ESSnet Big Data II

Grant Agreement Number:

NUMBER — 847375 — 2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Workpackage WPB

Implementation – Online Job Vacancies

CEDEFOP exploratory analysis

Version 22.10.2019

Prepared by: Martina Rengers, Jakob de Lazzer (DESTATIS, Germany)

Workpackage Leader:
Tomaž Špeh (SURS, SI)
e-mail address: tomaz.speh@gov.si
mobile phone: +38651672116

Table of contents

Figures

Tables

1 Data processing capabilities of the BDTI (Big Data Test Infrastructure)

Before showing any results, it is necessary to acknowledge the limitations imposed on the analysis by the BDTI. The BDTI limits each user to a maximum of 4 GB of RAM. This means that it is not possible to load the CEDEFOP OJA dataset for large countries like Germany into memory. Only one subset containing a maximum of 4 variables can be loaded at any time and therefore no analysis can take place across multiple variables for the full sample.

It also means that any analysis which requires involved computations (e.g. machine learning, string matching, deduplication) is not possible on the full sample. To emphasize this point, we have calculated that the current sample for Germany has an in-memory size of ~7700 MB. R programming guidelines recommend a data to operating memory ratio of at least 1:3 for basic operations and massively more for estimation, machine learning or matching procedures. When using parallel processing for higher computing speeds, these memory requirements will be roughly multiplied by the number of processing cores. Even when using clever low-memory programming and subsampling, we need to assume that more involved analysis will require a minimum of 128GB of memory and 8 processing cores per user.

2 Country identification, timeframe and sample size

Data for Germany refers to all observations in the BDTI Cedefop database for which $sourcecountry == DE$. ***16867910*** observations total.

Of these, ***15578258 are distinct*** (not duplicate) observations.

15220564 are ads for jobs ***located in Germany*** (country == DEUTSCHLAND):

BELGIQUE-BELGIË	21717
ČESKÁ REPUBLIKA	233
DEUTSCHLAND	15220564
ESPAÑA	4100
FRANCE	31151
IRELAND	3175
ITALIA	6490
LUXEMBOURG	1500
NEDERLAND	46313
ÖSTERREICH	167311
POLSKA	402
SVERIGE	522
UNITED KINGDOM	74780
Total	15578258

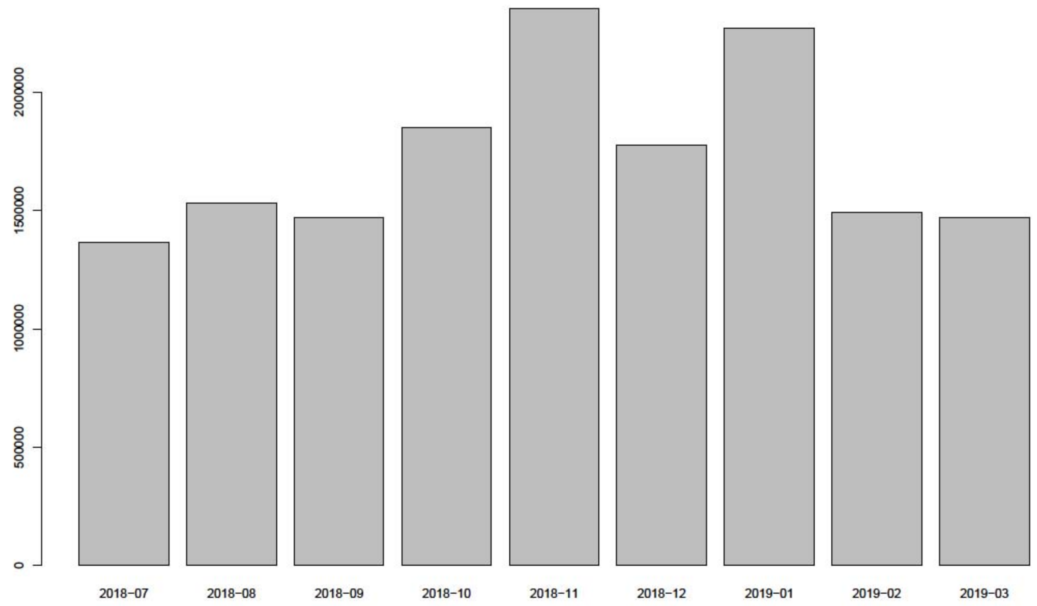
However, there seem to be some geographic consistency issues between *sourcecountry* and *source*. When *sourcecountry* is set to *DE* the sample also contains job ads from various sources outside of Germany. These sources include, for instance, FR_METEOJOB, CH_GIGAJOB or NL_JOBBIRD.

All observations were ***collected*** in the last two quarters of 2018 and the first quarter of 2019 (***01.07.18-31.03.19***).

194 ads were already expired by the time they were collected, according to the variable *expire_date*, and were dropped from the sample.

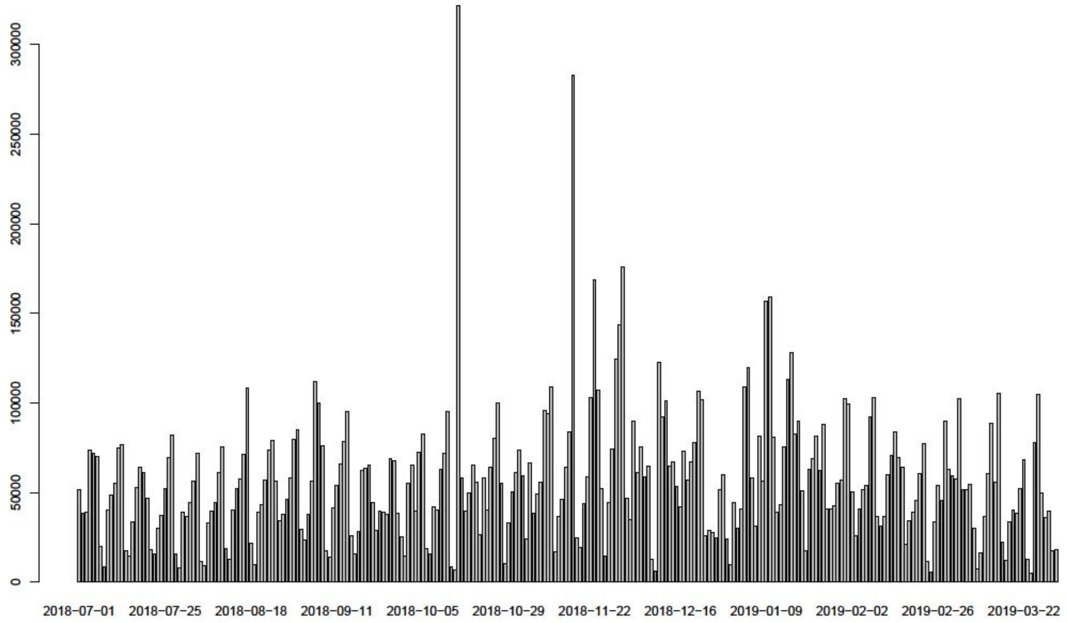
The frequency of scraped job ads varies by month

Figure 1 Total number of job ads by month



…and also by day of scraping:

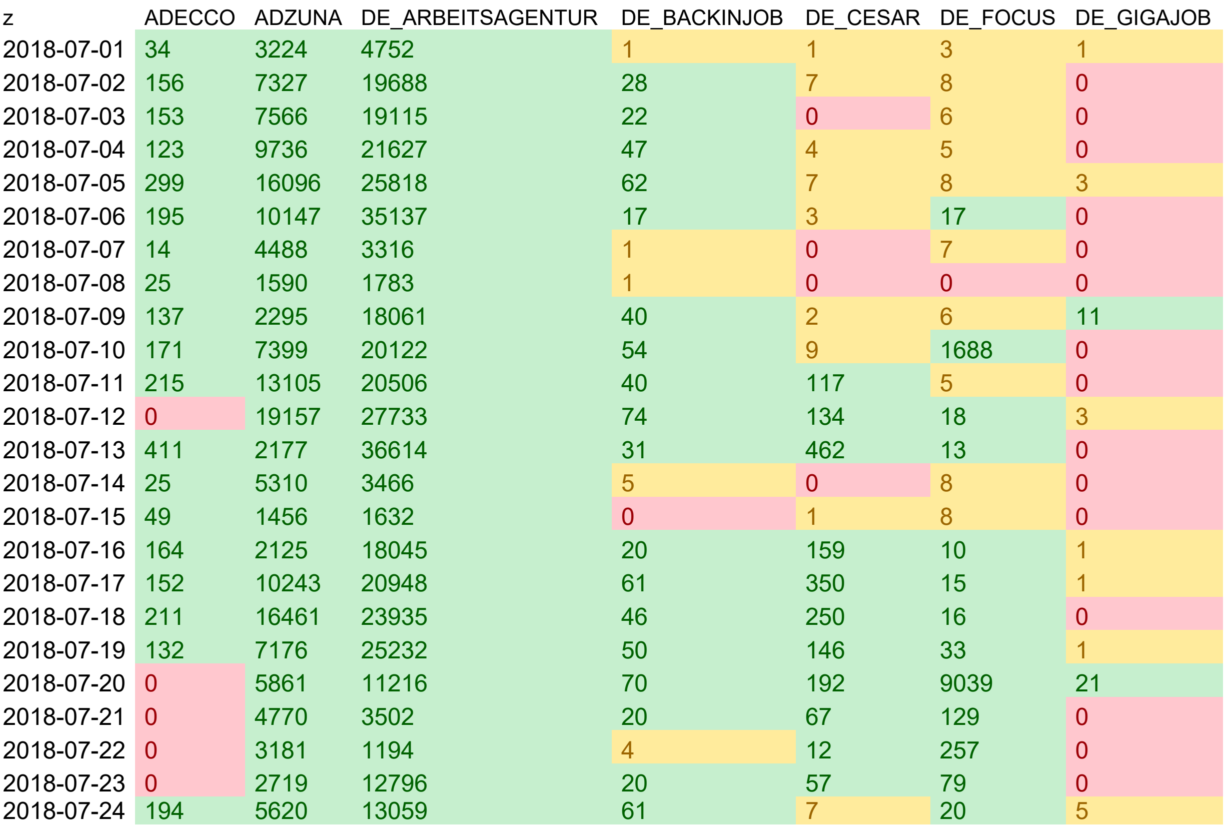
Figure 2 Total number of job ads per day



3 Coverage of sources and sites in data collection

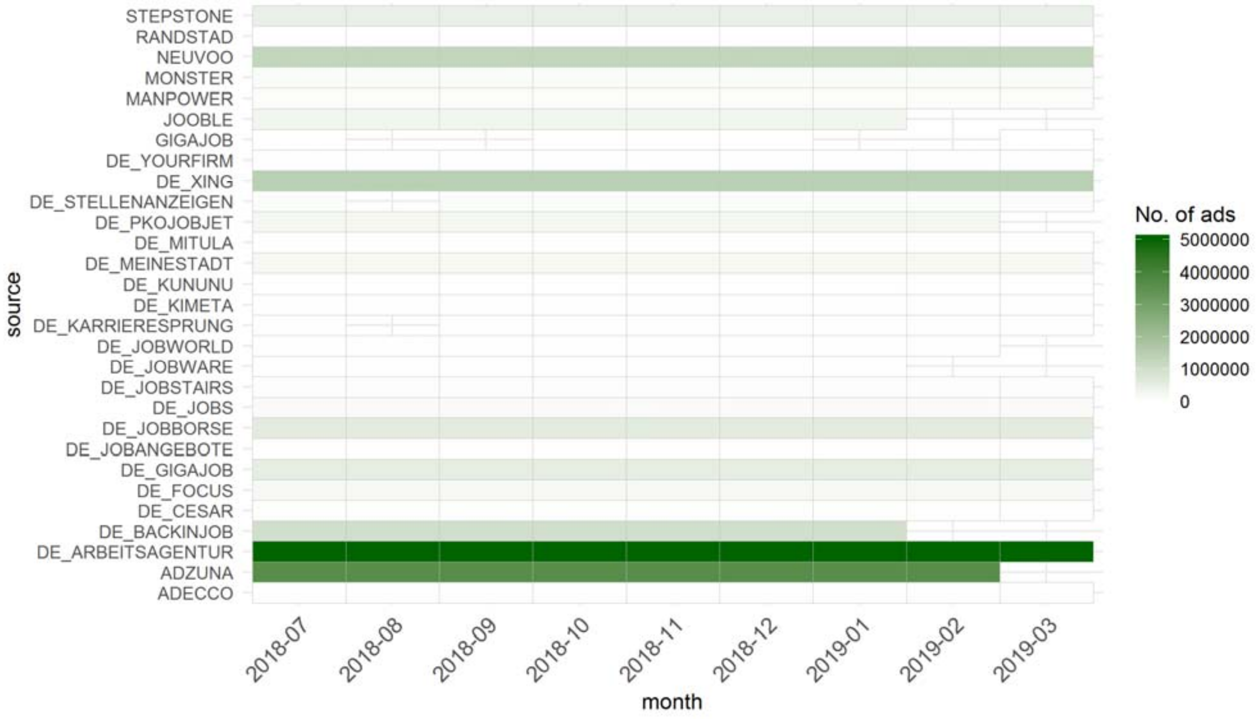
Not all ***sources*** were collected every day (Table 1). Most sources display large day-to-day variations in the number of collected job ads. Some sources have gaps in coverage of single days up to more than a month (Figure 3). Note that data collection for several sources, e.g. ADZUNA and DE_BACKINJOB, *stopped prematurely* after 01.2019.

Table 1 Different sources per day (excerpt)



z	ADECCO	ADZUNA	DE_ARBEITSAGENTUR	DE_BACKINJOB	DE_CESAR	DE_FOCUS	DE_GIGAJOB
2018-07-01	34	3224	4752	1	1	3	1
2018-07-02	156	7327	19688	28	7	8	0
2018-07-03	153	7566	19115	22	0	6	0
2018-07-04	123	9736	21627	47	4	5	0
2018-07-05	299	16096	25818	62	7	8	3
2018-07-06	195	10147	35137	17	3	17	0
2018-07-07	14	4488	3316	1	0	7	0
2018-07-08	25	1590	1783	1	0	0	0
2018-07-09	137	2295	18061	40	2	6	11
2018-07-10	171	7399	20122	54	9	1688	0
2018-07-11	215	13105	20506	40	117	5	0
2018-07-12	0	19157	27733	74	134	18	3
2018-07-13	411	2177	36614	31	462	13	0
2018-07-14	25	5310	3466	5	0	8	0
2018-07-15	49	1456	1632	0	1	8	0
2018-07-16	164	2125	18045	20	159	10	1
2018-07-17	152	10243	20948	61	350	15	1
2018-07-18	211	16461	23935	46	250	16	0
2018-07-19	132	7176	25232	50	146	33	1
2018-07-20	0	5861	11216	70	192	9039	21
2018-07-21	0	4770	3502	20	67	129	0
2018-07-22	0	3181	1194	4	12	257	0
2018-07-23	0	2719	12796	20	57	79	0
2018-07-24	194	5620	13059	61	7	20	5

Figure 3 Ads collected by source and month

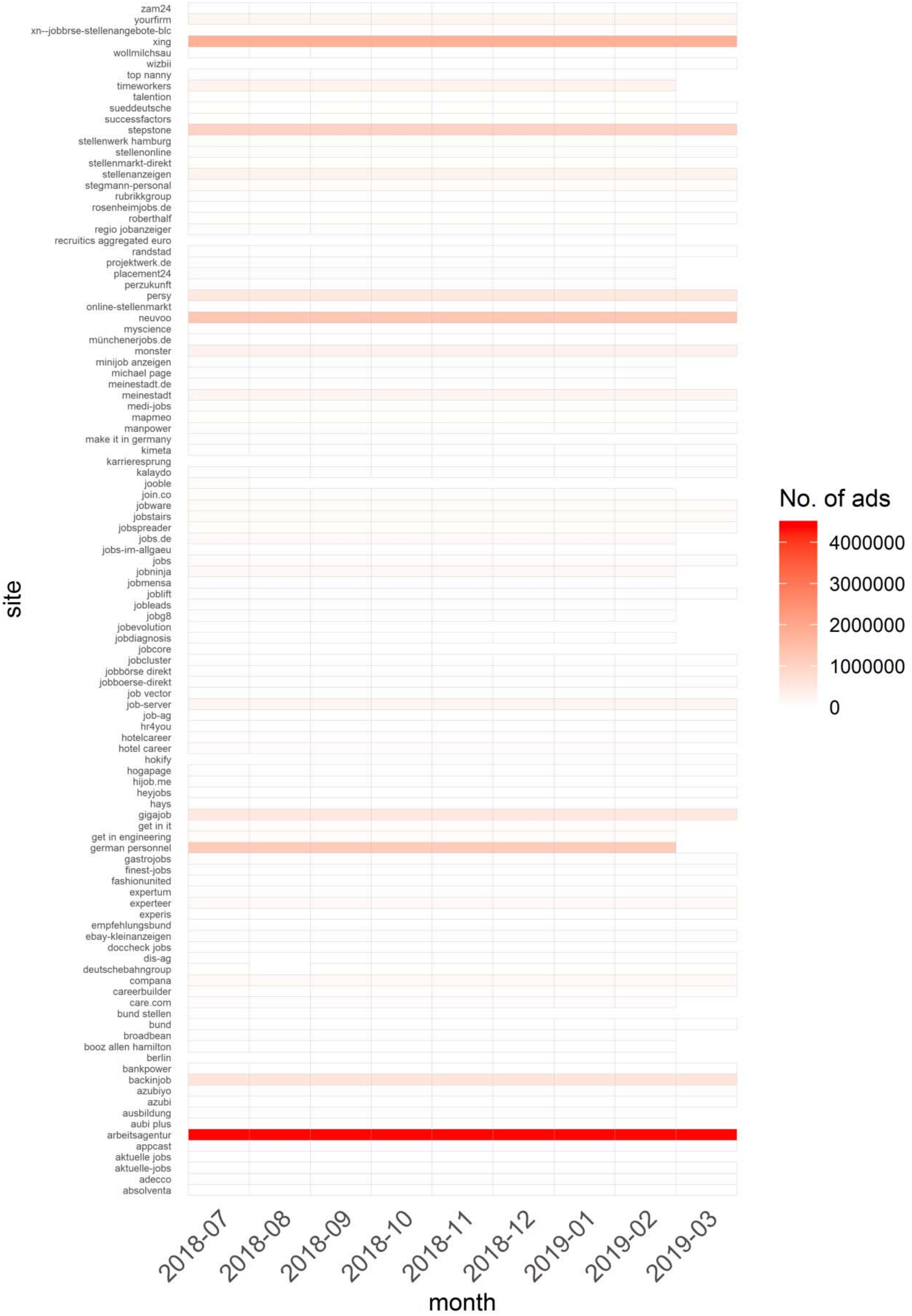


Nor were ads from every *site* consistently collected (Table 2, Figure 4)

Table 2 Different sites per day (excerpt)

grab_date	arbeitsagentur	backinjob	care.com	compana	ebay.kleinanzeigen	experteer	get.in.engineering
2018-07-01	3986	1	46	1	0	352	0
2018-07-02	17705	27	32	29	0	11	0
2018-07-03	16279	21	22	0	0	202	0
2018-07-04	18802	45	37	0	0	314	0
2018-07-05	22387	62	148	0	0	1221	0
2018-07-06	28022	16	43	2	0	991	249
2018-07-07	3075	2	78	0	0	160	0
2018-07-08	1728	1	69	0	0	31	0
2018-07-09	17370	39	14	1	1	283	0
2018-07-10	19219	53	81	0	1	358	0
2018-07-11	19265	38	0	0	0	1088	0
2018-07-12	26475	74	116	0	0	1157	107
2018-07-13	34116	31	38	1	0	111	0
2018-07-14	3082	5	72	0	0	377	0
2018-07-15	1111	0	70	0	0	68	0
2018-07-16	17014	19	69	1	0	99	0
2018-07-17	19489	61	40	2	0	984	68

Figure 4: Ads collected by site and month



4 Flow or stock?

In order to evaluate total labor demand based on online job ads (OJA), we need to know if OJA data refers to the stock of vacancies at a point in time or to the inflow of new vacancies relative to the previous period. However, ***Cedefop data represent neither the flow nor the stock of online job vacancies***.

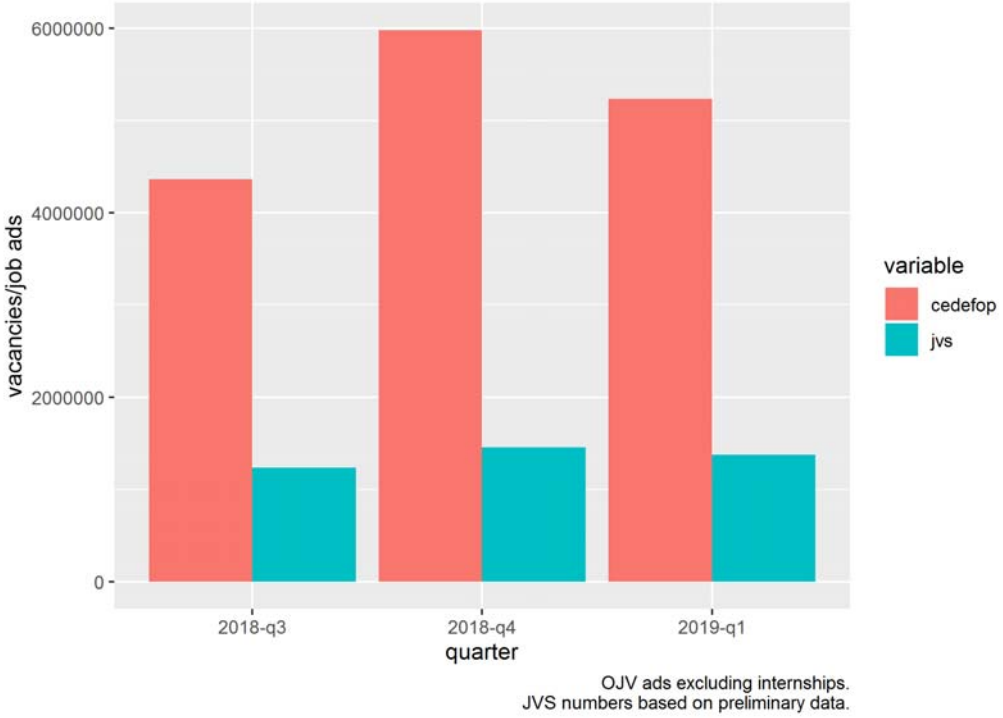
If it was flow data, we would see a large initial stock on which to base the flows of ads. This is not the case, as is evident by the low numbers of initial observations in Figure 1, Figure 2, and the second rows of Table 1 and Table 2.

If it were stocks, we would expect to observe similar stocks of observations for every day. This is also not the case, as many sites have large between-day fluctuations in the numbers of ads or gaps in their ad timeline (compare, for instance, column 8 in Table 2).

In order to do a plausibility-check for flow or stock data, we compare quarterly CEDEFOP data with the results from the German job vacancy survey (JVS). The JVS reflects the number of open positions at the time of the survey. Since companies have several weeks to fill out the survey, the results are more akin to the average stocks over the survey timeframe. While we don't expect OJA stocks to closely match the JVS, for a variety of reasons, they should at least be on the same order of magnitude.

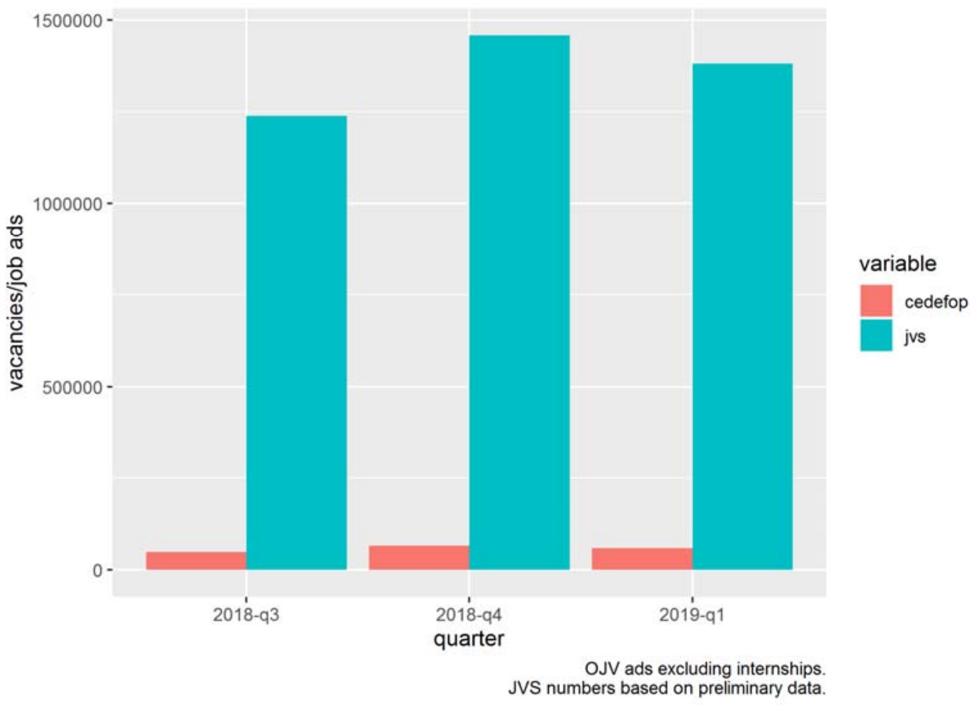
In Figure 5 we treat the CEDEFOP as flows of job vacancies, summing them over each quarter. The resulting quarterly aggregates are far too high to be plausible.

Figure 5: CEDEFOP-OJAs vs. JVS: quarterly sum of daily grab numbers



In Figure 6 we treat the CEDEFOP as stocks of job vacancies, taking the mean over days in each quarter to make them comparable to the JVS. The resulting quarterly averages are extremely low and therefore also not plausible.

Figure 6: CEDEFOP-OJAs vs. JVS: quarterly average over daily grab numbers



OJV ads excluding internships.
JVS numbers based on preliminary data.

5 Pseudo-stocks

The data layout and descriptive material from CEDEFOP hint at an unstructured approach to collecting OJA-data by scraping new ads whenever there happen to be resources available. This results in large day-to-day fluctuations in quantity, coverage and choice of source which we observe in the data. It also means that we can’t interpret the raw data as flow or stock.

In order to address this issue, ***we explore the possibility of generating pseudo-stocks from CEDEFOP data*** which are an approximation to the underlying stock of job ads. The idea behind pseudo-stocks is the following:

At a given point in time (T), we consider a job ad to be valid if it hasn’t expired yet and was posted a maximum of 30 days before T. Expiration of a job ad is determined by the variable *expire_date*. Preliminary analysis of other sources indicates that job ads stay relevant for roughly 30 days on average (see ESSnet WP 1). The number of all job ads which are considered at time T is the pseudo-stock of active job ads at time T.

This approach has the advantage that it can account for gaps in scraping coverage as long as the missing ads are scraped at some point within 30 days of posting.

One drawback of this approach is that for the first 30 days in the observed timeframe too few valid vacancies are available. Therefore pseudo-stocks are only useful from the 01.08.2018 onward.

In order to ease the comparison with JVS, we take the average of pseudo-stocks over the last month of each quarter and arrive at job ad stocks which are very roughly comparable with those of the JVS

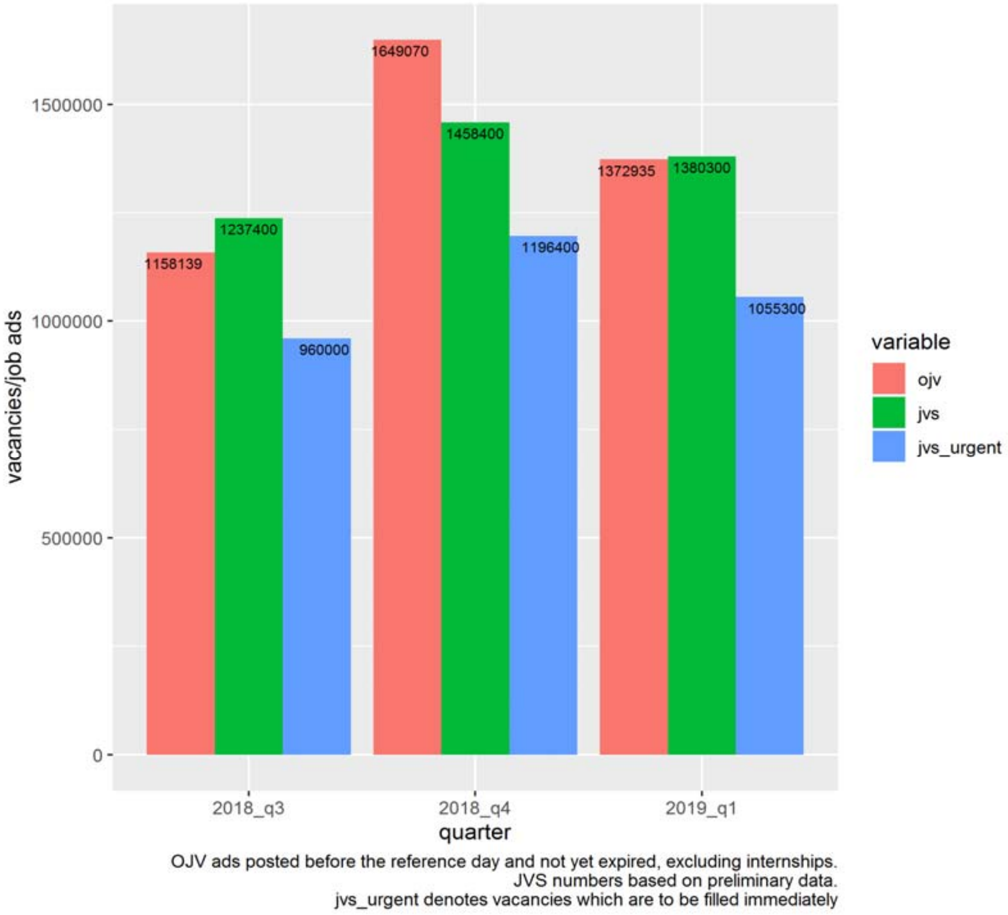
(Figure 7). The German JVS quantifies two types of vacancies, *general vacancies* which are to be filled at some indeterminate point in the future and *urgent vacancies* which are to be filled immediately.

For Q3 2018 and Q1 2019 the number of pseudo-stock job ads is close to the number of reported general vacancies in the JVS and in Q4 2018 it is substantially higher. Urgent vacancy numbers however lie below those of OJA and closely mimic their development over time. We consider urgent vacancies to be a better equivalent to OJA than general vacancies, since the publication of OJA indicates that the recruitment process has started and the position is to be filled at a fixed point in time.

OJAs refer only to a subset of all vacancies, therefore we would expect accurately measured OJA stocks to be lower than JVS vacancies. But despite the efforts of CEDEFOP at deduplication, we have to assume that the available OJA data contains a large fraction of duplicates. OJA data can also contain “ghosts” and ads without underlying vacancies.

Additionally, we can’t entirely rule out that the assumptions governing our calculation of pseudo-stocks may be inaccurate. It is also conceivable that the variable *expire_date* denotes something other than the actual expiration date of the job ad.

Figure 7: CEDEFOP-OJAs vs. JVS: pseudo-stocks average over last 2 month each quarter



6 CEDEFOP OJA data descriptives by geographical level

6.1 Missing location information

Location information is very likely at least partially generated or imputed by CEDEFOP. Geo information is missing to various degrees at the different regional levels (Table 3). The more fine-grained the location data, the more observations are missing. However, it is unclear why this should be the case. Sampling of job ads in Germany shows that if ads contain workplace information they always contain the city name (or province in sparsely populated rural areas). The only exception to that are ads by recruitment firms who only state a very vague location. Therefore, observations which contain federal country or region but no city or province information might indicate ads by recruitment firms.

Table 3: Missing geo-data by NUTS level

NUTS Level	Missing observations (of 15578258)	Missing percent
1 (federal country)	3347966	21.5%
2 (region)	3848241	25%
3 (province)	4894081	31%
City	5863889	38%

For reasons of simplicity, we initially focus our geographical analysis on the federal country level.

6.2 Missing by contract, month and source

Location data for the federal country is relatively uniformly missing by contract type (Table 4) and over time (Table 5).

Table 4: Missing geo-data by contract type

by contract type	Non missing	missing	share
	2770209	818361	0.23
Internship	2234451	627488	0.22
Permanent	2569378	620977	0.19
Self Employment	1797126	471267	0.21
Temporary	2859128	809873	0.22

Table 5: Missing geo-data by month

By month	Non missing	missing	share
2018-07	1075728	290174	0.21
2018-08	1221346	308749	0.20
2018-09	1136858	332385	0.23
2018-10	1454932	394931	0.21
2018-11	1848114	506226	0.22
2018-12	1373603	403940	0.23
2019-01	1813355	455526	0.20
2019-02	1180520	310976	0.21
2019-03	1125836	345059	0.23

There are, however, large differences in the availability of location data between sources. As Table 6 shows, some sources like ADECCO, JOOBLE or ARBEITSAGENTUR have few missing locations, while others, e.g. MEINESTADT or PKOJOBJET only provide locations for a small fraction of their ads.

Table 6: Missing geo-data by source

By source	Non missing	missing	share
ADECCO	24833	2162	0.08
ADZUNA	2868790	760433	0.21
DE_ARBEITSAGENTUR	4041328	970171	0.19
DE_BACKINJOB	858150	134418	0.14
DE_CESAR	18691	10070	0.35
DE_FOCUS	147175	40361	0.22
DE_GIGAJOB	407115	115590	0.22
DE_JOBANGEBOTE	3607	3291	0.48
DE_JOBBORSE	389271	216806	0.36
DE_JOBS	67728	42278	0.38
DE_JOBSTAIRS	39436	13592	0.26
DE_JOBWARE	16051	4778	0.23
DE_JOBWORLD	17755	4012	0.18
DE_KARRIERESPRUNG	29605	11281	0.28
DE_KIMETA	12927	11702	0.48
DE_KUNUNU	8415	2240	0.21
DE_MEINESTADT	81691	118304	0.59
DE_MITULA	8890	4240	0.32
DE_PKOJOBJET	61546	181957	0.75
DE_STELLENANGEBOTE	2240	2052	0.48
DE_STELLENANZEIGEN	36987	40104	0.52
DE_XING	1277806	208383	0.14
DE_YOURFIRM	13562	4483	0.25
GIGAJOB	4942	1603	0.24
JOOBLE	275640	17847	0.06
MANPOWER	55489	12822	0.19
MONSTER	101124	7054	0.07
NEUVOO	967708	310471	0.24
RANDSTAD	6494	1422	0.18
STEPSTONE	374718	87990	0.19

7 Pseudo stocks by federal country

The following figures plot the geographical distribution of OJA pseudo stocks by federal country for the three quarters of the observed time-period. We observe large differences in OJA intensity across federal countries, with job ads primarily concentrated in highly populated areas. For comparison, Figure 11, plots the population distribution across federal countries. In fact, the geographic distribution of ads closely mimics the geographic distribution of individuals.

The relative shares of job ads by federal country are quite stable over time. While total numbers of job ads vary between the quarters of the observed timeframe, they do so evenly across regions.

Figure 8: Map of job ad pseudo-stocks Q3 2018

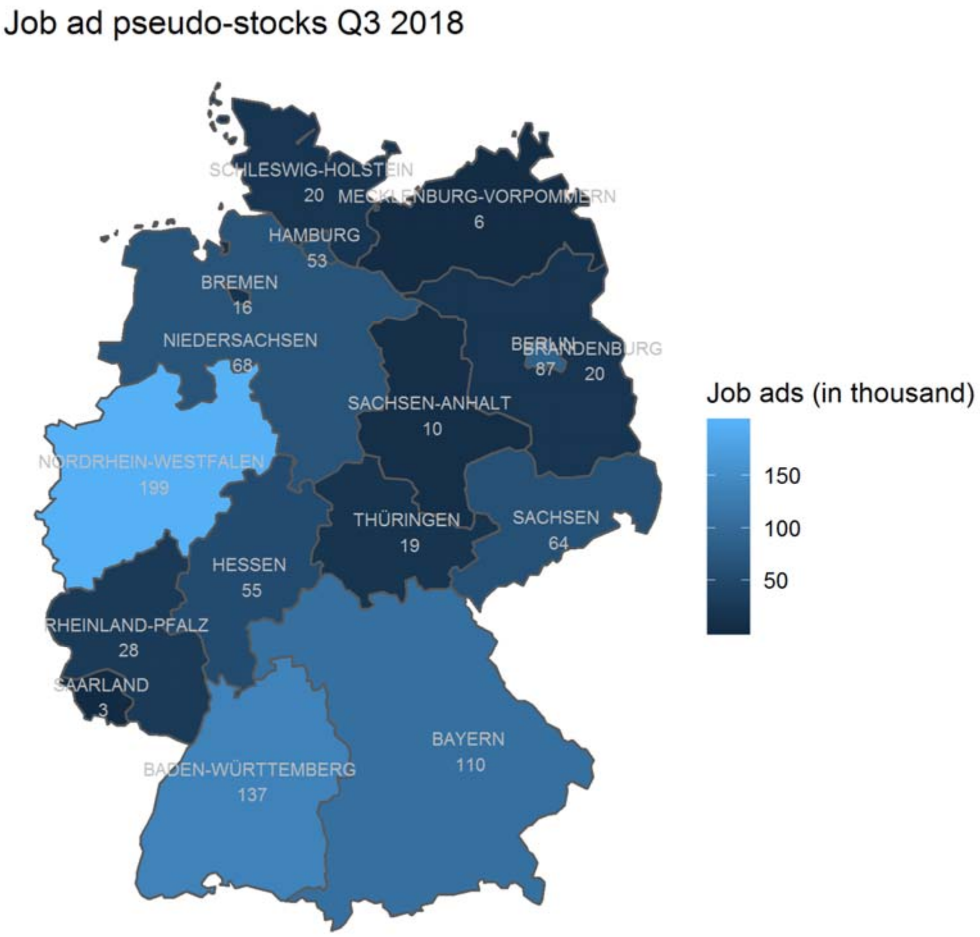


Figure 9: Map of job ad pseudo-stocks Q4 2018

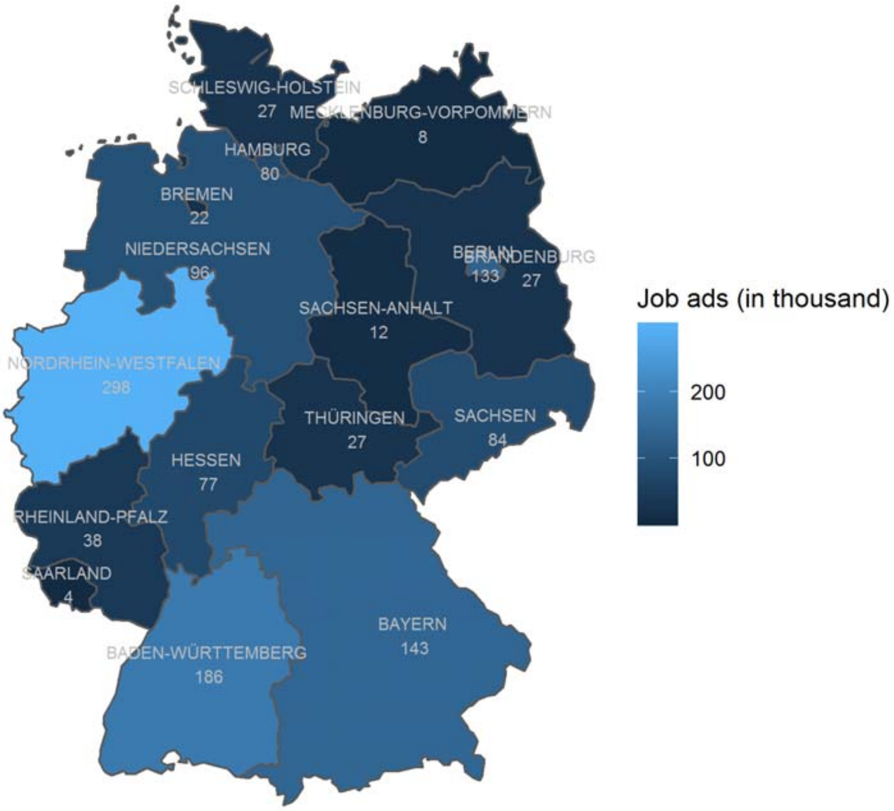


Figure 10: Map of job ad pseudo-stocks Q1 2019

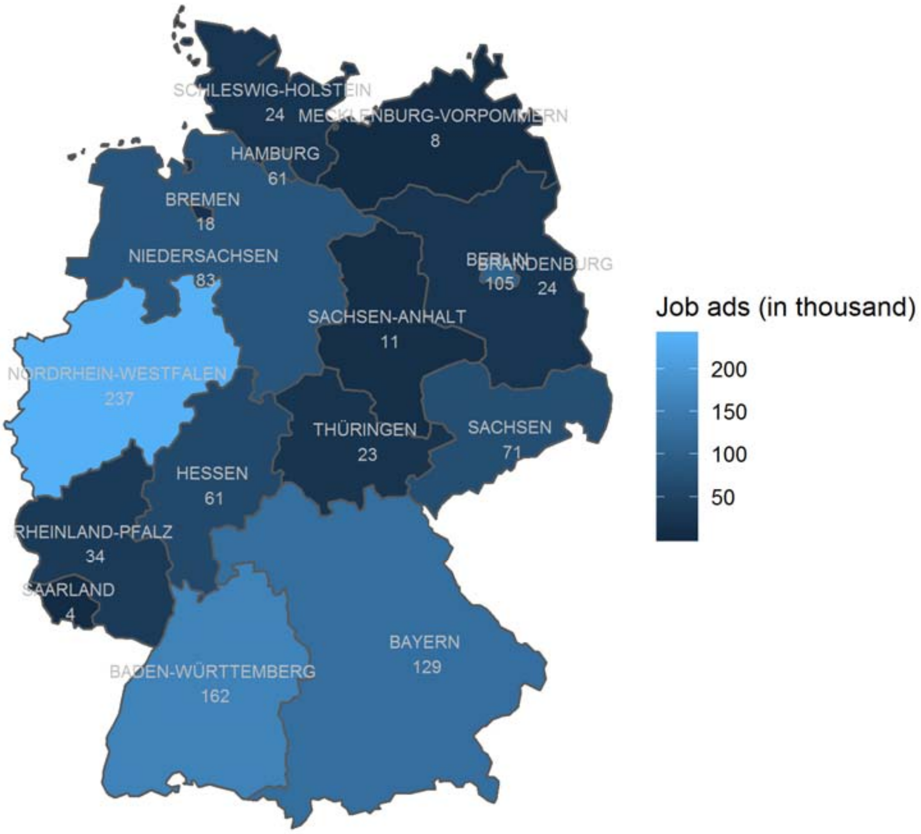
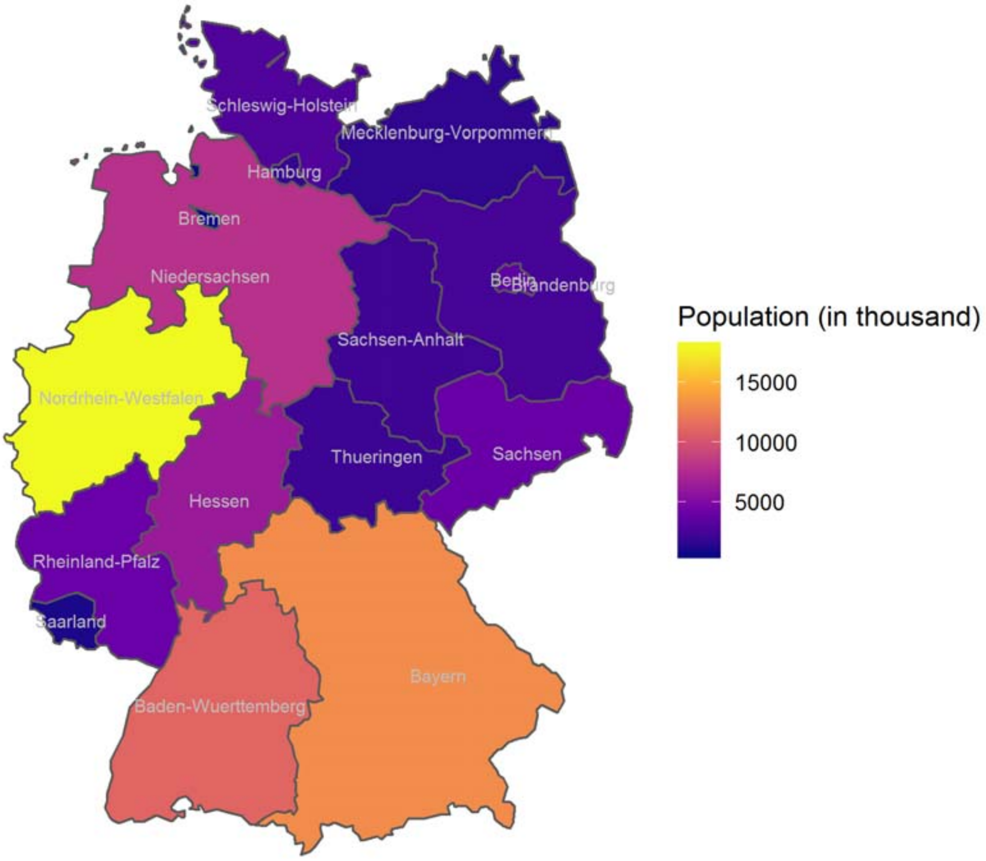


Figure 11: Map of population numbers in 2018

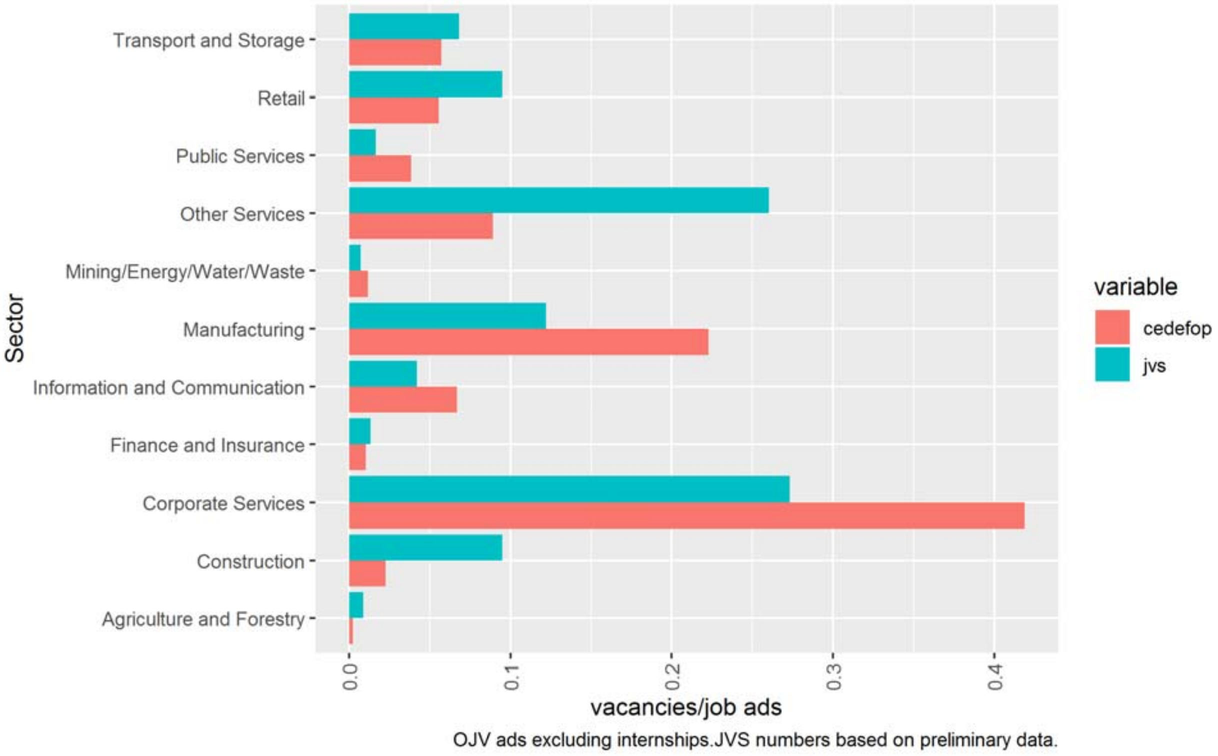


8 Industry sectors

As an additional plausibility check, we compare the relative shares of industry sectors (by NACE categories) in OJA and JVS (Figure 12). Deviations from the industry shares in the JVS can be caused by uneven coverage in the web-scraping process, by sector differences in online job-portal use and by sector specific variations in the efficiency of the CEDEFOP deduplication process.

We see an overrepresentation of the manufacturing, IT and corporate services sectors in the CEDEFOP data, compared with the JVS. On the other hand, retail, construction and other services are under-represented in CEDEFOP. The imbalance between corporate and other services might be caused by differences in classification of specific services between the two datasets.

Figure 12: CEDEFOP-OJA vs. JVS: relative share of job ads by industry sector



OJV ads excluding internships.JVS numbers based on preliminary data.