University of Amsterdam
Amsterdam Institute for Advanced labour Studies

Survey vs scraped data: Comparing time series properties of web and survey vacancy data

Pablo de Pedraza, Stefano Visintin, Kea Tijdens and Gábor Kismihók

WP 175
June 2017

AIAS
Working Paper Serie

AIAS works!

Rebalancing labour between market and regulation

June 2017

General contact: aias@uva.nl

Pablo de Pedraza
University of Amsterdam/AIAS and
Joint Research Centre
Unit I.1, Modelling, Indicators & Impact Evaluation
Via E. Fermi 2749, TP 361
Ispra (VA), I-21027, Italy
Tel: +39 0332 783805
pablo.depedraza@ec.europa.es

Stefano Visintin
Universidad Camilo José Cela
Facultad de Derecho y Economía,
Urb. Villafranca del Castillo,
Calle Castillo de Alarcón, 49, 28692
Villanueva de la Cañada, Madrid
svisintin@ucjc.edu

Dr Kea Tijdens
University of Amsterdam/AIAS
Postbus 94025
1090 GA Amsterdam
Netherlands
k.g.tijdens@uva.nl

Dr. Gábor Kismihók
Center of Job Knowledge Research
University of Amsterdam
Amsterdam Business School
Postbus 15953 | 1001 NL Amsterdam
The Netherlands
T +31 20 525 8668
g.kismihok@uva.nl
www.eduworks-network.eu

Bibliographical information

De Pedraza, P., Visintin S., Tijdens, K., Kismihók, G. (2017). Survey vs scraped data: comapring time series properties of web and survey vacancy data. Universiteit van Amsterdam, AIAS Working Paper 175.

This paper can be downloaded from our website uva-aias.net under the section: Publications/ AIAS Working Papers Series.
ISSN online: 2213-4980
ISSN print: 1570-3185

Survey vs scraped data: comparing time series properties of web and survey vacancy data

Pablo de Pedraza

University of Amsterdam, AIAS and Joint Research Centre

Stefano Visintin

Universidad Camilo José Cela,
Facultad de Derecho y Economía

Kea Tijdens

University of Amsterdam, AIAS

Gábor Kismihók

Center of Job Knowledge Research, University of Amsterdam

WP 175

Table of contents

Abstract

This paper studies the relationship between a vacancy population obtained from web crawling and vacancies in the economy inferred by a national statistics office using a traditional scientific method. We compare the time series properties of samples obtained by the Statistics Netherlands and a web scraping company between 2007 and 2014. We find that the web and national statistics office vacancies data present similar time series properties, suggesting that both time series are generated by the same underlying phenomenon: the real number of new vacancies in the economy. We conclude that in our case study web sourced data is able to capture aggregate economic activity in the labour market.

Keywords: web crawling, statistical inference, time series, vacancies

1 Introduction

Big Data is having a major impact on data production and analyses; however, there is uncertainty about whether it serves as a basis for credible science. Enthusiasm about Big Data has mainly come from the commercial sector, which is motivated by profit, rather than from social scientists, who are motivated by a search for the 'truth' (Lagoze 2014). The most popular characterizations of Big Data, the V-definitions (Laney 2001; Janowicz 2010), are not directly applicable to scientific research. Nevertheless, regarding the notions of Velocity, Variety, Volatility, Validity and Veracity, the latter two do have specific and fundamental connotations for the scientific community. They not only refer to the amount of noise, bias or accuracy of the data, but also to the data generation process and method. National Statistics Offices (NSOs) use scientific data generation processes to provide information on many important aspects of society according to scientific standards. At the European level, quality norms have been codified in a Statistics Code of Practice (Eurostat 2014). NSOs have fed social science research for many years and, in the context of Big Data, they can make a contribution, with their focus on quality, transparency and sound methodology, and can provide advice on the quality and validity of information from various types of Big Data sources (Struijs et al. 2014). In this paper, we use the term 'Big Data' for very large datasets collected from online environments, which cannot be directly managed using traditional statistical data management toolsets.

The literature has pointed out: 1. there is a need for specific case studies to address and identify emerging practices around managing data quality and validity, thus contributing to the prospects of Big Data in the social sciences (Taylor, Schroeder and Meyer 2014); and 2. there is a need to stimulate collaboration between NSOs, Big Data holders and universities (Struijs et al. 2014). This paper is the output of such a collaboration.1

In this paper, we propose to focus on a specific case study using data from an NSO to benchmark a very large dataset collected from the internet, with the aim of shedding light on the relationship between the population collected online and the population at large as inferred by traditional scientific methods. More specifically, we focus on the number of vacancies in the economy inferred by a statistical office compared to the number of vacancies obtained from web crawling.

In economics research, labour markets are among the areas in which Big Data is increasingly being used (Choi and Varian 2012; Askitas 2009; Artola and Galan 2012; Antenucci et al. 2014; Kureková, Beblavý and Thum-Thysen 2015; Lenaerts et al. 2016). Posting vacancies, through which employers look for workers, and workers look and apply for jobs, is increasingly being conducted online, producing large quantities of information on the matching processes as a byproduct.

The number of vacancies posted is an important indicator of the state of the economy, more specifically, of the state of labour demand. The number of vacancies is extensively used in applied economics, for example to estimate the matching function (Pissarides 2000, 2011, 2013; Petrongolo and Pissarides 2001; Pedraza, Tijdens and Visintin 2016) or to draw the Beveridge curve (Pissarides 2013), both of which are cornerstones of macroeconomic models. Many statistical offices obtain a figure for the total number of vacancies in the economy by surveying a probabilistic sample of

1 http://www.eduworks-network.eu/

employers. Statistics Netherlands (CBS) conducts this kind of survey for every quarter and infers the quarterly numbers of vacancies at aggregate and industry levels. On the company side, those interested in commercially exploiting labour market information can scrape vacancies posted on the internet (Rothwell 2014). This data can be parsed (Barnichon 2010) and aggregated (e.g. quarterly) to make it comparable to the NSO number of vacancies. There is increasing interest from National Statistics Institutes in exploring Big data and particularly web job advertisements that are considered to be a good starting point for official statistic to tap into Big Data (Swier 2016, Rengers 2016).

In the following, we compare the time series properties of the number of vacancies obtained by the CBS and by a web scraping company between 2007 and 2014, with the aim of benchmarking the two datasets. We approach the time series comparison from two perspectives. First, we test the hypothesis that there is a statistically significant long-term connection between the online and the offline time series. Second, we decompose the observed time series into their three main constituents (trend, seasonal and irregular components) and proceed with an exploratory comparison of the web and the CBS data components.

With respect to the total number of vacancies in the economy, we find that the web and the CBS vacancies data present similar time series properties. Our results suggest that, in both cases, the time series for vacancy data could have been generated by the same underlying phenomenon: the real number of new vacancies appearing in the Dutch labour market every quarter. This demonstrates that web sourced data is able to capture aggregate economic activity.

We consider our exercise to be a quality test of a specific example of a Big Data set, web vacancy data for a specific country, the Netherlands, and a specific time period, 2007-2014. We consider that our approach is scalable to other countries and can be implemented at industry level; however, we do not intend to generalize our conclusions to other similar data sets, countries or periods. The information contained in web data is much richer and granular than the information contained in the official data (Swier 2016), and can be obtained and processed in real time. Tapping into this additional information and good characteristics is not the goal of this paper. Our approach can be seen as a necessary preliminary step, prior to tapping into Velocity and Variety. Despite the promising results, we emphasize the importance of scientifically generated benchmarking data, rather than claiming Big Data as a substitute for Official Statistics (OS).

The remainder of the paper is structured as follows. Section two explains the data generation processes for both data sources. Section three explains our comparison strategy and the time series concepts used. Section four explains the results and in Section five we present our conclusions.

2 Data generation processes

The national statistics office of the Netherlands (CBS) conducts a quarterly postal and telephone survey of employers aiming to measure the number of vacancies at the end of each quarter, as well as the number of new vacancies, the number of vacancies filled and the number of vacancies cancelled during the quarter. For the survey, a stratified random sample of companies and institutions with employees is drawn from approximately 22,000 institutions and companies, of which some 21,000 are private companies and 900 public sector institutions.[2] Population numbers are inferred using firm size weights. The CBS complies with international codes and models and applies sound and scientifically valid statistical methods.[3] We use this data as a benchmark to compare and evaluate the validity and quality of vacancy data crawled from the web.

Web vacancy data is not generated as a result of a scientific method. It is a byproduct of employers' activity on the web, posting vacancy advertisements online in their search for suitable workers. Since 2007, the private company Textkernel[4] has been scraping online vacancies in the Netherlands. Although there is no way to establish whether all online vacancies are being crawled, Textkernel claims to be crawling all websites with vacancies known to them based on their many years of experience, as well as new websites with vacancies. Textkernel is by far the largest vacancy crawling company in the Netherlands, and they provide figures on scraped vacancies to the Netherlands Employment Office, among other organizations. Whether or not the data is comprehensive, it is the result of crawling several hundreds of millions of websites, aiming to capture all the vacancies posted online. Note that by definition Textkernel does not include vacancies not posted online. Moreover, we are unable to estimate the share of online vacancies to all vacancies in the economy. However, in our view, the share of offline vacancies is likely to be small because almost all newspapers with vacancy advertisements also have online editions. Nevertheless, vacancies and recruitment by word of mouth are not captured, and we are unable to estimate its size. Vacancies posted in supermarkets predominantly aim at recruiting domestic staff for private households. However, these are not included in the definition of vacancies by the CBS or Textkernel.

The data generated falls within two Big Data definitions, one by Einav and Levi (2013) and the other by Schroeder (2014), which are commonly accepted by the scientific community. First, it has the four characteristics pointed out by Einav and Levi (2013): i) available in real time, ii) large in size, iii) it contains aspects of labour demand that are difficult to observe using the traditional methods, and iv) it is unstructured. The data also changes the scale and scope of the sources of material available and the tools for manipulation, as signalled by Schroeder (2014). Information contained in job advertisements can be cleaned, structured and aggregated to give it a meaningful structure for our purpose.

Both sources produce different kinds of information about vacancies. While CBS produces information about flows of vacancies (new vacancies emerging during a quarter), the stocks of

2 See for details https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/vacatures-kwartaalenquete

3 Details about data collection method and data quality are available here https://www.cbs.nl/en-gb/our-services/methods

4 http://www.textkernel.com/nl/

vacancies (total number of vacancies that are open at the end of the quarter), the number of vacancies for which a match was found and consequently filled during the period, and the number of vacancies whose selection process was cancelled, Textkernel only identifies new vacancies. For our comparison of the two data sources, we thus concentrated only on the type of information collected by both data sources: the data describing the surge of new vacancies during each period. In other words, the CBS quarterly time series on new vacancies that emerged during each period are compared with the time series of new vacancies posted on the web. Below, we will refer to the former as NSO (National Statistics Office) data and to the latter as WEB data.

For our study, we parsed and structured the WEB vacancy data by quarters, in total, covering seven years (28 quarters). By identifying the date when the vacancy was posted, we can aggregate vacancies from a given quarter and obtain the total number of new vacancies during that period. This transformation allows comparison of aggregates from the WEB data to numbers inferred from the NSO data. This is in line with other curated synopses of huge data sets that have been commonly used to make predictions in the field of 'nowcasting' (Choi and Varian 2012; Askitas 2009; Artola and Galan 2012; Antenucci et al. 2014; Artola et al. 2015). In terms of Tambe's microscope analogy (Taylor, Schroeder and Meyer 2014), we are not tapping into granularity to look at the labour demand organism at a new level of detail. Rather, we are examining whether the organism whose granularity we can observe behaves similarly to the organism we can observe using the scientific method, considering that, when using the latter, we cannot observe the new level of detail in its entirety.

In summary, both data sources are created at the micro level, from which macroeconomic aggregates are obtained. The scientific method builds upon a sampling process from which a random sample is obtained. The web data method builds upon the assumption that a web crawler is able to explore the whole web and scrape information about every vacancy posted online. That information, more specifically, the date the vacancy was posted, has been used to curate and structure the data via aggregation by quarter.

3 Comparison strategy: time series components and multivariate estimates

Our time series properties comparison approach involves two steps. In the first part of our analysis, we test the hypothesis that there is a statistically significant connection between the WEB and the NSO time series. Cointegration analysis supports this step. In the second part, we decompose the observed time series into their three main constituents (trend, seasonal and irregular components) and proceed with an exploratory comparison of the WEB and NSO data components. This twofold methodological approach addresses the main aim of the paper: to produce a comprehensive contrast of the two data sources by combining statistical inference and exploratory analysis.

Cointegration is a statistical property of two time series that share the same underlying trend. We make use of this property to assess whether there is a statistically significant relationship between the NSO and WEB vacancy data. The basic idea behind the cointegration concept is that two non-stationary time series, such as those representing the number of new vacancies in the economy, are cointegrated when we can form a linear combination of them such that this new series is stationary, with fixed means and variances. The stationarity property of the new series allows the time series to exhibit some short-term disruptions, in which the values differ from each other and from the mean; however, they will eventually return to the mean over time.

The interpretation of cointegration in our specific case is as follows: if the WEB and NSO time series are cointegrated, then we can assume that they have been generated by the same long-term phenomenon, namely the real number of vacancies in the Dutch labour market. The rationale behind this interpretation is that if their long-term values are likely to be the same, or at least in equilibrium, each time series must be the result of the sum of the real number of vacancies and a measurement error, the first being a component common to both time series and the second specific to each data production method. Short-term differences between the series might arise as a consequence of the difference between the measurement errors that affect each series randomly to a different extent. However, given that measurement errors are supposed to be temporary, the two time series eventually return to their equilibrium value once the error is resolved over the long term. In turn, the long-term difference in trends (if any) might represent the extent to which each data gathering methodology is capable of covering the real vacancies phenomenon.

In practice, and are the number of new vacancies in period *t* according to the statistics office and the web data respectively. If the combined series [1]

$$aNSO_t + bWEB_t \qquad [1]$$

is stationary, given a constant and real *a* and *b*, then we can affirm that and are cointegrated. For a time series to be considered stationary, all of the roots of the equation producing the observed values, in this case [1], must exceed unity. Therefore, to detect whether a time series is stationary or not, we need to implement a statistical test for the presence of a unit root in an observed sample.

Several statistical tests are proposed in the literature for this purpose. One of the most widely used tests for unit roots is the Augmented Dickey-Fuller (ADF) test (Said and Dickey 1984). It is based on the idea that if a time series is stationary, current levels are proportional to the previous period's

levels and the series is characterized by mean reversion. The ADF tests the hypothesis that the process is a random walk, and therefore not mean reverting. Another widely spread unit root test in the literature is the Philips-Perron (PP) test (Philips and Perron 1988), which shares the same null hypothesis as the ADF test.

To proceed with the tests, it is necessary to estimate the coefficients *a* and *b* of equation [1]. We do this by estimating the coefficient of the following model by linear regression:

Having obtained the coefficients' values, we can compute the linear combination of the two time series [1] and proceed with the ADF and PP unit root tests. If we are able to reject the null hypothesis, we can infer that the two time series are cointegrated. This implies that both the WEB and the NSO time series share the same underlying trend and, according to the interpretation presented above, have been generated by the same phenomenon, namely the real number of vacancies in the Dutch labour market.

If the cointegration analysis is capable of testing the existence of a common underlying trend, it becomes of great interest to estimate, represent and compare the underlying trends beyond the observed time series. Therefore, in the second part of our analysis, we aimed to identify and represent these trends. We proceeded by decomposing the observed time series into their three main constituents: the trend, the seasonal and the irregular components. The visual comparison of each pair of components complements the cointegration analysis and provides useful insight into the similarities and differences between the two time series.

A time series (Y) can be decomposed into three components (Maravall 1985; Findley et al. 1998; Ladiray and Quenneville 2001; Maravall 2005), and it can be assumed that these components are in multiplicative form as in [2].

$$NSO_i = -\frac{b}{a} WEB_i + \varepsilon_i \qquad [2]$$

The Trend-Cycle (TC) component reflects long-term movements and cyclical fluctuations, showing periodic movements over long periods, mainly caused by structural and permanent effects on data periods that are generally longer than one year, without the noise produced by periodic movements or short-term shocks. In quarterly data, the TC component usually shows periodic movements related to periods longer than the year. The TC component of labour market-related quarterly time series may reflect, for example, the effect of the economic cycle on the labour market (when trends over periods longer than one year are identified) or changes in the potential growth of the economy (when trends over several years/periods are identified).

The Seasonal (S) component represents the volume and direction of the cyclical movements repeated within one quarter and cancelled out over the year, caused mainly by economic seasonality and habits, for example, the effect of weekends on the weekly number of hours worked. In our case, we expect to observe quarterly seasonal effects on the number of vacancies in the labour market. These are due to the impact of sectors with strong seasonal cycles on the demand for labour. Agriculture and tourism are two examples of sectors with a strong seasonal component that might create seasonal peaks and low points in the demand for labour. Similar seasonal components in the NSO and WEB time series would confirm the hypothesis that both series have been generated by the same underlying phenomenon.

Finally, the Irregular (I) component represents extraordinary movements caused by random events. In particular, it catches all the time series movement not due to the two other components. These can be produced by measurement errors in the case of the NSO data, or by failures of the scraping software to cover the entire spectrum of the labour market (e.g. if a big player changes name and the software does not immediately take this into account). Extraordinary events, such as a government intervention subsidizing appointments during a specific period of time, are also reflected here.

There are several time series decomposition methods (e.g. see Wei 2006 for an introduction). However, they can be classified into two families: parametric methods based on model-based filters and non-parametric methods based on ad-hoc filters. In both cases, the three components mentioned above can be presented graphically for visual analysis.

We decomposed the NSO and WEB time series (into seasonal effects, trends and irregular terms) using the non-parametric technique of LOESS smoothing, proceeding through a two-step strategy. In the first step, we estimated the seasonal effects (in our case quarterly effects) on the time series. These were computed as the coefficients (plus the constant term) of a regression of the series on quarterly dummies. As a result of this exercise, we split the original series into two components, the quarterly effects and a 'corrected' series unconditioned by these effects. In the second step, the corrected series was further decomposed by means of non-parametric techniques into the *TC* and *I* components. We applied the LOESS methodology (local regression technique) at this stage. This consists in considering each point of the time series subsequently and fitting a low-degree polynomial using weighted least squares, giving more weight to points near the moment at which the response is being estimated, on a subset of the time series. The value of the regression function for each point is thus obtained from the estimated local polynomial (for more details on this technique see Cleveland et al. 1990). The process described can easily be implemented in various statistical packages.[5]

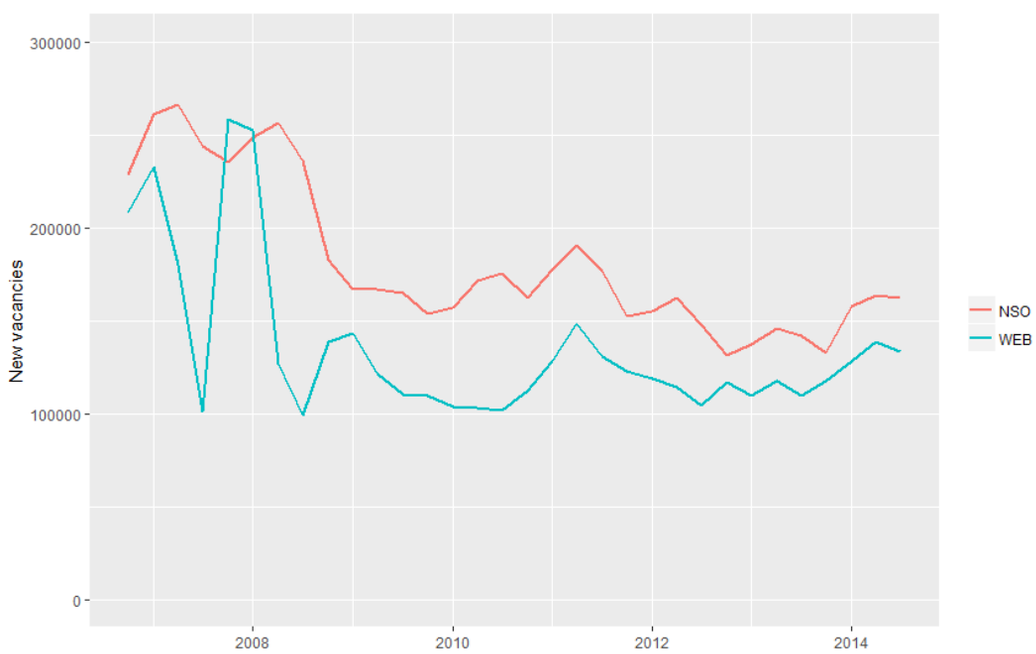
5 Seasonal decomposition can be implemented within the R-project package via the stl built-in function.

4 Results

Figure 1 represents the NSO and WEB time series. Quarterly data covering the period from the end of 2006 (Q4) to the third quarter (Q3) of 2014 are presented. Several points can be made on the basis of a visual inspection of the two time series:

- First, the similar behaviour of the two time series over time is worth noting. For example, both series maximums are reached at the beginning of the period considered and both series present local peaks in 2011 Q2 and 2014 Q2.
- The WEB series variance for the 2007-2009 period is markedly higher than the NSO variance. Given that at the beginning of the web data collection, the methodology used to scrape the web and collect information was still in development, it is reasonable to think that the evidence obtained could be affected by errors (which, in turn, would result in high variance values). However, from 2009 onwards the variance of the two time series is comparable.
- The difference between the two series is glaring: the NSO time series presents higher values over the entire period observed (except in 2007 Q3 and 2008 Q1).

Figure 1 Quarterly NSO (red) and WEB (blue) data on new vacancies, 2007 Q1–2014 Q3



Although the first impression obtained observing the two time series suggests that there could be a long-term relationship between the two, we statistically evaluated the existence of this long-term relationship by testing the series for cointegration. To this end, we constructed the spread according to [1] and then tested it for a unit root.

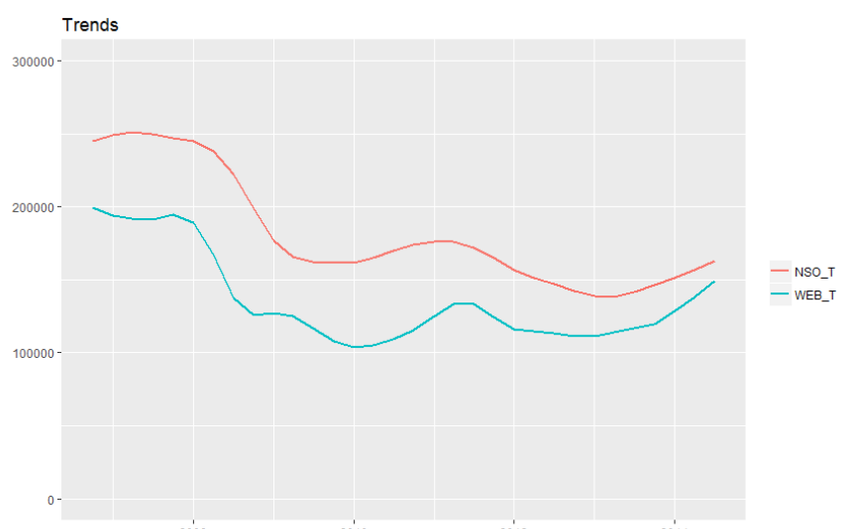
Table 1 Spread regression and unit root test results

Spread regression			Unit root tests	
Dep.	NSO		ADF test	PP test
coeff. -b/a	0.743	test statistic	-4.94	-21.457
		p-value	> 0.01	0.0198
R^2	0.946			

Table 1 presents the main results of the spread regression exercise and both the statistics and the p-values for each unit root test run: the ADF and the PP tests. In both cases, we can clearly reject the null hypothesis that the spread is non-stationary. The probability that we could have observed these spread values, given the assumption that the spread is non-stationary are very low. In other words, this result implies that there is a statistically significant relationship between the WEB and the NSO vacancies time series. The cointegration analysis results suggest that the NSO and the WEB time series share the same underlying trend, which, in turn, can be interpreted as an indication that they may have been generated by the same phenomenon: the actual number of real vacancies arising in the Dutch labour market each quarter.

It is therefore of great interest at this point to decompose the original time series into their *TC*, *S* and *I* components, as explained in Section 3. In particular, a comparison of the *TC* components obtained from the NSO and WEB data is at the centre of this second part of our analysis. Through this comparison we can visually identify differences and similarities between the two time series, once most of the noise that affects the data has been removed. Similar trends support the hypothesis that both time series have been generated by the same underlying process. Furthermore, the visual analysis of the *TC* components allows us to observe long-term differences between the two data gathering methodologies. The comparison of the NSO and the WEB *TC* component time series are presented in Figure 2.

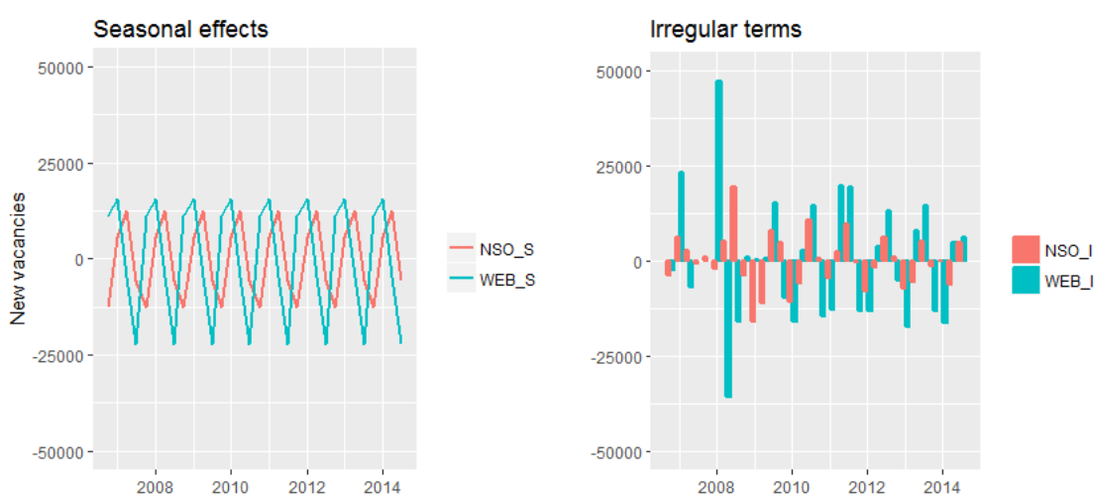
Figure 2 WEB and NSO quarterly vacancies time series decomposition: TC component



The visual inspection of the *TC* component of the NSO and WEB time series confirms the outcomes of the cointegration analysis. The underlying trends of the NSO and WEB data show very similar behaviour over the period observed. Both series show maximums during the 2007-2008 period, before the crisis. Likewise, both series present a noticeable decline in the volume of new vacancies during the 2008-2009 period. As of 2014, neither of the series had recovered to the pre-crisis levels, although both show the same long-term cyclical behaviour, with a peak around the beginning of 2011 and one at the end of the period observed. This visual inspection confirms the hypothesis that both series have been generated by the same phenomenon: new vacancies emerging in the Dutch labour market each quarter.

It is also worth mentioning that the NSO *TC* component is constantly higher than the WEB *TC*, although this difference seems to steadily decay over time. One hypothesis in line with this behaviour is that, given the proven methodology implemented by national statistics offices, the NSO data covers the whole spectrum of new vacancies in the Dutch labour market, while the web data collection only identifies a large fraction of it. Nevertheless, the WEB produced data seems to draw closer to the NSO over time, as internet penetration continues to evolve.

Figure 3 WEB and NSO quarterly vacancies time series decomposition: Seasonal



Seasonal effects also provide important information, as shown in Figure 3. First, they are of similar magnitude, although slightly higher in the case of the WEB series, which might capture vacancies produced in sectors affected by seasonality to a greater extent, such as hotel and restaurant vacancies. The difference between the maximum and minimum peaks accounts for more than 20% of the time series average in the case of the WEB series, and less than 15% in the case of the NSO series, reflecting a lower variance of the latter. Second, it is also interesting to observe how the WEB vacancies exhibit a seasonal maximum during the first quarter of the year (Q1), while the NSO vacancies reach the seasonal maximum during the second quarter (Q2). Similar behaviour is observed in the case of the seasonal minimum, which occurs during the third quarter (Q3) of the WEB data and in the fourth (Q4) of the NSO data.

A further confirmation of a relationship between the two time series is given by the observation of the irregular terms. They are also of similar magnitude, apart from the high values registered in the WEB series at the beginning of the data collection period. They represent approximately no more than 6-8% of the trend values. It is also interesting to note that since 2009 the *I* components of both series share the same sign in each quarter observed.

5 Conclusions

A comprehensive contrast of web and national statistics office vacancies data revealed that they present similar time series properties. This result was obtained through cointegration analysis of the corresponding vacancies time series and by visually comparing the underlying trends in the series obtained through a non-parametric decomposition method. The cointegration analysis suggested that, although the two time series might show different behaviour over a short period of time, the web sourced vacancies and the vacancies data produced by the CBS represent similar underlying trends in the long term. This was visually confirmed by the decomposition of the time series. The components of the two trends were found to behave similarly over a seven year period, reflecting a similar impact of the economic crisis and business cycles.

It is also worth noting that the CBS data produced a higher value of vacancies over the entire period, indicating greater coverage of the phenomenon. However, the gap between the two trends decreased over time, suggesting the web series was catching up. Seasonal effects and irregular terms perceivably confirmed the strong relationship maintained by the web and the CBS vacancies time series.

These results suggest that the web and CBS time series vacancy data may have been generated by the same underlying phenomenon: the real number of new vacancies appearing in the Dutch labour market each quarter. This demonstrates that web-sourced data is able to capture aggregate economic activity. Finally, both the CBS and the web data allow for breakdowns by industry. Our future research agenda includes exploring how the comparison above might apply to each sector of the economy.

References

Antenucci, D., Cafarella, M., Levenstein, M. C., Ré, C., Shapito, M. D. (2014). Using social media to measure labor market flows. *NBER Working Papers Series* no. 20010. http://www-personal.umich.edu/~shapiro/papers/LaborFlowsSocialMedia.pdf

Artola, C. and Galan, E. (2012). Tracking the future of the web: construction of leading indicators using internet searches. *Banco de España, Documentos Ocasionales Nº1203.*http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf

Artola, C., Pinto, F., and Pedraza, P. de (2015). Can Internet Searches Forecast Tourism Inflows? *International Journal of Manpower,* vol. 36, 1: 103-116.

Askitas, N. and Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *IZA Discussion Paper No. 4201,* June 2009.

Barnichon, R. (2010). Building a composite Help Wanted Index. *Economic Letters,* 109: 175-178.

Choi and Variant (2012). Predicting the Present with Google Trends. *The Economic Record, vol. 88,* Special Issue, June, 2012: 2-9.

Eurostat (2011). European Statistics Code of Practice: Revised edition 2011, ISBN: 978-92-79-21679-4, see the link http://goo.gl/Z0xArw

Eurostat (2015). ESS Guidelines on Seasonal Adjustment, Luxembourg, European Communities. ISBN 978-92-79-45176-8 [http://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf]

Einav, L. and Levi, J. D. (2013). The Data Revolution and Economic Analyses. *NBER Economic Papers Series*, Paper 19035 http://www.nber.org/papers/w19035

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B.-C. (1998). New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program. *Journal of Business and Economic Statistics*, 16: 127-177.

Hitzler, P. and Janowicz, K. (2010). Linked data, Big data and the 4th Paradigm. *Semantic Web* 0 (0) 1. IOS Press. http://www.semantic-web-journal.net/system/files/swj488.pdf

Kureková, L. M., 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: 18. DOI 10.1186/s40172-015-0034-4

Ladiray, D. and Quenneville, B. (2001). *Seasonal Adjustment with the X-11 Method*, Springer: New York.

Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data & Society*, July-December: 1-11.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. In Meta Group. Available at: (accessed 30 June 2016). http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

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

Maravall, A. (1985). On Structural Time Series Models and the Characterization of Components. *Journal of Business & Economic Statistics*, American Statistical Association, vol. 3(4): 350-355.

Pedraza, P. de, Tijdens, K., and Visintin, S. (2016). The role of the short-term employed in the matching process before and after the crisis: Empirical evidence from the Netherlands. *AIAS Working Papers* no. 165, December 2016. https://aias.s3-eu-central-1.amazonaws.com/website/uploads/1490258513430WP-165-1---de-Pedraza,-Tijdens,-Visintin.pdf

Pissarides, C. A. (2000). *Equilibrium Unemployment Theory,* MIT second ed.

Pissarides, C. A. (2011). Equilibrium in the Labour Market with Search Frictions. *American Economic Review,* 101 (June): 1092-1105.

Pissarides, C. A. (2013). Unemployment in the Great Recession. *Economica,* 80: 380-403.

Petrongolo, B. and Pissarides, C.A. (2001). Looking into the black box: A survey of the matching function. *Journal of Economic Literature,* vol. XXXIX (June): 390-431.

Rengers, M. (2016) Big Data in Official Statistics: Estimation of job vacancies by using web scraping techniques. European Conference on quality in Official Statistics 2016, 03-06-2016

Rothwell, J. (2014). Still Searching: Job Vacancies and STEM Skills. Metropolitan Policy Program at Brookings, July 2014.

http://www.brookings.edu/research/interactives/2014/job-vacancies-and-stem-skills#/M10420

Schroeder, R. (2014). Big Data: Towards a More Scientific Social Science and Humanities? In In: Graham M and Dutton WH (eds) *Society and the Internet*, *How Networks of Information are Changing our Lives* Oxford University Press, chapter 10, p. 164, DOI:10.1093/acprof:oso/9780199661992.003.0011

Struijs, P., Braaksma, B., Daas, P. J. H. (2014). Official Statistics and Big Data. *Big Data and Society,* April-June: 1-6.

Swier, N. (2016) Webscraping for Job Vacancy Statistics. Presented in Big Data ESSNet Workshop. Sofia. 24-25 February 2017 Downloaded last time 18/05/2017 from http://socialstats2016.eu/speaker/mr-nigel-swier

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,* July-December: 1-10.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. 2nd ed. Boston: Pearson.

AIAS Working Paper Series

The AIAS working paper series consists of several publications of AIAS staff and AIAS guests on a wide variety of subjects in the fields of labour economics, sociology of work, labour law, and health and safety.

ISSN online 2213-4980
ISSN print 1570-3185

Information on AIAS

AIAS is an institute for multidisciplinary research and teaching at the University of Amsterdam. Founded in 1998, it brings together the University's expertise in labour studies.

AIAS *research* focuses on the analysis of labour markets, social security incomes, institutions and governance. It combines various approaches from sociology, law, economics, medical sciences and political sciences. Our research programmes include studies on employment relations, labour market issues, inequality, institutions and the welfare state. Many studies take an international comparative perspective and are conducted in co-operation with academic partners in other countries.

AIAS offers tailor-made in-company *courses* in the field of HRM, Inequality and solidarity, labour market development, labour relations etc.

AIAS and its staff contribute to *society* on many subjects, for different audiences and in varying formats (articles, books, reports, interviews, presentations etc.) Next to this 'Working Papers' Series, we also have the 'Industrial Relations in the Netherlands' Series and the 'GINI Discussion Paper' Series which also addresses a great variety of topics.

Annually AIAS organizes conferences about ongoing research and current trends.
Furthermore several (lunch) seminars and workshops take place during the year, offering interesting opportunities for the exchange and deliberation of research on labour issues between researchers from all over the world. AIAS has a major collection of academic socio-economic data in the field of labour relations, labour organizations, employment and working conditions in the Netherlands and abroad.

See for more information our website www.uva-aias.net

University of Amsterdam
Amsterdam Institute for Advanced labour Studies

Postal address

PO Box 94025
1090 GA Amsterdam
The Netherlands

Visiting address

Roetersstraat 31
1018 WB Amsterdam
The Netherlands

+31 20 525 4199
+31 20 525 4301
aias@uva.nl
www.uva-aias.net