 2017, with not enough time series. Furthermore, the data in this period follow a relatively stable trend and pattern, whereas the use of real time online data can provide most value in situations with sudden changes to this pattern, providing a quick feedback for decision makers. It would thus be interesting to re-visit the ideas explored in this project in the future, when a few years' worth of data is built up and available for analysis.

Finally, it should be noted that the nature of this work is still highly experimental, with little guarantees around the produced statistical outputs. The issues around representivity of the online data and their coverage of the true population are still very much in need of further research, especially in cases where a direct comparison to the JVS cannot be made due to the lack of relevant variables in JVS (e.g. location).

7.7 References

Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. "*Querying and mining of time series data: experimental comparison of representations and distance measures*". Proc. VLDB Endow. 1, 2 (August 2008); As of 13.4.2018 available at: https://pdfs.semanticscholar.org/47ab/b668c6205a64bddd9573078d4aa002eea926.pdf.

Nigel Swier, Frantisek Hajnovic, Ingegerd Jansson, Dan Wu, Boro Nikic, Christina Pierrakou, Martina Rengers. 2017. "*Work Package 1, Web scraping / Job vacancies, Deliverable 1.3, Final Technical Report (SGA-1)*". As of 13.4.2018 available at:

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/20/Deliverable_1_3_main_report_final_1.0.pdf.

Gerard Salton, Christopher Buckley. 1988. "*Term-weighting approaches in automatic text retrieval*". Information Processing & Management, Volume 24, Issue 5, 1988, Pages 513-523. As of 13.4.2018 available at:

http://www.cs.bilkent.edu.tr/~canf/CS533/saltonBuckley1988.pdf.

Sepp Hochreiter, Jürgen Schmidhuber. 1997. "*Long Short-term Memory*". Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735. As of 13.4.2018 available at:

http://www.bioinf.jku.at/publications/older/2604.pdf.

James Durbin, Jan Siem Koopman. 2012. "*Time Series Analysis by State Space Methods*". OUP Catalogue, Oxford University Press, edition 2, number 9780199641178. The applied methods are from *statsmodels* Python package:

http://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html#r80.