Journal Title
XX(X):2–23

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Using online job vacancies to predict key labour market indicators

Miroslav Štefánik[1], Štefan Lyócsa[1] and Matúš Bilka[12]

Abstract

We explore data collected as a side product of administering an online job advertisement portal with dominant market coverage in Slovakia. Namely, we process the information on the aggregated, quarterly registered number of online job vacancies. We provide an assessment of the potential of this information in predicting the official vacancy statistics and the development of other indicators relevant in labour market analysis. In analysing the online data and official vacancy statistics' cross-correlation, we found similarities with comparable studies reported for the Netherlands and Italy. Additionally, we assess the online job vacancy data's predictive potential by comparing out-of-sample estimates of three simple linear models. We show that vacancy data are not only attractive in now-casting, but combining it with an auto-regression model, substantially improved the precision of predictions, especially in the case of medium-term predictions (exceeding six months). Besides a favourable performance in improving predictions of the official vacancy statistics, we have also revealed a promising potential of online vacancy data in predicting the number of employed and unemployed persons and the usual working time.

Keywords

Job vacancy statistics, Online job-search, Time series, Predictive modelling, Employment

Introduction

As online activities penetrate our everyday activities, more big data is being collected in behind. The presence of this data tempts scientists in various fields. This paper explores data collected as a side product of administering an online job advertisement portal—namely, the aggregated, quarterly registered number of online job vacancies (OJV). We aim to assess its ability to capture and predict the development of the official job vacancy statistics together with selected key labour market indicators.

We see our contribution mainly in assessing the predictive potential of OJV data in a specific modelling exercise. Additionally, also in extending the attention from predicting official job vacancy statistics to predict other, labour-market relevant indicators, such as employment, unemployment, working time or the GDP. Finally, we join an emerging stream of case studies bringing evidence on the correlation of the OJV data to the official job vacancy statistics.

The following section provides an overview of studies dealing with online data and specifically OJV data. Explored data and the comparison strategy are described in the third section. The fourth section presents a more detailed comparison of OJV data to the official job vacancy statistics. The fifth section provides an assessment of the predictive potential of OJV data through a modelling exercise. We conclude in the final, sixth section.

Online data in labour market analysis

With the Internet penetrating our everyday life, new, big-data sources emerge to be explored for potential use (Askitas and Zimmermann, 2015). The development of data storage capacities enabled the emerging of the concept of big-data[1], potentially driving a paradigm

[1]Institute of Economic Research of the Slovak Academy of Sciences, Šancova 56, 811 05 Bratislava, Slovakia
[2]Faculty of National Economy, University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia

Corresponding author:
Miroslav Štefánik, Institute of Economic Research of the Slovak Academy of Sciences, Šancova 56, Bratislava, Slovakia, http://www.ekonom.sav.sk
Email: miroslav.stefanik@savba.sk

Prepared using sagej.cls

shift in social science (Hitzler and Janowicz, 2010). Such development motivates towards exploring potential uses of big data in social sciences Taylor et al. (2014). Specific attention in this respect is paid to big data's potential in substituting for information collected through the surveys of official statistics Struijs et al. (2014). With respect to studies covering internet data, together with Hooley et al. (2012), we distinguish those covering the Internet and those using the Internet to conduct research. An example of the later, present web-based surveys such as Glassdoor or the WageIndicator, designed to collect wage information from internet users[2]. Research presented here classifies as the first type; we explore Internet data collected as a side product of the Internet's everyday operation.

In this research line, the pioneering stream of studies explored trends in online search data to forecast (or now-cast) labour market development. For example, Choi and Varian (2012) or Schmidt and Vosen (2013) predict the development of the economic cycle. For predicting unemployment, online search data are explored at the European level by Tuhkuri (2016), at the country level by Askitas and Zimmermann (2009), Fondeur and Karamé (2013) and specifically in the context of the COVID-19 pandemic by Caperna et al. (2020).

Another, more recent, stream of studies utilises data collected by social networks. For example, Twitter data became popular because of their high frequency and rich content (Barberá and Rivero, 2015), (Blank, 2017), (Rafail, 2018). Antenucci et al. (2014) use it to create indexes of job search, job loss and job posting in the US, with references to the Beveridge curve positioning. An overview of the studies employing non-vacancy data for labour market analysis was prepared by Lenaerts et al. (2016).

With the expansion of platform work, Internet data gain additional importance in labour market analysis. Data from online platforms were used to create the Online Labour Index as a measure of online labour demand[3]. Moreover, this data's complexity offers the opportunity to perform advanced structural analysis, for example, in terms of skills demanded online (Stephany, 2020). In this aspect, online platform data compare to online job vacancy data.

Online job vacancy data

Online job vacancy (OJV) data are created as a side-product of an online job search. Although commercial providers dominate the OJV market (Cedefop, 2019), examples of data exported for research purposes are becoming more numerous[4]. From the perspective of labour market analysis, the OJV data present an even richer source of information than the data acquired from online search (e.g. Google Trends) or social networks (e.g. Twitter), because they document a substantial share of the hiring process. Additionally, Kuhn (2014) shows that the importance of the Internet in the job search is increasing in time, suggesting an increasing relevance of OJV data for labour market analysis.

The first issue of interest in exploring the OJV data is the question of their representativeness, addressed by multiple earlier studies. Kureková et al. (2015) provide an overview of studies exploring OJV data listing their strategies to assess or increase the representativeness of their data. Although each OJV data source is specific, some common features can be identified. For example, jobs in the public sector or traditional occupations[5] are often underrepresented. Available studies try to assess the representativeness against the structure of employment known from the official statistics. Revealed shortcomings are, in some cases, addressed by observations weighting. Alternatively, other studies pick a labour market segment, which is less problematic from the point of representativeness (e.g. Fabo et al. (2017)).

The issue of representativeness is less concern at the level of analysing aggregate trends in OJV data. This approach was taken, for example, by De Pedraza et al. (2019) or Lovaglio et al. (2020), which are exploring the potential of country-level OJV data in substituting for and predicting the official vacancy statistics. They both use data from commercial OJV providers dominating in their countries and found a strong correlation between OJV and vacancy statistics in time. Additionally, Lovaglio et al. (2020) performs the analysis at the level of economic sectors, pointing at sectors where OJV perform better than in others.

In this context, we present an additional case study exploring country-specific OJV data's potential in predicting the official

statistics indicators. We report evidence comparable to De Pedraza et al. (2019) and Lovaglio et al. (2020). In comparison to them, we go beyond comparing time series components and correlations and assess the predictive potential of OJV data by comparing the precision of out-of-sample predictions employing OJV data. Moreover, we explore the prediction potential of vacancy statistics and additional relevant indicators of the official statistics[6].

The comparison strategy

Our objective is to provide an assessment of the association and predictive potential of OJV data. Since our OJV data only offer evidence on one country (Slovakia), we aim for comparisons with similar empirical *"case studies"*; namely the De Pedraza et al. (2019) providing evidence on the Netherlands and Lovaglio et al. (2020) providing evidence on Italy. To the best of our knowledge, these are the only two studies exploring the predictive potential of aggregated OJV data. Both are limiting their attention to the comparison of OJV data to the job vacancy statistics (JVS) produced by Eurostat[7]. In this paper, additionally, we explore the association and predictive potential of OJV data with selected key labour market indicators, namely: the gross domestic product[8] (GDP), the number of employed persons (EMPL), the number of unemployed persons (UNE) and the average usual working time (HOURS). The last three key labour market indicators are acquired from the Labour Force Survey (LFS). All four of them are used with quarterly periodicity.

Our observation period's start is restricted by the provided OJV data time series, starting with the first quarter of 2010. The end of our observation period is the third quarter of 2020. The OJV data are available instantly, while the JVS and LFS based indicators are published with a delay of approximately 10 weeks[9]. Thus, our observation period involves a relatively homogeneous period of steady post-2009 economic-crisis development disrupted by a massive shock caused by the COVID-19 Pandemics. The impact of the COVID-19 Pandemics can be spotted at the very end of the observation period, starting with the second quarter of 2020.

Following both of the studies mentioned above, we first decompose the time series of interest to its three main components, the trend-cycle (T-C), seasonal (S) and irregular (I) components. This is done following the classical literature on time-series decomposition[10] implemented through the STL function in R[11]. We report correlation coefficients with OJV data for all time series (raw and differentiated) and each of their components. This provides a first-hand assessment of the association in the development of the number of OJV and other indicators of interest.

In the second step, we graphically explore the auto-correlation function, together with the cross-correlation (with OJV data) function for each of the concerned indicators. Both functions are plotted through variable lags and reported for raw and differentiated data, enabling references to the two empirical case studies (Dutch and Italian). Together with the two earlier case studies, we hope this approach will reveal potential repeating patterns in the development of the time series and help identify a prediction model's best specification. Unlike De Pedraza et al. (2019) we do not report results from the cross-spectral density and squared coherency since these were inconclusive[12].

Before exploring similarity in the development of the time series, we test for stationarity. Stationarity, represented by unit root, would mean the need to change the data structure before proceeding to the analysis (logarithmic or differentiating operations). For these purposes, we used the ADF-GLS test, which is a variant[13] of the modified Dickey-Fuller test for a unit root, suitable in cases where the variable is assumed to have a non-zero mean or to exhibit a linear trend. We expect all of the considered time series to have a non-zero mean[14]. Results of the ADF-GLS test suggests that there is an unpredictable systematic pattern within the time series used. Therefore, to draw more robust results, the differentiation of the data is necessary (see Table 2).

Finally, we assess the OJV time series as a potential predictor of the indicators of interest. Our assessment is based on a comparison of predictions acquired from thee models. Model 0 is a simple auto-regression model, predicting the value of the indicator of interest (Y_t)

based on its lagged value Y_{t-l}, with lags (l) ranging from 1 to 4. All models are complemented by a set of seasonal dummies $S_{Q_{(1,2,3)}}$.

$$Y_t = \alpha + \beta Y_{t-l} + \sigma S_{Q_{(1,2,3)}} + \epsilon \tag{1}$$

In Model 1, the lagged value of the dependent variable is replaced by the OJV data. We use the OJV figure observed for time period of $l + 1$, taking advantage of the earlier availability of the OJV data. The predictive performance of Model 1 provides an idea of the potential of OJV data in now-casting the indicator of interest during the period after the end of the reference period and its publication.

$$Y_t = \alpha + \gamma OJV_{t-(l-1)} + \sigma S_{Q_{(1,2,3)}} + \epsilon \tag{2}$$

Model 2 combines Models 0 and 1. It utilizes both the auto-regressive element as well as the OJV data.

$$Y_t = \alpha + \beta Y_{t-l} + \gamma OJV_{t-(l-1)} + \sigma S_{Q_{(1,2,3)}} + \epsilon \tag{3}$$

We perform an assessment of the predictive potential of OJV data through generating out-of-sample predictions of Y_t[15] for the period of the first quarter of 2013[16] until the end of our observation period (the third quarter of 2020).

The out of sample predictions are assessed based on their root means square errors (RMSE)[17]. We report the RMSEs for Model 0 and the difference of RMSEs of Models 1 and 2 to the RMSE of Model 0 by the four lags considered. The difference of RMSE is expressed in relative terms to the RMSE of Model 0. Thanks to this specification (see equation 4), acquired figures can be interpreted as percentage improvement in the prediction's precision related to the inclusion of OJV data.

$$Improvement = \frac{RMSE_{Model0} - RMSE_{Model1}}{RMSE_{Model0}} \tag{4}$$

Job vacancy data

This section looks specifically at the association between the number of vacancies collected under the system of official, job vacancy statistics (JVS), and a dominant online job advertisement portal in Slovakia OJV data.

In the case of JVS, the number of vacancies presents the stock of jobs vacant at the end of each quarter[18]. Data collection covers employers of all sizes in all the economic sectors[19]. Different countries across the EU use various data collection techniques; in the case of Slovakia, JVS is collected via an electronic reporting system covering all employers with more than 100 employees, a sampling survey collects information from smaller employers.

In the case of OJV, the number of vacancies presents a flow measure with monthly recording, since each vacancy has to be renewed after one month if it remains open for a period preceding one month[20]. A vacancy open through the whole quarter is, therefore, counted three times. Because of the monthly frequency of observation, the OJV quarterly figure is comparable to a sum of three stock observations. In fact, the number of OJV vacancies is three times higher than the number of JVS vacancies. OJV is collected via a dominant commercial provided covering the region of Slovakia; no group of employers is explicitly excluded. Fees related to OJV advertising are not significant. Nevertheless, there are differences in the composition of OJV compared to the JVS in terms of occupational structure or coverage of economic sectors; for example, the health sector or teaching jobs are underrepresented in the OJV data[21].

Comparison of the JVS and OJV time series

We observe the aggregated number of vacancies collected by JVS and OJV between 2010 and 2020[22]. The number of vacancies was, jointly, growing for both indicators (JVS and OJV) until the end of 2019. The subsequent decline was deepened by the impact of the COVID-19 Pandemics (Figure 1).

The decomposition of both time series, the JVS and the OJV confirms the existence of a trend shared by both time series (Figure 2). In line with the Dutch evidence (De Pedraza et al., 2019), but in contrast to the Italian experience (Lovaglio et al., 2020), the

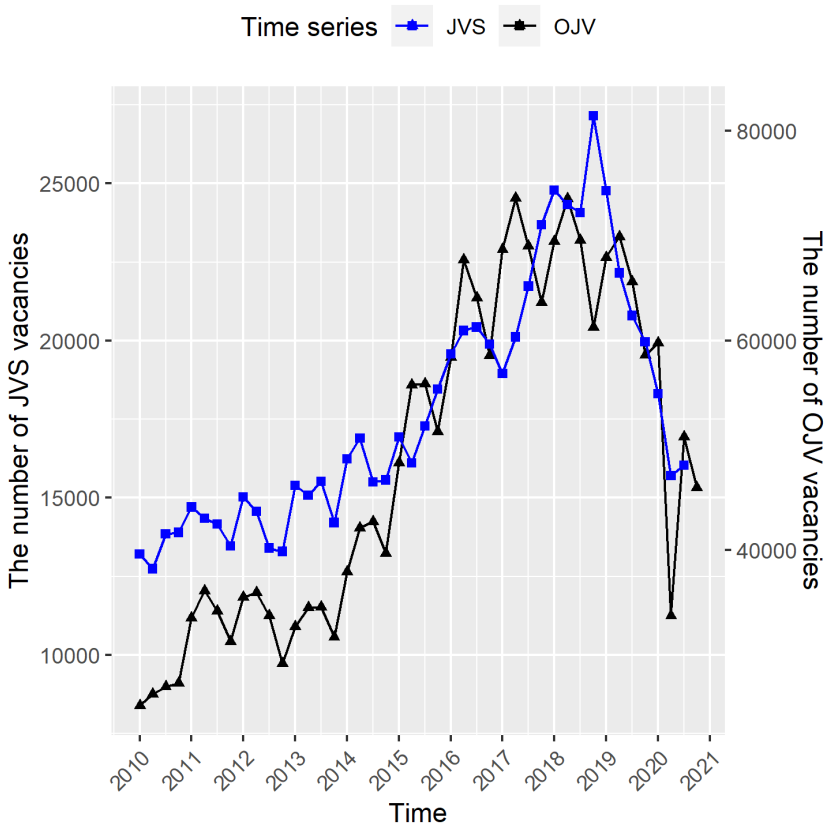

Figure 1. OJV and JVS timeseries

seasonal component of OJV and JVS is not synchronised in our case. While the OJV data show a seasonal peak in the second quarter, the JVS data in the first quarter. In terms of magnitude, the remaining component usually is comparable to the seasonal component[23], except for the substantial hit of the COVID-19 Pandemics at the end of the observation period[24].

The ADF-GLS test speaks clearly in favour of non-stationarity of both of the time series[25] Non-stationarity speaks in favour of using differentiated values of the time series, when building a prediction model. For this reason, we report the following auto-correlation and

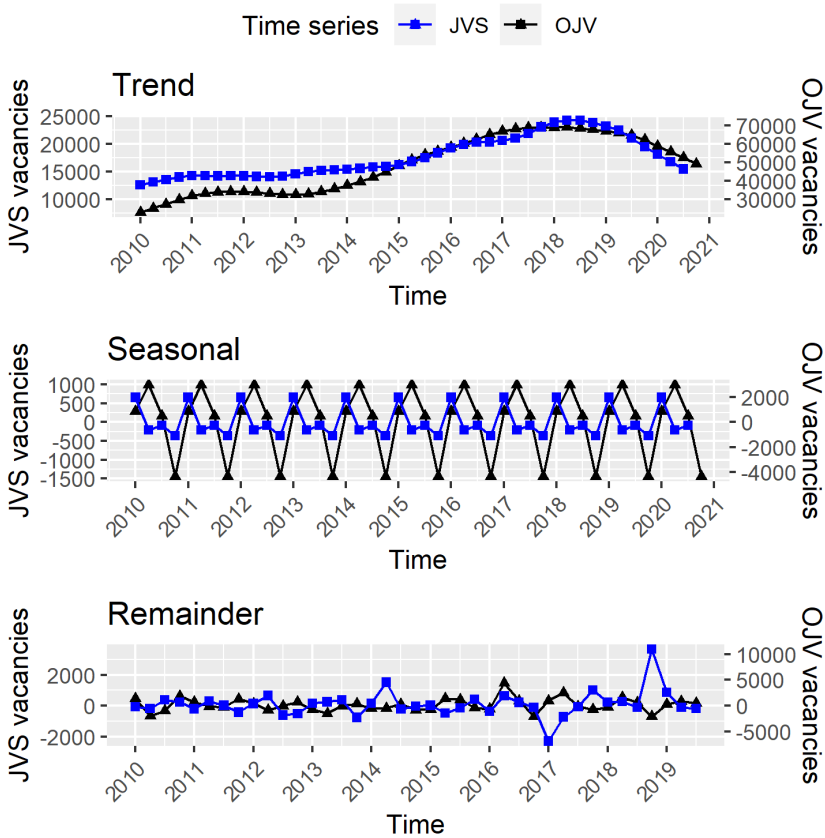

Figure 2. OJV and JVS timeseries - decomposition

cross-correlation patterns of the time series in their original - raw, as well as differentiated forms[26].

In comparison to De Pedraza et al. (2019) we observe auto-correlation in the raw data series for a longer period (Figure 3). In our case, it lasts for nine (OJV) and eight (JVS) quarters. In the case of the Dutch data, it is only for one (OJV) and five (JVS) quarters[27]. In the case of the differentiated time series, our auto-correlation pattern is similar to the Dutch case. Differences are auto-correlated after one year (4 quarters) and, in the case of OJV, also marginally significant after two years. Potentially interesting is a marginally significant, negative auto-correlation coefficient observed only for OJV in the first

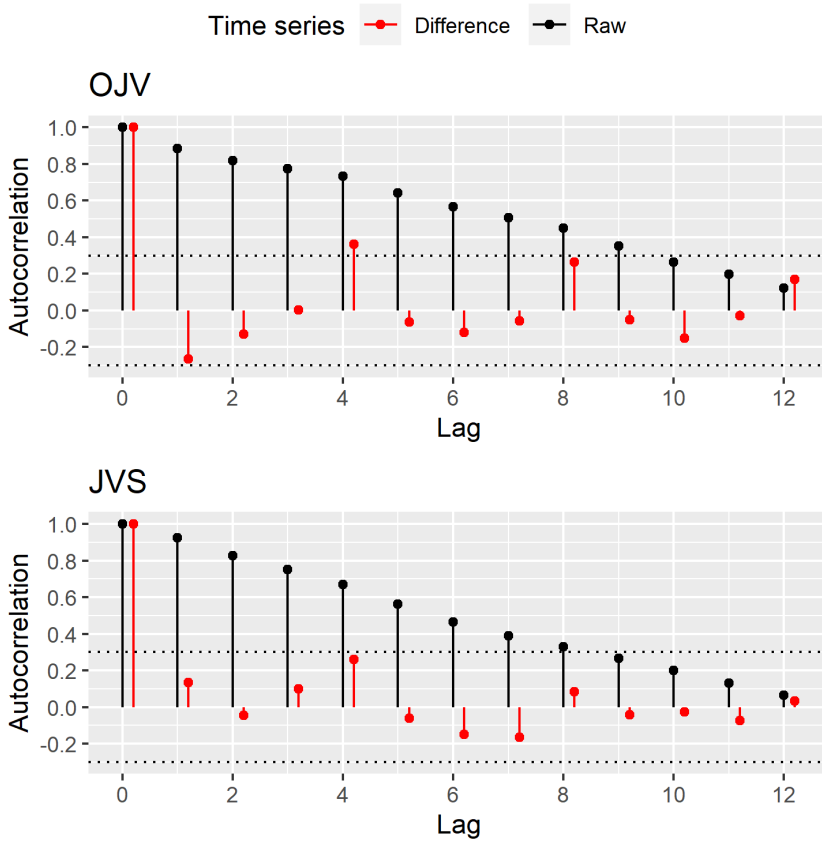

Figure 3. auto-correlation of raw and differentiated time series (JVS and OJV), Lags in quarters

quarter; it was also reported by the Dutch study. It suggests that in the case of OJV data, quarters with higher numbers of vacancies follow after quarters with fewer vacancies.

We reveal the synchronisation of the two time series by their cross-correlation (Figure 4). Alike in De Pedraza et al. (2019), the raw time series show a strong cross-correlation centred around lag 0. The significance of the lag zero suggests that there is positive and synchronised co-movement between the time series. In other words, none of the time series is leading the other. Nevertheless, we bear in mind that the OJV data are available almost one quarter ahead of the

JVS information about the same reference period. This enables now-casting the most recent JVS observations with the OJV observation. Moreover, the lagged OJV data show strong cross-correlation up, at least four lags (one year). We will explore the predictive potential of this cross-correlation in the later text.

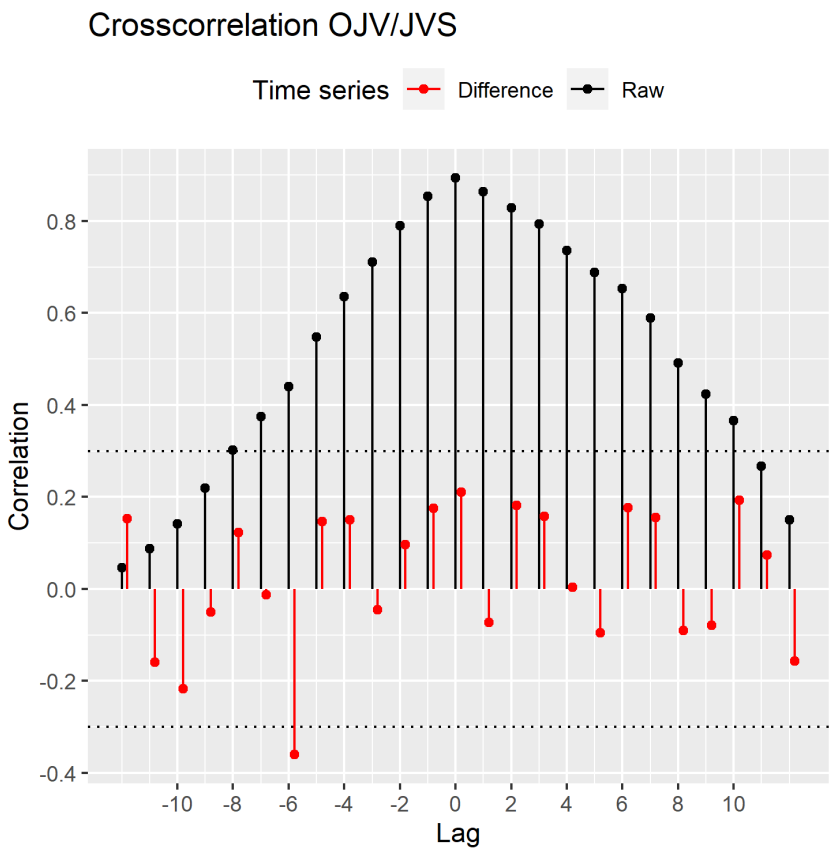

Figure 4. cross-correlation of raw and differentiated time series (OJV lagged)

Similarly to the Dutch case[28], the differentiated time series do not appear to be cross-correlated, with only one coefficient slightly above the 95 percent significance level, suggesting that the differentiated JVS time series six quarters ahead are negatively correlated with the differentiated OJV tie series.

Table 1. Correlation of the indicators used and OJV time-series - (Period: 2010 Q1-2020 Q3), Source: Labour Force Survey (Employment/Unemployment and Working time), National Accounts (GDP)

Time series	JVS	GDP	Employment	Unemployment	Working tim
Raw	0.893	0.451	0.913	-0.848	-0.829
Differentiated	0.21	0.485	0.306	-0.341	0.067
Trend	0.949	0.569	0.968	-0.925	-0.924
Seasonal	0.329	0.826	-0.437	-0.645	-0.503
Remainder	0.075	0.023	0.402	-0.01	0.193

Employing OJV data to predict key labour market indicators

After a more detailed exploration of the association between OJV and JVS time series, we extend our interest to additional selected indicators. In relation to them, we focus on exploring the predictive potential of OJV data. Although the Gross domestic product[29] is not considered as a labour market indicator, its central position in the economic debate makes it interesting also in our context. It is also the only indicator considered here, which is not collected by the Labour Force Survey (LFS). As it originates from the National Accounts data collection, it is also published almost two months earlier[30] than the LFS based indicators (EMPL, UNE and HOURS).

Instead of a graphical display of the association between the key labour market indicators and OJV time series, we report coefficients of correlation with the OJV, separately for raw and differentiated for as well as all the three components for each indicator (Table 1).

The highest correlation coefficient is observed for the trend component of the number of employed persons [31], followed by the already explored JVS and negative correlation coefficients observed for the number of unemployed persons and the usual working time.

Before proceeding to the assessment of the predictive potential of OJV data towards these policy-relevant indicators, we test for stationarity of the considered time series (Table 2). Since none of the test values was below the critical value (-3.19 at 5 percent significance level), we conclude that none of the tested indicators was stationary. For this reason, we report summary statistics of predictive models estimated on raw and differentiated time series.

Table 2. Test statistics of ADF-GLS tests

Timeseries	Test statistic
OJV	-1.1627
JVS	-1.5996
UPSVAR	-2.2845
Employment	-1.8416
Unemployment	-2.6725
GDP	-0.0844
Hours worked	-0.4381

Finally, we assess the prediction potential of OJV data in relation to the indicators of interest. The assessment is based on a comparison of out-of-sample predictions yielded from three simple models. Models are designed to capture the situation of a modeller during the period when OJV data are already available, but the indicator of interest (collected under the official statistics) is not published yet[32]. For this reason, we use as predictors either a lagged value of the predicted indicator or the one-period-more recent OJV value ($lag - 1$). We compare the prediction errors (in terms of RMSE) against an auto-regression model and look at the improvement of predictive performance after adding the OJV figures to the model.

In predicting the JVS indicator, we can see that the model relying only on OJV data (Model 1) predicts OJV development one quarter ahead with a 36.7 percent higher error, in comparison to the auto-regression model (Model 0) (Table 3). Negative RMSE improvement turns to positive figures, in the case of Model 1, since lag 2, suggesting that replacing lagged values of the dependent variable (JVS) with OJV data in yields a more precise prediction when predicting two and more lags ahead. When predicting four quarters ahead, replacing with the OJV data lead to a 25.7 percent decline in RMSE. The observed decline is driven mostly by a drop in the precision of the auto-regression related to the prolongation of the prediction period (see the absolute RMSE values for Model 0 in Table 3).

Model 2 employs both the auto-regression element, as well as OJV data. The assessment of its prediction precision is even more favourable. An improvement of 14.2 percent against Model 0 is

Table 3. RMSE improvement after using OJV to predict JVS (RMSE difference to Model 0)

	Lag 1	Lag 2	Lag 3	Lag 4
Model 1	-0.367	0.074	0.204	0.257
Model 2	0.142	0.172	0.251	0.319
Model 1 (diff)	-0.018	0.071	0.108	-0.034
Model 2 (diff)	0.060	0.111	0.145	0.043
RMSE of Model 0	1222.619	1825.942	2155.406	2421.409
RMSE of Model 0 (diff)	1221.176	1287.411	1272.740	1236.893

The second indicator of interest is the GDP. In this indicator's case, the improvement of the prediction precision linked to including OJV data is less favourable. Replacing the auto-regression element by OJV data leads to a substantial decline (see Model 1 in Table 4) and adding the OJV data (Model 2) only to a less substantial improvement in the prediction precision.

Table 4. RMSE improvement after using OJV to predict GDP (RMSE difference to Model 0)

	Lag 1	Lag 2	Lag 3	Lag 4
Model 1	-2.894	-1.937	-1.086	-0.677
Model 2	0.075	0.053	0.080	0.106
Model 1 (diff)	0.044	-0.012	0.010	-0.254
Model 2 (diff)	0.128	0.032	0.038	0.045
RMSE of Model 0	0.628	0.711	0.924	1.073
RMSE of Model 0 (diff)	0.634	0.641	0.664	0.507

observed when predicting one quarter ahead; it grows gradually to 31.9 percent when predicting four quarters ahead.

Employment was the indicator showing the highest correlation with OJV data. It is also reflected in the improvement of prediction precision (Table 5). The pattern observable is similar to the JVS, with Model 1 showing a loss in precision when predicting one period ahead but a substantial gain (39.5 percent) when predicting 4 quarters ahead. Adding the OJV data together with the lagged dependent variable pays even more (Model 2). In the case of Slovak data, adding OJV contributed to a substantial increase in the precision of employment predictions.

Table 5. RMSE improvement after using OJV to predict the number of employed persons (RMSE difference to Model 0)

	Lag 1	Lag 2	Lag 3	Lag 4
Model 1	-1.393	0.062	0.317	0.395
Model 2	0.341	0.355	0.417	0.440
Model 1 (diff)	0.200	0.048	0.111	-0.013
Model 2 (diff)	0.229	0.063	0.139	0.062
RMSE of Model 0	15.352	21.944	24.831	28.653
RMSE of Model 0 (diff)	14.949	15.372	15.134	13.524

Table 6. RMSE improvement after using OJV to predict the number of unemployed persons (RMSE difference to Model 0)

	Lag 1	Lag 2	Lag 3	Lag 4
Model 1	-4.225	-1.107	-0.360	0.066
Model 2	0.347	0.314	0.368	0.412
Model 1 (diff)	0.088	0.026	0.082	0.022
Model 2 (diff)	0.136	0.053	0.073	0.090
RMSE of Model 0	9.152	14.200	17.703	22.031
RMSE of Model 0 (diff)	8.639	9.378	8.999	8.191

Although highly correlated, in the case of unemployment, predicting only with OJV data (Model 1) leads to a lower improvement in the prediction precision (Table 6); the auto-regression model (Model 0) beats Model 1 in the first three lags. Nevertheless, adding OJV data together with an auto-regression element, leads to a gain in precision, comparable to predicting employment. Interestingly, the OJV data seem to fare less poorly in the differentiated form of the models.

The usual weekly working hours do not show as much variability as the previously analysed indicators (Table 7). This results in relatively lower absolute RMSE values of Model 0. Improvement of predictions related to the inclusion of OJV copies the pattern observable for other indicators.

Table 7. RMSE improvement after using OJV to predict usual working time (RMSE difference to Model 0)

	Lag 1	Lag 2	Lag 3	Lag 4
Model 1	-0.954	-0.050	0.128	0.182
Model 2	0.048	0.064	0.173	0.225
Model 1 (diff)	0.114	0.036	0.009	0.016
Model 2 (diff)	0.129	0.071	0.030	0.049
RMSE of Model 0	0.088	0.121	0.159	0.186
RMSE of Model 0 (diff)	0.093	0.092	0.093	0.093

Concluding remarks and discussion

The potential of big data collected while administrating online services is tempting for multiple scientific fields. This paper documents one case study of employing such data to improve potential predictions of the official statistics. We look at online job vacancy data collected while running a job advertisement web portal with dominant market coverage in one country. Our analysis explored time variation in the aggregate number of vacancies collected by the web portal. First, we display the auto-correlation and cross-correlation patterns of the online vacancy data to the official vacancy statistics. Later, we demonstrate the attractiveness of employing this information in predicting key labour market indicators, such as the number of employed or unemployed persons.

We are aware of two studies are exploring comparable data for the Netherlands (De Pedraza et al., 2019) and Italy (Lovaglio et al., 2020). They focused on exploring the now-casting potential of OJV to the official vacancy statistics (JVS). Together with them, we show a strong correlation with a synchronised cross-correlation of OJV and JVS time series. Another stream of earlier studies explored online search data potential in predicting unemployment (Askitas and Zimmermann, 2009; Fondeur and Karamé, 2013; Tuhkuri, 2016; Caperna et al., 2020). We contribute to this literature by exploring OJV data's potential in predicting other than JVS indicators with labour market relevance. Our findings point at the potential of OJV in predicting employment data. Additionally, our comparison strategy shows that while an auto-regression-based approach has some advantages in producing short-run predictions, OJV data can improve

auto-regression-based predictive models, especially when creating predictions in the longer run.

Being aware of the limitations in generalising a case study's experience based on evidence for one country, we stress the European perspective by relying on indicators collected by methodologies applied by Eurostat in practically all European countries. It is reasonable to assume that a substantial coverage of the OJV data source is a necessary precondition for capturing the correlations of OJV data with country-level values of indicators collected by the official statistics. Especially for larger countries, instead of using data provided by one particular job advertisement provider, data from online job vacancy aggregators are at hand (such as in the one explored by the Dutch case study (De Pedraza et al., 2019)).

Acknowledgements

We would like to thank a private company Profesia ltd. for providing us the data on the number of online job vacancies registered in their system.

Author Biographies

Matúš Bilka graduated from the MSc. Finance at the University of Groningen in the Netherlands and also a master's degree in Finance at the University of Economics in Bratislava. He is currently a PhD. student of the University of Economics in Bratislava with a focus on the pharmaceutical market in Slovakia. In addition, he participates in an international research project within the Slovak Academy of Sciences - Institute of Economics, where he works as a researcher.

Miroslav Štefánik is a senior researcher at the Slovak Academy of Sciences. He focuses on processing big data (web based and administrative) to produce policy-relevant information, such as in program impact evaluations or forecasting. Among the topics covered are active labour market and training program evaluation as well as analysing and predicting future labour market development.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publica-tion of this article.

Funding

This research is a result of the Diskow project, supported by the European Commission under the Erasmus+ programme (http://www.diskow.eu) and was also supported by the Slovak Research and Development Agency under the contract no. APVV-17-0329.

Supplemental material

Supplemental material for this article is available online, at: www.lmevidence.sav.sk

Notes

1. Among many definitions of Big Data, the one using their V-characteristics: volume, velocity, variety, value, veracity appears rather informative. (Hitzler and Janowicz, 2010)
2. For an overview of these studies, see the fifth section of Lenaerts et al. (2016).
3. http://ilabour.oii.ox.ac.uk/online-labour-index/
4. Data scrapping also is an option, in the case of acquiring OJV data from some web providers.
5. Such as, e.g. medical doctors, priests or teachers.
6. Namely, the GDP, employment, unemployment, and working time.
7. https://ec.europa.eu/eurostat/cache/metadata/en/jvs_esms.htm
8. Gross domestic product at market prices expressed as an index (with implicit deflator), 2015=100, euro.
9. https://ec.europa.eu/eurostat/news/release-calendar
10. See Maravall (1985)
11. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/
12. http://web.stanford.edu/class/earthsys214/notes/series.html
13. The difference from the classic Augmented Dickey-Fuller (ADF) test is that the de-meaning or de-trending of the variable is done using the Generalized Least Squares (GLS) procedure. This gives a test a greater power than the standard Dickey-Fuller approach.
14. Tests are conducted with the maximum lag number 4. As the critical value for the 10 percent alpha is -2.89, we can not reject the unit root's null hypothesis for any of the tested indicators.
15. The indicators considered are: JVS, GDP, employment, unemployment and the usual number of working hours.

16. Since our observation period starts in the first quarter of 2010, our "earliest" predictions were based on models estimated to $12 - l$ observations. While the "latest" set of models was estimated at $43 - l$
17. Relying on the *rsme* function in the *Metrics* package in R. See: https://www.rdocumentation.org/packages/Metrics
18. Downloaded from Eurostat Database, table:[jvs_q_nace2].
19. https://ec.europa.eu/eurostat/cache/metadata/EN/jvs_esqrs_sk.htm
20. Data provided by a private company Profesia a.s. administrating a commercial job advertisement portal at https://www.profesia.sk/.
21. For a more detailed analysis related to this source, please visit, (Štefánik, 2012), or on OJV data in general (Kureková et al., 2015).
22. In the case of JVS only until the third quarter of 2020.
23. In line with the pattern observed by De Pedraza et al. (2019).
24. The year 2020 was cut off the time series of the remaining component because of extreme values, to keep Figure 2 readable.
25. With AFD-GLS test values -1.163 for OJV and -1.6 for JVS against the critical value for the 10 percent alpha is -2.89 and suggests the unit-root presence in the raw time-series. Non-stationarity is also not rejected by the ADF-GLS test for any other indicator used.
26. Differences present the annual change: the number of vacancies in quarter t minus the number of vacancies in the same quarter of the previous year (t-4).
27. Visit Figure 5 in De Pedraza et al. (2019) for a comparison (OJV=WEB and JVS=NSO)
28. De Pedraza et al. (2019) page 15.: cross-correlation is less clear within the differentiated data
29. Eurostat Database [namq_10_gdp]
30. First quarterly values of GDP are published approximately four weeks after the end of the reference period; the LFS based indicators, after approximately eleven weeks.
31. Even slightly higher than the one observed for OJV, already explored in the previous section and included here for the sake of comparison.
32. In the case of the LFS-based indicators this period presents a dominant part (approximately ten weeks) of the quarter following the reference period.

References

Antenucci D, Cafarella M, Levenstein M, Ré C and Shapiro M (2014) Using Social Media to Measure Labor Market Flows. *National Bureau of Economic Research* DOI:10.3386/w20010. URL http://econprediction.eecs.umich.edu/.

Askitas N and Zimmermann KF (2009) Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly* 55(2): 107–120. DOI:10.3790/aeq.55.2.107.

Askitas N and Zimmermann KF (2015) International journal of manpower the internet as a data source for advancement in social sciences. *International Journal of Manpower* 36(1): 2–12. DOI: 10.1108/IJM-02-2015-0029.

Barberá P and Rivero G (2015) Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review* 33(6): 712–729. DOI:10.1177/0894439314558836. URL http://journals.sagepub.com/doi/10.1177/0894439314558836.

Blank G (2017) The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review* 35(6): 679–697. DOI:10.1177/0894439316671698. URL http://journals.sagepub.com/doi/10.1177/0894439316671698.

Caperna G, Colagrossi M, Geraci A and Mazzarella G (2020) Googling Unemployment During the Pandemic: Inference and Nowcast Using Search Data. *SSRN Electronic Journal* DOI:10.2139/ssrn.3627754.

Cedefop (2019) The online job vacancy market in the EU Driving forces and emerging trends DOI:10.2801/16675. URL www.cedefop.europa.eu.

Choi H and Varian H (2012) Predicting the Present with Google Trends. *Economic Record* 88(SUPPL.1): 2–9. DOI:10.1111/j.1475-4932.2012.00809.x. URL http://doi.wiley.com/10.1111/j.1475-4932.2012.00809.x.

De Pedraza P, Visintin S, Tijdens K and Kismihók G (2019) Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data. *IZA Journal of Labor Economics* 8(1): 1–23. DOI:10.2478/izajole-2019-0004.

Fabo B, Beblavý M and Lenaerts K (2017) The importance of foreign language skills in the labour markets of Central and Eastern Europe: assessment based on data from online job portals. *Empirica* 44(3): 487–508. DOI:10.1007/s10663-017-9374-6. URL https://link.springer.com/article/10.1007/s10663-017-9374-6.

Fondeur Y and Karamé F (2013) Can Google data help predict French youth unemployment? *Economic Modelling* 30(1): 117–125. DOI: 10.1016/j.econmod.2012.07.017.

Hitzler P and Janowicz K (2010) Linked data, big data, and the 4th paradigm editorial. URL http://www.w3.org/TR/rdf-primer/.

Hooley T, Marriott J and Wellens J (2012) *What is online research?: using the Internet for Social Science research.* Bloomsbury Academic. DOI:http://dx.doi.org/10.5040/9781849665544. URL http://hdl.handle.net/10545/247782.

Kuhn P (2014) The internet as a labor market matchmaker. *IZA World of Labor* DOI:10.15185/izawol.18. URL http://ftp.iza.org/dp5955.pdf.

Kureková LM, Beblavý M and Thum-Thysen A (2015) Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics* 4(1). DOI:10.1186/s40172-015-0034-4. URL http://dx.doi.org/10.1186/s40172-015-0034-4.

Lenaerts K, Beblavý M and Fabo B (2016) Prospects for utilisation of non-vacancy Internet data in labour market analysis—an overview. *IZA Journal of Labor Economics* 5(1): 1–18. DOI:10.1186/s40172-016-0042-z. URL http://www.

Lovaglio PG, Mezzanzanica M and Colombo E (2020) Comparing time series characteristics of official and web job vacancy data. *Quality and Quantity* 54(1): 85–98. DOI:10.1007/s11135-019-00940-3. URL https://doi.org/10.1007/s11135-019-00940-3.

Maravall A (1985) On Structural Time Series Models and the Characterization of Components. *Journal of Business & Economic Statistics* 3(4): 350. DOI:10.2307/1391721.

Rafail P (2018) Nonprobability Sampling and Twitter. *Social Science Computer Review* 36(2): 195–211. DOI:10.1177/0894439317709431. URL http://journals.sagepub.com/doi/10.1177/0894439317709431.

Schmidt T and Vosen S (2013) Using Internet Data to Account for Special Events in Economic Forecasting. *SSRN Electronic Journal* DOI:10.2139/ssrn.2200402. URL https://papers.ssrn.com/abstract=2200402.

Stephany F (2020) Does it Pay off to learn a new skill? Revealing the economic benefits of cross-skilling. DOI:10.2139/ssrn.3717077. URL https://papers.ssrn.com/abstract=3717077.

Struijs P, Braaksma B and Daas PJ (2014) Official statistics and Big Data. *Big Data & Society* 1(1): 205395171453841. DOI:10.1177/2053951714538417. URL http://journals.sagepub.com/doi/10.1177/2053951714538417.

Taylor L, Schroeder R and Meyer E (2014) Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society* 1(2): 205395171453687. DOI:10.1177/2053951714536877. URL http://journals.sagepub.com/doi/10.1177/2053951714536877.

Tuhkuri J (2016) ETLAnow: A Model for Forecasting with Big Data (54). DOI:10.4995/carma2016.2016.4224.

Štefánik M (2012) Internet job search data as a possible source of information on skills demand (with results for slovak university graduates). In: *Building on skills forecasts comparing methods and applications*. Luxembourg: Publications Office of the European Union. ISBN 978-92-896-0892-3. URL http://www.cedefop.europa.eu/EN/Files/5518_en.pdf.