de Pedraza, Pablo; Visintin, Stefano; Tijdens, Kea Gartje; Kismihók, Gábor

**Article**

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

IZA Journal of Labor Economics

**Provided in Cooperation with:**
IZA – Institute of Labor Economics

*Suggested Citation:* de Pedraza, Pablo; Visintin, Stefano; Tijdens, Kea Gartje; Kismihók, Gábor (2019) : Survey vs scraped data: Comparing time series properties of web and survey vacancy data, IZA Journal of Labor Economics, ISSN 2193-8997, Sciendo, Warsaw, Vol. 8, Iss. 4, pp. 1-23, https://doi.org/10.2478/izajole-2019-0004

This Version is available at:
https://hdl.handle.net/10419/222150

*Pablo de Pedraza[1]*, Stefano Visintin[2], Kea Tijdens[3] and Gábor Kismihók[4]*

# Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data

## 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 (NSO) using a traditional method. We compare the time series properties of samples obtained between 2007 and 2014 by Statistics Netherlands and by a web scraping company. We find that the web and NSO vacancy 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 are able to capture aggregate economic activity in the labor market.

**Current version:** April 03, 2019
**Keywords:** web crawling, statistical inference, time series, vacancies, Labor demand, data collection
**JEL codes:** J23, J63, C22, C80
**Corresponding author:** Pablo de Pedraza
pablo.depedraza@ec.europa.eu
pablodepedraza@usal.es

1 University of Amsterdam and European Commission, Joint Research Centre (JRC), Unit I.1, Modelling, Indicators & Impact Evaluation, Via E. Fermi 2749, TP 361, Ispra (VA), I-21027, Italy. E-mail: pablo.depedraza@ec.europa.es. Tel: +39(0)332 783805

2 University of Amsterdam/AIAS and Universidad Camilo José Cela, Facultad de Tecnología y Ciencia, Urb. Villafranca del Castillo, Calle Castillo de Alarcón, 49, 28692, Villanueva de la Cañada, Madrid, Spain. E-mail: svisintin@ucjc.edu

3 University of Amsterdam/AIAS, Postbus 94025, 1090 GA Amsterdam, The Netherlands. E-mail: k.g.tijdens@uva.nl

4 Leibniz Information Centre for Science and Technology, Welfengarten 1 B, 30167 Hannover, Germany. E-mail: Gabor.Kismihok@tib.eu. Tel.: +49(0)511 7621 4705


Cite as: Pedraza et al. *IZA Journal of Labor Economics* (2019) 8:4.
https://doi.org/10.2478/izajole-2019-0004

# 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; Hitzler and Janowicz, 2010), may not be 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 refer to the data generation process and method. National Statistics Offices (NSOs) use data generation processes to provide information on many important aspects of society according to traditional scientific standards. At the European level, quality norms have been codified in a Statistics Code of Practice (Eurostat, 2011). 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 data sets collected from online environments, which cannot be directly managed using traditional statistical data management toolsets.

The literature has made clear that first 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 et al., 2014). Second, 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 collaboration.[1]

Literature has evaluated data quality of non-probabilistic online sources such as Twitter (Barbera and Rivero, 2015; Blank, 2017; Rafali, 2018), Wikipedia (Martin, 2018), Google searches (Choi and Variant, 2012; Butler, 2013; Lazer et al., 2014), voluntary web surveys (Pedraza et al., 2010), Open Data Government portals (Sáez Martín et al., 2016), the combination of sources (Revilla et al., 2017), and the resulting new sampling and recruiting modes and methods (Stern et al., 2016; Head et al., 2016; De Leeuw, 2018; Revilla et al., 2015). Conclusions about quality and representativeness are diverse and depend on the context, target population, and research goal. In this paper, we propose to focus on a specific case study using data from an NSO to benchmark a very large data set 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 survey methods by a statistical office compared to the number of vacancies obtained from web crawling.

In economics research, labor markets are among the areas in which Big Data is increasingly being used (Choi and Variant, 2012; Askitas and Zimmermann, 2009; Artola and Galan, 2012; Artola et al., 2015; Antenucci et al., 2014; Kureková et al., 2015; Lenaerts et al., 2016). Posting vacancies, through which employers look for workers, and workers look and apply for

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

jobs, is increasingly being conducted online, producing large quantities of information on the matching processes as a by-product.

The number of vacancies posted is an important indicator of the state of the economy and, more specifically, of the state of labor 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 et al., 2016, 2019) 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 employers. Statistics Netherlands Centraal Bureau voor de Statistiek (CBS) conducts this type of survey each quarter and infers the quarterly numbers of vacancies at aggregate and industry levels. On the company side, those interested in commercially exploiting labor market information can scrape vacancies posted on the Internet (Rothwell, 2014). The amount of information obtained from the Internet is much more detailed. These data can be parsed (Barnichon, 2010) and aggregated (e.g., quarterly) to make it comparable to the NSO information about vacancies. The main aim of this paper is to compare the vacancy data collected by an NSO, using a survey method, with the vacancy data collected by scraping the Internet.

In the following sections, we compare the time series properties of the number of vacancies obtained between 2007 and 2014 by the CBS and by a web scraping company, with the aim of benchmarking the two data sets. We approach the time series comparison combining preliminary visual inspection, exploratory analysis, and statistical inference. First, 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. Second, we compare their autocorrelation, cross-correlation patterns, and their synchronization by means of cross-spectral analyses.

With respect to the total number of vacancies in the economy, we find that the web and the CBS vacancy 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 labor market every quarter.[2] This suggests that web-sourced data are able to capture aggregate economic activity.

We consider our exercise to be a quality test of a specific example of a Big Data set, the web vacancy data, for a specific country, the Netherlands, and a specific time period, 2007–2014. We realize that our approach is scalable to other countries and can be implemented at industry level, but before generalizing conclusions to other similar data sets, countries, or periods, similar quality tests as the one conducted here would be needed. We do not aim to explore the richness and granularity of the information contained in web data compared to the information contained in the official data. Our approach can be seen as a necessary first step. Prior to tapping into velocity, variety, volume (Laney, 2001), and data granularity, we propose an exploration that focuses on validity and veracity of data and their ability to reflect economic realty. There are several research lines using web vacancy data at other aggregation levels. For example, there are efforts to establish semantics-based bidirectional matching between job descriptions and job vacancies, to improve the quality of job advertisements (Chala et al., 2016). Furthermore,

2 We arrive at this conclusion studying both, original and transformed (first difference), time series.

research covers language, information and communication technology (ICT), and platform economy-related skills analyses in a number of European countries on the basis of vacancy text analysis (Fabo et al., 2017; Lenaerts et al., 2016). In the field of work and organizational psychology, vacancy text analysis is contributing to the better understanding of jobs (job analysis; Kobayashi et al., 2016). Users of Textkernel are, for example, public employment service in the Netherlands; data are also used by headhunters, temp agencies, occupational career websites, among others for predicting the chances to find a job, given a preferred occupation.

Despite the promising results, and the fact that new data sources could either supplement nonresponse or fully abandon costly vacancy surveys, we emphasize the importance of traditionally generated benchmarking data, rather than proclaiming Big Data as a substitute for official statistics (OS).

The remainder of the paper is structured as follows. Section 2 explains the data generation processes for both data sources. Section 3 explains our comparison strategy and the time series concepts used. Section 4 explains the results, and in Section 5 we present our conclusions.

## 2 Data generation processes

The NSO of the Netherlands (CBS) conducts a quarterly postal and telephone survey of employers that aims 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 are public sector institutions.[3] Population numbers are inferred using firm size weights. The CBS complies with international codes and models and applies sound and valid statistical methods.[4] Approximately 80% of companies and institutions respond to the survey. Response rates are stable over the time we study; data are not influenced by nonresponse shocks.[5] We use these data as a benchmark to compare and evaluate the validity and quality of vacancy data crawled from the web.

Web vacancy data are not generated as a result of a traditional scientific method. It is a by-product of employers' activity on the web, posting vacancy advertisements online in their search for suitable workers. Since 2007, the private company Textkernel[6] 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 that are known to them, based on their many years of experience, as well as new websites with vacancies. The Textkernel algorithm is continuously updated and improved. With the exception of an anomaly, due to early stages of the algorithm development in 2008, Textkernel also claims de-duplication of all vacancies collected. This is done by methods of machine learning, where algorithms are trained to identify high degree of similarity between two job

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

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

5 https://opendata.cbs.nl/statline/#/CBS/en/dataset/80474eng/table?ts=1550482026366, see the box Table explanation, accessed 18FEB2019.

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

advertisements. These methods include shingling and min-wise permutations (Jijkoun, 2016; Broder et al., 1997). Textkernel is by far the largest vacancy crawling company in the Netherlands, and it provides figures on scraped vacancies to the Netherlands Employment Services and to other organizations. Whether or not the data are 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, and therefore 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. It is certain that vacancies with recruitment by word of mouth are not captured, nor are we able to estimate its size. Vacancies posted in supermarkets predominantly aim at recruiting domestic staff for private households. These are not included in the definition of vacancies by the CBS or Textkernel.

The data generated fall 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 possesses the four characteristics pointed out by Einav and Levi (2013): (1) available in real time, (2) large in size, (3) it contains aspects of labor demand that are difficult to observe using the traditional methods, and (4) it is unstructured. The data also change the scale and scope of the sources of material available and the tools for manipulation, as signaled by Schroeder (2014). Information contained in job advertisements can be cleaned, used to de-duplicate vacancies posted in more than one site, structured, and aggregated to give it a meaningful structure for our purpose.

The two sources produce different kinds of information about vacancies. Although CBS produces information about flows of vacancies (new vacancies emerging during a quarter), the stocks of 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 canceled, Textkernel only identifies new vacancies. For our comparison of the two data sources, we thus concentrated only on the type of information that is 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 refer to the vacancy time series produced by CBS as NSO data and to the vacancy time series produced by Textkernel scraping the web as web data.

For our study, we parsed and structured the web vacancy data by quarters, covering 8 years in total (31 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 a comparison of aggregates from the web data with 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 Variant, 2012; Askitas and Zimmermann, 2009; Artola and Galan, 2012; Antenucci et al., 2014; Artola et al., 2015). In terms of Tambe's microscope analogy (Taylor et al., 2014), we are not tapping into granularity to look at the labor 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 traditional 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 microlevel, from which macroeconomic aggregates are obtained. The traditional 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 entire web and scrape information about every vacancy posted online. That information, or more specifically, the date that 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 comparison is based on decompositions of specific properties of the observed times series using two complementary methods. First, we performed a well-known decomposition of both time series into their three main constituents (trend, seasonal, and irregular components) and proceeded with an exploratory comparison. Second, we implemented a cross-spectral analysis to infer whether both series display a similar autocorrelation pattern and whether they are correlated and synchronized. This twofold methodological approach provides a comprehensive contrast of the two data sources by combining visual, exploratory analysis, and statistical inference.

The first time series decomposition we performed was first proposed by Maravall (1985) and further developed by, among others, Findley et al. (1998); Ladiray and Quenneville (2001) and Maravall (2005). According to these authors, a time series ($Y$) can be decomposed into three components, and it can be assumed that these components are in multiplicative form as in equation (1).

$$Y = TC \times S \times I \tag{1}$$

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 1 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 labor market-related quarterly time series may reflect, for example, the effect of the economic cycle on the labor market (when trends over periods longer than 1 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 movements repeated within one quarter and canceled out over the year, caused mainly by economic seasonality and habits: for example, the effect of the summer season on the number of hours worked in the tourist sector. In our case, we expected to observe quarterly seasonal effects on the number of vacancies in the labor market. These are due to the impact of sectors with strong seasonal cycles on the demand for labor. Agriculture and tourism are two examples of sectors with a strong seasonal component that might create seasonal peaks and troughs in the demand for labor. 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. 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 labor market (e.g., if a big player changes name and the software does not immediately take this into account). Extraordinary events, such as government intervention subsidizing appointments during a specific period of time, are also reflected here. In this case, both series would be affected similarly.

There are several time series methods to perform the traditional decomposition (e.g., see Wei, 2006 for an introduction). They can be classified into two families: parametric methods based on model-based filters and nonparametric methods based on ad hoc filters. In both cases, the three components can be presented graphically for preliminary visual inspection. We used a combination of both. On the one hand, 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, we split the original series into two components, the quarterly effects and a "corrected" series unconditioned by these effects. On the other hand, the corrected series was further decomposed by means of nonparametric techniques into the TC and *I* components. For this purpose, we applied the locally estimated scatterplot smoothing (LOESS) methodology. It consists of 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 (Cleveland et al., 1990).[7]

After comparing time series components of NSO and web, we implemented a cross-spectral analysis to infer whether the two series display similar autocorrelation patterns and whether they are correlated and synchronized. The cross-spectral analyses approach (and no other approach, such as co-integration) is supported by the results of stationarity (unit root) tests.

Spectral and cross-spectral analyses are frequency domain methods that can be applied to study the relationship between time series. They are commonly used to decompose time series into several periodic components, corresponding to frequency bands or timescales (e.g., low frequencies correspond to the long-term, high frequencies to very short-time periods).

An important difference with the traditional method presented above is that in spectral and cross-spectral analysis timescales are not a priori imposed by the researcher. Structural characteristics and cyclical behavior are extracted from the data itself and identified at different timescales. Periodicity of short-, medium-, and long-term components is endogenous from the analysis (Granger and Hatanaka, 2015; Iacobucci, 2005) and the statistical comparison of frequency bands extracted from the data gives information on the synchronization, or lack of it, between two series (Granger and Hatanaka, 2015).

Decomposition of series into short-, medium-, and long-term frequencies is often exploited by researchers when working with economic variables. For example, frequency domain methods are used to observe the lengths of different business cycles. Cross-spectral analysis has also been employed in economic literature to observe co-movements of economic variables at several frequencies. Jayaram et al. (2009), for example, analyzed business cycle synchronization between India and a set of industrial economies; Fidrmuc et al. (2008) studied the links in

7 The process described can easily be implemented in various statistical packages. Seasonal decomposition can be implemented within the R-project package via the *stl* built-in function.

the business cycles of European economies, India, and China. These techniques have also been used to quantitatively evaluate the relationship between the business cycle and other aspects of the economy. Costas and Eckels (2011) observed the dynamic correlation between the business cycle and tourism in Switzerland. We implemented this approach to test the hypothesis that NSO and web series are correlated and synchronized. The extent of their synchronization at time frequencies that emerged from letting data speak provides valuable information about the extent to which both series are actually measuring the same phenomenon.

We implemented the cross-spectral analysis exploring several parameters: autocorrelation, cross-correlation, cross-spectral density, and squared coherency. We computed autocorrelation to observe the presence of repeating patterns, if any, and to understand whether both series behave similarly regarding the impact of past vacancy values on future ones. Cross-correlation expresses the correlation of the two time series with each other and across time lags: it represents a preliminary attempt to better understand the series' correlation with each other in a frequency domain. The actual cross-spectral analysis begins by measuring the series' cross-spectral density. The observation of their cross-spectral density allows understanding of the frequencies at which the series are more similar. Finally, we computed the squared coherency to identify to which extent the series are correlated at each frequency. Cross-spectral density and squared coherency parameters together make it possible to identify co-movements and synchronization. We can accept that both series move with high synchronicity if, for example, both series present high cross-spectral density at frequencies of one and, simultaneously, they present high squared coherency at the same frequency. The capability to uncover synchronization between time series is one of the reasons cross-spectral analysis is gaining increasing attention in applied economics.

Although the frequency domain analysis can provide a meaningful insight into co-movements of the series, the analysis is subject to sample size limitation. Our information duration may be considered short (31 quarterly data point corresponding to almost 8 years), which may bias cross-spectral estimations. In general in the literature, more data points are used but there are also empirical examples using similar samples sizes to those in this paper (Leon and Eeckels, 2011).

## 4 Results

This section is structured according to the methodology presented above. First, we present the time series decomposition exercise by presenting a graphical depiction of the three main components. Second, we proceed with the cross-spectral analysis in which results are represented graphically. The visual representation of the estimated values was chosen (1) according to the empirical literature cited (e.g., Fidrmuc et al., 2008; Iacobucci, 2005) and (2) to maintain the readability of the paper.

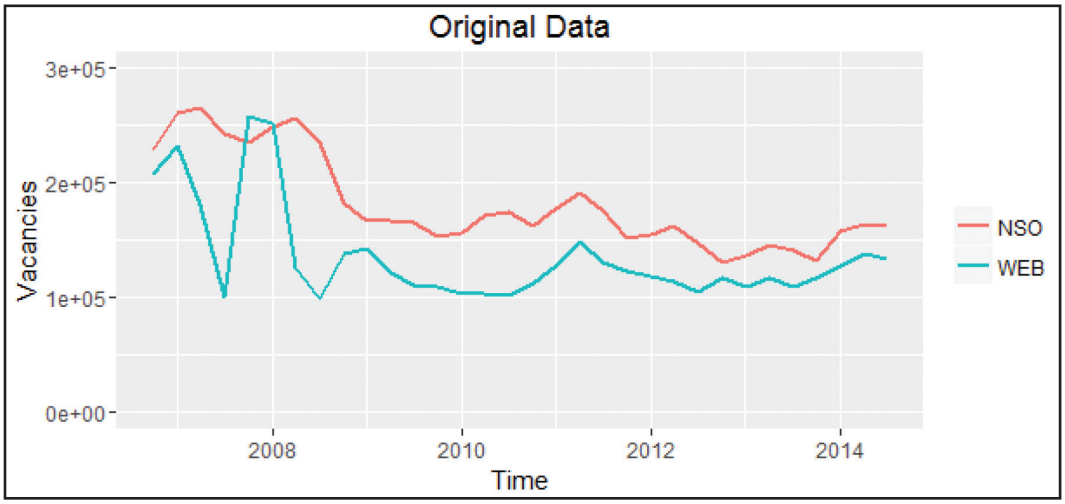
Figure 1 represents the NSO and web time series. Quarterly data covering the period from the end of 2006 (Q4) to the third quarter (Q3) of 2014 are presented. Several points can be made on the basis of a preliminary visual inspection of the two time series:

- First, the similar behavior 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 2011Q2 and 2014Q2.

- 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 onward, 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).

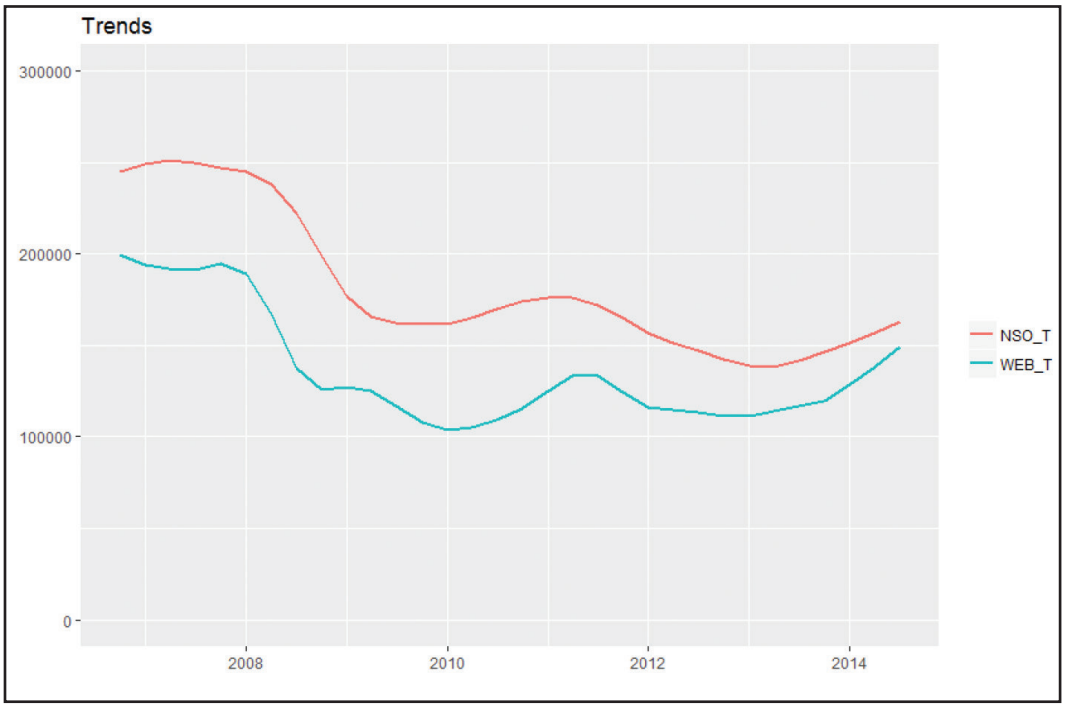
Regarding the decomposition of the original time series into their TC, *S*, and *I* components, Fig. 2 presents the comparison of the NSO and the web TC component. Through

**Figure 1** NSO and web-extracted data on vacancies in the Dutch economy, 2007–2014.




NSO, National Statistics Office

**Figure 2** Web and NSO quarterly vacancies, time series decomposition: TC components.




NSO, National Statistics Office; TC, trend–cycle

the comparison of the TC components, we can visually identify differences and similarities between the two time series, once most of the noise that affects the data has been removed.

The visual inspection of the TC component of the NSO and web time series shows very similar behavior 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 around the mid-2008. As of 2014, neither of the series had recovered to the precrisis levels, although both show the same long-term cyclical behavior, with a peak around the beginning of 2011 and one at the end of the period observed. This visual inspection supports the hypothesis that both series have been generated by the same phenomenon: new vacancies emerging in the Dutch labor 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. Table 1 presents the trend

**Table 1** Web and NSO quarterly vacancies, time series decomposition: TC components

| Time | Web trend | NSO trend | % Difference (decimal form) |
|---|---|---|---|
| 2006Q4 | 199344 | 245027 | 0.19 |
| 2007Q1 | 193759 | 248898 | 0.22 |
| 2007Q2 | 191909 | 251124 | 0.24 |
| 2007Q3 | 191448 | 250038 | 0.23 |
| 2007Q4 | 194738 | 247137 | 0.21 |
| 2008Q1 | 189334 | 244879 | 0.23 |
| 2008Q2 | 166867 | 238482 | 0.30 |
| 2008Q3 | 137499 | 222269 | 0.38 |
| 2008Q4 | 126055 | 199127 | 0.37 |
| 2009Q1 | 126930 | 177155 | 0.28 |
| 2009Q2 | 125408 | 165567 | 0.24 |
| 2009Q3 | 117155 | 162485 | 0.28 |
| 2009Q4 | 108189 | 161601 | 0.33 |
| 2010Q1 | 103809 | 161966 | 0.36 |
| 2010Q2 | 105235 | 164912 | 0.36 |
| 2010Q3 | 109167 | 169850 | 0.36 |
| 2010Q4 | 115662 | 174378 | 0.34 |
| 2011Q1 | 125033 | 176205 | 0.29 |
| 2011Q2 | 133092 | 175901 | 0.24 |
| 2011Q3 | 133149 | 172350 | 0.23 |
| 2011Q4 | 124728 | 165096 | 0.24 |
| 2012Q1 | 116527 | 157101 | 0.26 |
| 2012Q2 | 114696 | 151530 | 0.24 |
| 2012Q3 | 113612 | 147465 | 0.23 |
| 2012Q4 | 111037 | 142652 | 0.22 |
| 2013Q1 | 111303 | 138776 | 0.20 |
| 2013Q2 | 114218 | 138519 | 0.18 |
| 2013Q3 | 117096 | 141902 | 0.17 |
| 2013Q4 | 119665 | 146736 | 0.18 |
| 2014Q1 | 128575 | 151675 | 0.15 |
| 2014Q2 | 138235 | 157164 | 0.12 |
| 2014Q3 | 149538 | 162732 | 0.08 |

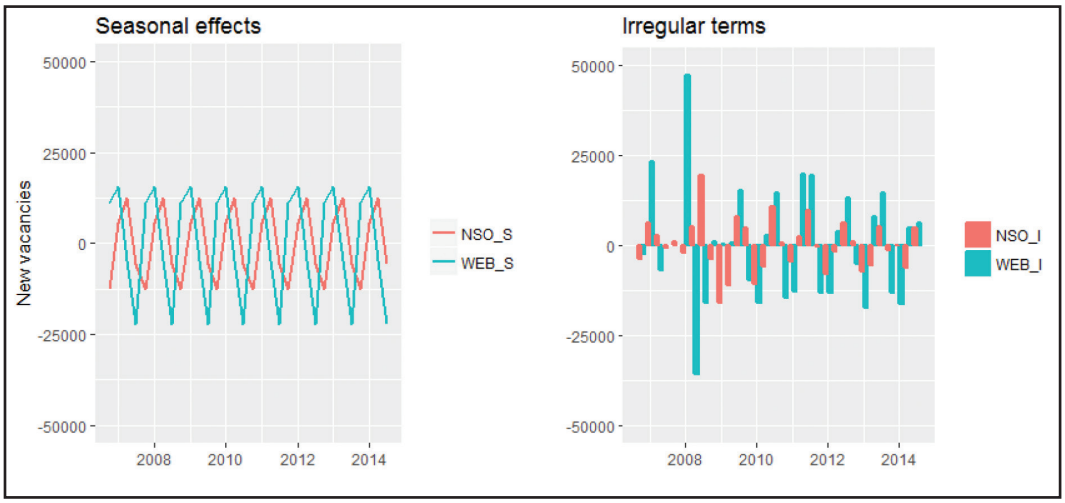
NSO, National Statistics Office; TC, trend–cycle.

component values for the NSO and the web series; in addition, the relative difference between the two is exposed. In the first half of the table, corresponding to the period going from the fourth quarter of 2006 until the fourth quarter of 2010 the NSO values are, on average, 30% higher than the web values. In the second half of the table, corresponding to the period going from the first quarter of 2011 until the third quarter of 2016, the NSO values are, on average, 20% higher than the web values. The last values observed show a percentage difference of around 10%. One hypothesis that could explain this behavior is that, given the proven methodology implemented by NSOs, the NSO data cover the whole spectrum of new vacancies in the Dutch labor market, while the web data collection only identifies a large proportion of it. Nevertheless, the web data seem to draw closer to the NSO over time, as Internet penetration continues to evolve.

The comparison of seasonal (*S*) effects is reported in Fig. 3 where we can observe the following. 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 behavior 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. This might highlight the ability of the web to lead NSO in capturing ups and downs in labor demand. The reason for the lead–lag relationship may be that the web data identify the posting of a vacancy on the first day, whereas NSO data measure vacancies at the day of survey, which is always later.

However, these differences in seasonal patterns (suggesting a lead–lag relationship) are not conclusive at this stage of the analysis. The seasonal decomposition is part of an explanatory approach that aids the understanding of the behavior of the time series and provides some suggestions regarding possible similarities between the series that have to be confirmed by

**Figure 3** Web and NSO quarterly vacancies, time series decomposition: seasonal. NSO, National Statistics Office



other statistical methods. The main limitations so far are given by the fact that the web data are quite erratic at the beginning of the observed period and seasonal patterns might be driven by this fact. Data over longer time spans will be made a clearer view possible.

A further confirmation of a relationship between the two time series is given by the observation of the irregular (*I*) terms. The irregular components have the same signs but the magnitudes are significantly different, especially during the first half of the observed period. In both cases, they show a decreasing pattern. In addition, the web irregular term is larger than the NSO, although the difference is decreasing over time. It is also worth noticing 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. Once more, the web data seem to draw closer to the NSO over time.

Before proceeding with the cross-spectral analyses, we explored whether or not both time series were stationary. If two series presented a clear long-run trend (which implies they are nonstationary), and if this trend was similar, we could explore their long-run relationship by means of co-integration analysis. If two series presented no clear trend over the long-run (which implies that they are stationary), their relationship could be explored by the means of other techniques, such as cross-spectral analysis.

We conducted the following stationarity (unit root) tests: the augmented Dickey–Fuller (ADF) test (Said and Dickey, 1984), a modified ADF test estimating the optimal number of lags, the Phillips and Perron (PP) test (Philips and Perron, 1988) and, finally, the Zivot and Andrews (ZA) test (Zivot and Andrews, 2002). The results are presented in Table 2.

The ADF and PP tests present puzzling results although, when considering all the results, they seem to suggest the nonstationarity of the series. In detail, the modified ADF test does not reject the null hypothesis of nonstationarity in both cases, which allows us to think of the two series as nonstationary. The PP test seems to reject the null hypothesis of nonstationarity in the case of the web data, but not when applied to the NSO time series. Therefore, according to both ADF tests and the PP test, the NSO series is nonstationary while results are not conclusive in the case of the web series.

However, these results are not definite because level shifts might bias ADF and PP results. Both series undergo a leap around the mid-2008 as can be observed in Fig. 1. The visual

**Table 2** Unit root test statistics

| | Unit root test statistics | | | | |
|---|---|---|---|---|---|
| | **ADF test** | **ADF test (modified)** | **PP test** | **ZA test (and critical values)** | |
| Web | −4.6675*** (4) | −1.0229 (3) | −17.177* (2) | −6.1109 ***(6) | 0.01= −5.34 0.05= −4.8 0.1= −4.58 |
| NSO | −2.884 (4) | −1.6402 (3) | −8.0429 (2) | −5.2097**(8) | 0.01= −5.34 0.05= −4.8 0.1= −4.58 |

*Notes*: *, **, and ***Rejection of the null hypothesis at the 10%, 5%, and 1%, respectively. Lags are given in parenthesis. ADF test (modified) estimates the optimal number of lags to use. Truncation lags parameter is given in parenthesis. Potential break position is given in parenthesis. ADF, augmented Dickey–Fuller; NSO, National Statistics Office; PP, Phillips and Perron; ZA, Zivot and Andrews.

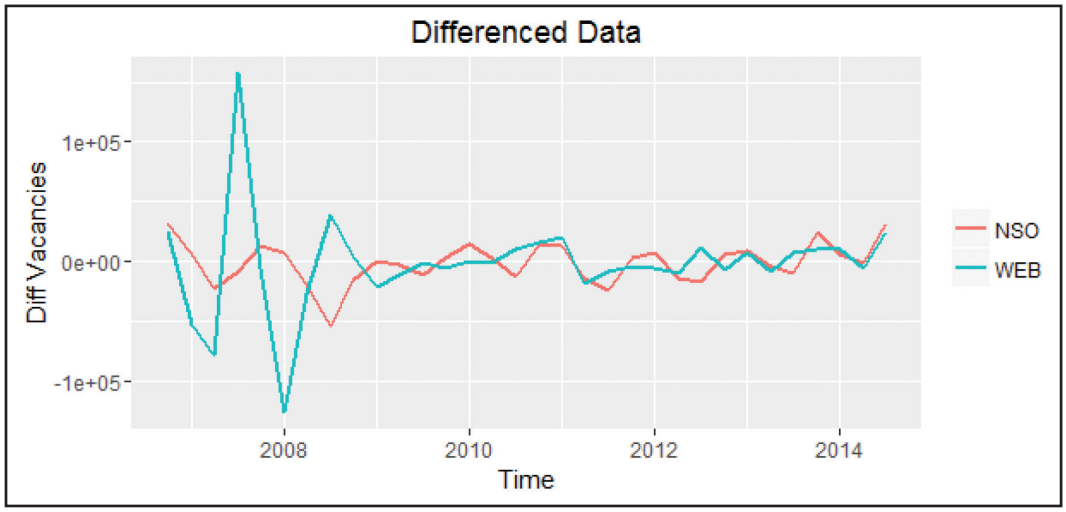
inspection suggests a break in the intercept (a level shift). Due to the presence of this kind of shift, the ADF tests experience power losses (see Cavaliere and Georgiev, 2007 among others).

The ZA unit root test (Zivot and Andrews, 2002) is a unit root test robust to level shifts; it allows a break at an unknown point in the series either in the intercept, in the linear trend, or in both. The ZA also identifies when the level shift takes place. The ZA test results for the web time series confirm that the process is stationary with a level shift (Table 2). Nonstationarity hypothesis can be rejected with more than 99% confidence. The break is identified in the web series at 2008Q1. Similarly, the ZA test confirms that the NSO time series process is stationary with a level shift. We reject the nonstationarity hypothesis with a probability close to 99%. The break is actually identified in the NSO series at 2008Q3. Both breaks appear to occur almost simultaneously with web leading. This is an additional signal indicating that the series could have been produced by the same phenomena, with the web somehow being able to capture shifts earlier than NSO which is in line with the preliminary findings obtained by the observation of the seasonal patterns.

As the ZA test confirmed the stationarity of both series, we proceeded with the spectral analysis[8] on both, the original and a transformed pair of series. The stationarity of the series suggested by the ZA test results allows us to proceed with the spectral analysis directly on the observed values, however, for the sake of robustness (given the uncertainty suggested by the results obtained in ADF and PP), we proceed with some series transformation that further guaranteed the stationarity. We expressly differentiate the two series.

Implementing the analyses on the transformed series as well has several advantages. First, it is a way to guarantee stationarity and, therefore, the appropriateness of the method. Second, it rules out the effect of the observed level shifts. Third, it allows us to observe the relationship between the two series focusing on the changes over time, especially on the direction of the changes of the series. Fourth, since our data are not seasonally adjusted, some of the synchronization might be due to the effects that we already observed in the previous section. By differentiating, seasonal effects are removed. Figure 4 graphically presents the differentiated series.

**Figure 4** NSO and web-extracted, differentiated data on vacancies in the Dutch economy, 2007–2014. NSO, National Statistics Office




8 Given the results obtained by ADF and PP, we also conducted co-integration analyses: results are available upon request. They strongly support the conclusions obtained from cross-spectral analyses.

Therefore, we proceeded measuring autocorrelation, cross-correlation, cross-spectral density, and squared coherency of raw and differentiated data of both series. Figure 5 presents the raw (top row) and differentiated (bottom row) autocorrelograms, that is, correlations of each series with its own past values. The horizontal axis represents the number of lagged time units. It shows that all the series are correlated with their past values. In the very short run (one quarter), all series show a positive autocorrelation. In the case of the original series, a high number of vacancies at a given moment are followed by a high number in the next period and vice versa. However, the effect of past values on present values lasts less than 1 year in the case of the row data and no longer than one quarter for the differentiated series. When observing the differentiated series autocorrelograms, a feeble yearly seasonal effect emerges (this year change is in line with the change that took place the previous year during the same quarter).

Autocorrelations' quicker reduction in differentiated data supports the differentiated series approach. Studying differentiated data, we remove most of the impact of past vacancies on present vacancies and concentrate on the relationship between the two series.

A negative significant autocorrelation appears for differentiated web data in the short run (after two quarters). This implies that if web vacancies steadily grew during a period, we might expect a decrease 6 months later and vice versa. It is interesting to observe that the sign of the autocorrelation of the growth values (differentiated series) is the same in both series (although the magnitude is slightly different). Good semesters (growth) follow bad semesters (decreases) and so on in both series: more intense in the web series (autocorrelation values are higher) and more persistent (autocorrelation values drop to zero more slowly) in the NSO data.

In summary, our series show a positive autocorrelation in the very short run that promptly fades. More interestingly for the purpose of our analysis is that web and NSO data display very similar autocorrelation patterns in both raw and differentiated data.

Cross-correlation is generally considered a proper approach to relate two time series in terms of co-movements. It expresses the correlation of two time series across time lags. We used it as a preliminary step to test the hypothesis that the two series are synchronized by exploring whether they are independent or have a statistically significant relationship.

**Figure 5** Autocorrelation of the original and differentiated series, top and bottom rows, respectively, lags expressed in years

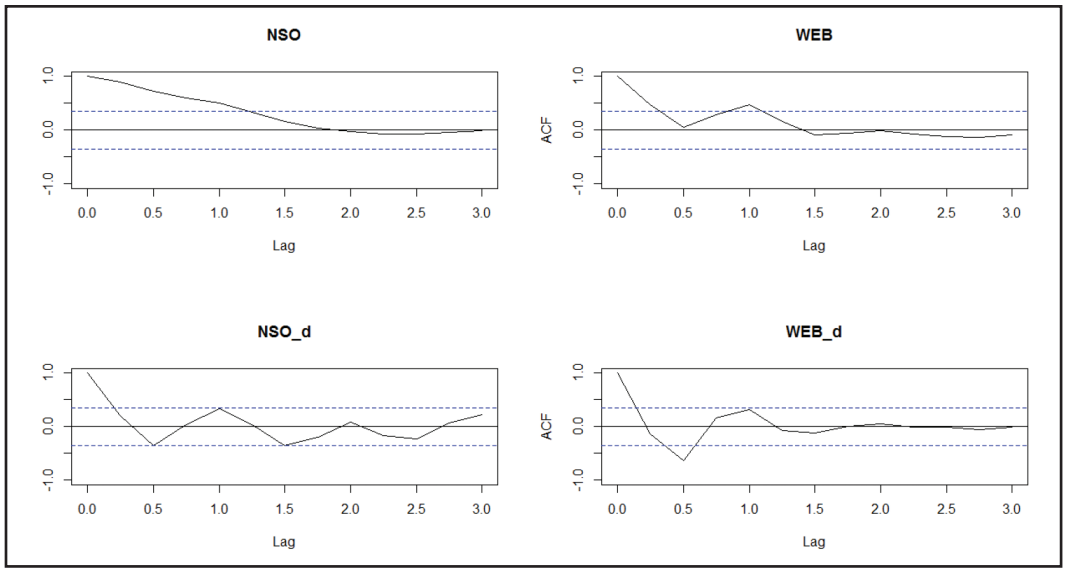


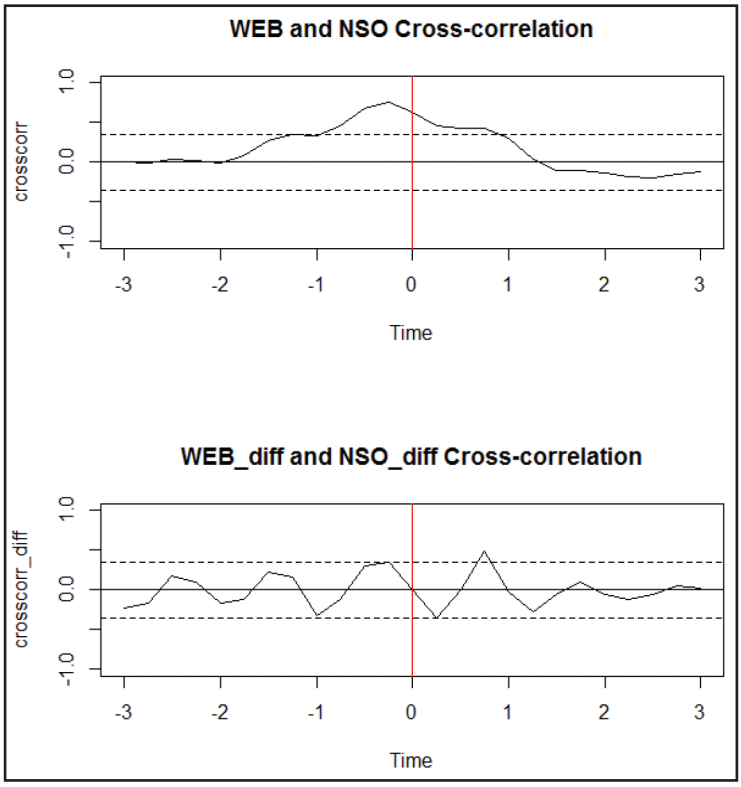
Figure 6 shows the cross-correlograms between the two original series (top) and the differentiated series (bottom), respectively. Cross-correlograms are obtained by estimating the cross-correlation coefficients between the two series. The solid lines describe the coefficient behavior with respect to the different time lags. Intervals of confidence at 95% are also presented (dotted horizontal lines).

Original data show, at time lags close to 0, a positive and statistically significant correlation between the observed series. This behavior suggests a positive, significant, and almost simultaneous co-movement between the series. The fact that the highest observed correlation takes place at lags smaller than zero, and given that we analyzed the cross-correlation of web and NSO in this order, suggests that web data are leading indicator of NSO. Again, although this is in line with some of the results presented above, it is not possible to conclude that there is a lead–lag relationship by the simple study of the cross-correlation function.

When it comes to changes in the series, the analysis on the differentiated series, we observe a less clear cross-correlation between them. On the differentiated series, the cross-correlation is still positive and significant at a time lag of three quarters.[9] The reason behind the lead–lag relationship may be that the vacancy web information is collected in real-time, that is, the first day that it is published online, whereas NSO only produces information once at the end of each quarter.

We can conclude that the hypothesis of independence of the two series is rejected: web and NSO data are not independent, neither in the original form nor in the differentiated one.

**Figure 6** Cross-correlation between NSO and web-extracted data on vacancies in the Dutch economy. Raw data (top) and differentiated data (bottom). NSO, National Statistics Office




9 The fact that the highest observed correlation takes place at lags larger than zero suggests that NSO is the leading series in this case. These results suggest once more that the lead–lag relationship should be explored further.

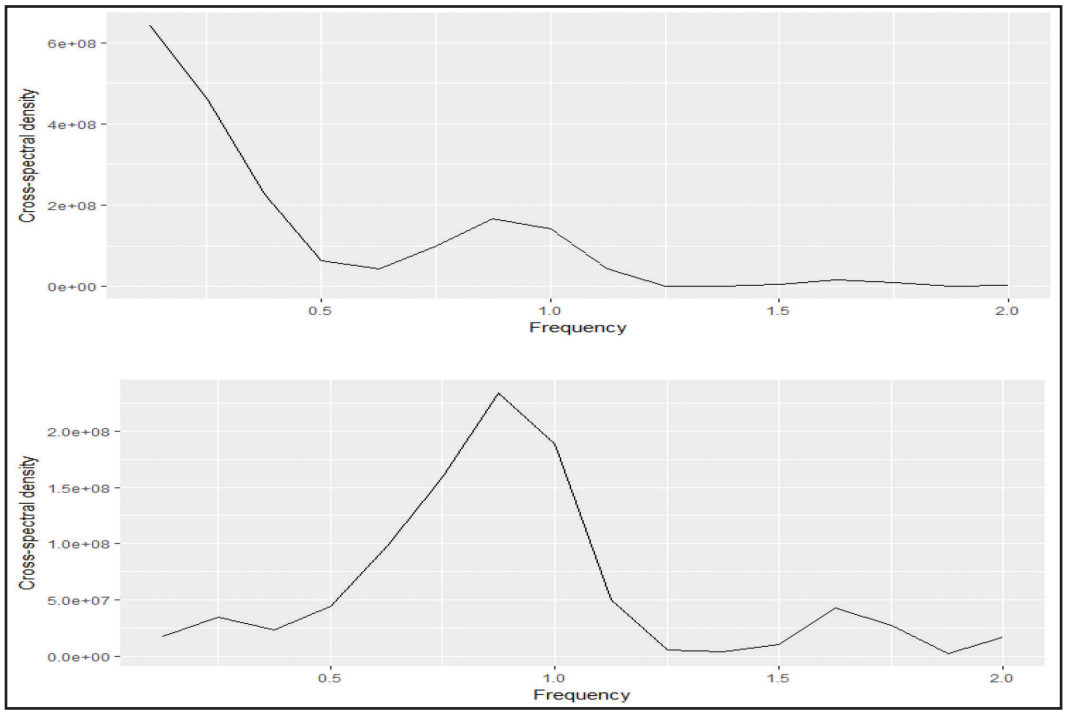
In addition, there is a certain level of synchronization of the series that is less evident in the differentiated series. A possible interpretation combining these two considerations is that, in the short run, web data might be faster in capturing seasonal changes, while in the longer run the two series follow similar trends and present co-movements over time spans longer than a year.

Spectral analysis is based on the decomposition of time series into their cycle components, i.e., a time series can be the sum of two or more cycles at different frequencies, such as a cycle in the long term and a cycle in the short term. Cross-spectral density is a measure of the synchronization of the series at different frequency values. When two time series have cycles at the same frequency, they present high cross-spectral density at that specific frequency and they are synchronized. Figure 7 shows cross-spectral density of the original series (top) and differentiated series (bottom). High cross-spectral density at low frequencies (graphically, the high level of the curve on the left side of a cross-spectral density chart), means that in the long term (when it takes a longer period to complete a cycle) the series are tied together; correspondingly, high cross-spectral density at high frequencies (a high level of the curve on the right side of the graph), implies that in the short run the series are tied together.

Figure 7 shows a certain amount of synchronization of the original series. The maximum coherency appears for the very long term. The short length of our data series and the fact that both series begin and end at similar levels justify the high cross-spectral density in the "long" run (left side of the graph). It is however very interesting to observe a local maximum around frequencies close to 1.

The same pattern (maximum at frequencies close to 1) is present and more evident in the case of the differentiated series and confirms the high synchronization of the series. High cross-spectral density at frequencies close to 1 (both, the local maximum of the original series and the maximum cross-spectral density for the differentiated series, occur at a frequency of 0.875) means that changes in the series occur fairly simultaneously. Given the quarterly nature

**Figure 7** Cross-spectral density, original series (top), and differentiated series (bottom)



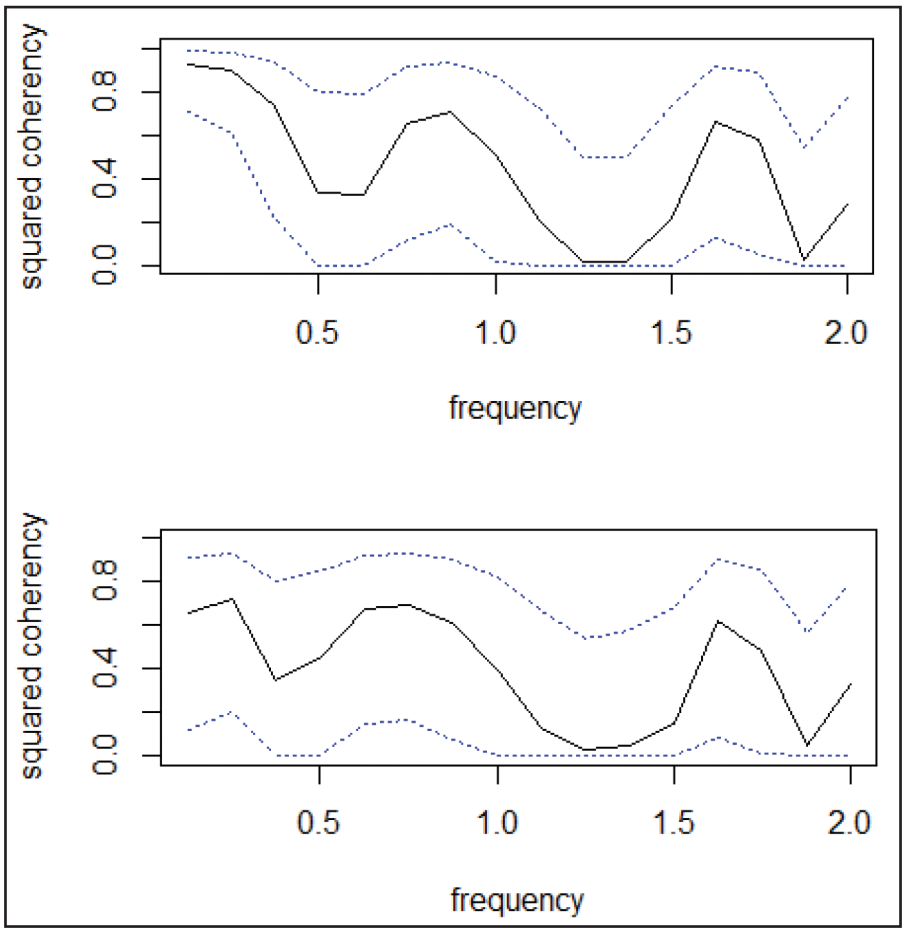
of our data, the high cross-spectral density at this frequency indicates that both series share cycles of 4.4 months in length.[10] In short, the cross-spectral density analysis suggests co-movements of the series. The time series of web and NSO move synchronized over time.

Squared spectra coherence is (in a frequency domain) analogous to what the correlation coefficient is in statistical inference: it measures the strength of the linear relationship between two series with the peculiarity that it does it for each frequency. High squared coherency means significant/considerable linear relationship at the corresponding frequency. Therefore, again, it can help in observing the relationship of two time series in the long and short term. It is calculated as the cross-spectrum divided by the square root of the product of univariate spectra. It is the percentage of the common variance due to common cycles.

Of all the squared coherency values, those corresponding to the frequencies where the series have higher values of cross-spectral density are the most meaningful, the relevant ones, because they measure the series correlation of shared cycles. In our case, the coherency values correspond to the frequencies close to 1.

Figure 8 represents the squared coherency across different frequencies of the original data (above) and of the differentiated series (below). According to the definition presented in the previous paragraph, the peaks corresponding to low frequencies (left side of the graph) imply

**Figure 8** Squared coherency, original series (top), and differentiated series (bottom)




10 The frequency value of 0.875 corresponds to cycles of approximately 1.14 periods. When the period is a quarter, this frequency corresponds to cycles of 3.4 months in length (3-month times 1.14).

long-run synchronization. The peak observed at higher frequencies (above 1.5, right side of the graph), present in both analysis but more evident in the differentiated series case, suggests that differentiated series present co-movements at frequencies which roughly correspond to 2-month cycles.[11]

In addition, and more interestingly for our analysis given the results of the cross-spectral density, is the central peak present in both panels of Fig. 8. This peak implies high synchronization at frequencies close to 1 for both pairs of series.

In other words, the latter finding, along with the results of the cross-spectral density, can be interpreted as evidence of highly correlated co-movements between the time series of web and NSO.

Autocorrelation, cross-correlation, cross-spectral density, and squared coherency measures produce important insights into the relationship between the web and the NSO time series. In particular, the results produced by the last two measures support the assumption of high synchronicity between the series. Although we cannot consider this synchronicity as proof, it supports the hypothesis that the two series measure the same underlying phenomenon: the real number of new vacancies appearing in the Dutch labor market each quarter. Both time series are almost the same but reflect slightly different realities. These small differences may be related to a compositional effect, i.e., a situation by which differences between the characteristics of two groups are attributable to the group composition (Duncan et al., 1992*)*. A time series comparison by sectors is beyond the scope of this paper. A preliminary sectoral comparison shows that in some sectors both time series are more similar than for others. In addition, these analyses suggest the potential presence of a lead–lag relationship between the series. Both results need to be investigated further.

## 5 Conclusions and limitations

A comprehensive contrast of web and NSO vacancy data revealed that they present similar time series properties. This result was obtained through visual comparison of traditional time series decomposition, a set of stationarity tests and cross-spectral analyses. The components of both were found to behave similarly over a 7-year period. TCs reflected a very similar impact of the economic crisis, and business cycles, seasonal effects (*S*), and the irregular terms (*I*) were of very similar magnitude, confirming the strong relationship maintained by the web and NSO vacancy time series. Both displayed very similar autocorrelation patterns in both raw and differentiated data. Cross-correlation analyses suggest a positive, significant and almost simultaneous co-movement between the series that is less clear in the differentiated series. The results produced by the cross-spectral density and squared coherency analysis support the assumption of high synchronicity between the series.

Both the traditional decomposition and the spectral analysis suggest the presence of a lead–lag relationship between the series; however, no conclusion could be drawn on this aspect.

It is also worth noting that the NSO 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.

11 The frequency value of 1.5 corresponds to cycles of approximately 0.667 periods. When the period is a quarter, this frequency corresponds to cycles of 2 months in length (3-month times 0.667).

One important limitation of this study is the availability of the vacancy data and algorithms. One could think that this type of data is in the open domain, but this is not always the case, because vacancy scraping is predominantly in the domain of commercial companies. We use data from Textkernel, a professional company with high-quality scraping mechanisms to collect vacancy data from the web. They continuously scrape, de-duplicate, and process all vacancy data in the Netherlands. Their service is proprietary, but since they were collaborating in the EU-funded Eduworks project, the authors were given access to this very rich data set (by signing a Non-Disclosure Agreement). This agreement expired after the conclusion of the Eduworks project, and therefore we could not enlarge the web data with more recent years, nor could we access their algorithms. However, there are other new projects that may provide vacancy data. For instance, the EU Agency Cedefop runs a project to set up vacancy scraping and provide an analytical service for the European Commission. The reality is that data access is mainly dependent on contacts and seniority of researchers. If available, data are very often offered subject to non-disclosure agreements as companies may danger their own business if their data and algorithms end up in hands of their competitors. Openness like Textkernel should be very much appreciated. Many companies may "not like the prospect of negotiating access with individuals in case by case basis, and decide not to make data available to everybody or to nobody" (H. Varian in Taylor et al., 2014 page 8). Access and transparency of algorithms is also an unresolved issue (European Commission, 2016; Connolly, 2016; Scott and Young, 2018; Barzic et al., 2018). The scientific community also needs to find the way to obtain agreements that make it possible further explorations of algorithms.

Another note on limitation is that researchers in this area should count with the need of significant (technical) processing power when dealing with vacancy data. The high-volume and real-time data feed options require a strong, stable, and secure information processing infrastructure.

In summary, our results gave several robust pieces of evidence that suggest that the web and NSO time series vacancy data were generated by the same underlying phenomenon: the real number of new vacancies appearing in the Dutch labor market each quarter. This supports the idea that web-sourced data are able to capture aggregate economic activity. Finally, both the NSO and the web data allow for breakdowns by industry. Our future research agenda includes exploring the possibility of lead–lag relationships, the possibility that the comparison presented above might apply to each sector of the economy, put the difference between the two time series into relation with other variables such as Internet penetration, and the development of new real-time labor market measures. In general, future research focusing on the use of web data will be determined by researchers' access to data and algorithms.

## 6 Availability of data and material

Data are available under request at pablo.depedraza@ec.europa.es. The empirical analysis had been performed within the R software environment.

- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

The Unit root tests were implemented through the tseries package using the *adf.test* function for the ADF test and the *pp.test* function for the PP test and through the *ur.za* function of the urca package for the ZA test.

- Adrian Trapletti and Kurt Hornik (2017). Tseries: Time Series Analysis and Computational Finance. R package version.
- Pfaff, B. (2008) Analysis of Integrated and Cointegrated Time Series with R. Second Edition. Springer, New York. ISBN 0-387-27960-1 0.10-37.

The decomposition analysis was performed through the STL (seasonal decomposition of time series by LOESS) function. Autocorrelation and cross-correlation values were obtained with the ACF (auto-covariance and cross-covariance and cross-correlation function estimation) function. Cross-spectral values and squared coherency values were obtained with the spectrum (estimated frequency spectrum) function.

**Competing interest**
The authors declare that they have no competing interests.

**Funding**
The authors acknowledge the financial contribution of the SERISS project (H2020 No: 654221).

**Author contributions**
PP has coordinated this work and has participated in every step from the development of the idea to final writing including data curation and analyses. SV has leaded the time series analyses and contributed to final writing. KT was the principle investigator of the EDUWORKS project and has been the supervisor of this work. GK was the EDUWORKS project proposer and coordinator; he was in charge of relations and agreements with Textkernel and has participated in data processing. All the authors read and approved the final manuscript.

**Acknowledgments**
The authors acknowledge (www.webdatanet.eu; Cost action IS1004) the comments and suggestions from participants at the Amsterdam institute for Advanced Labour Studies (AIAS) lunch seminar on April 23, 2015, and at the Conference on Intelligent Machines and the Future of Recruitment, organized by http://www.textkernel.com/nl/. We also thank Casper Kaandorp from the University of Amsterdam/AIAS, Bauke Visser and Jakub Zavrel from Textkernel, see http://www.textkernel.com/nl/, Dirk ter Steege and Linda Muller-Geuzinge from Statistics Netherlands (CBS), and Vladimer Kobayashi from the University of Amsterdam/Amsterdam Business School. The scientific output expressed in this paper does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication. The authors acknowledge the Marie Curie Initial Training Networks (FP7-PEOPLE-2013-ITN) project "EDUWORKS: Crossing borders in the comprehensive investigation of labor market matching processes: an EU-wide, transdisciplinary, multilevel, and science-practice-bridging training network."

**Biography**
Pablo de Pedraza is an economist at the European Commission Joint Research Center and was an EDUWORKS Marie Curie post doc researcher at the University of Amsterdam. He has published topics related to the use of web-based data to study life satisfaction, job satisfaction, and the labor market matching process.
Stefano Visintin is an applied economist with a passion for the analysis of (large amounts of) data and all the tools that make this type of analysis possible, from econometrics to the application of machine learning algorithms for prediction. Presently he is a lecturer at the Universidad Camilo José Cela in Spain where he also directs the Business Administration department.
Kea Tijdens is a sociologist and a senior researcher at the University of Amsterdam. Since 2000, she is also the scientific coordinator of the global WageIndicator data collections on work and wages. She has published on issues related to labor markets, vacancies, occupations, and wage setting.

Gábor Kismihók, PhD, is the head of the Learning and Skills Analytics research group at Leibniz Information Centre for Science and Technology in Hannover. His main focus of research is on matching processes between individuals, education and the labor market. He is also advising a number of European and international organizations on educational and research policies.

**References**


Antenucci, D.; M. Cafarella; M. C. Levenstein; C. Ré; M. D. Shapito (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.; E. Galan (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.; F. Pinto; P. de Pedraza (2015): Can Internet Searches Forecast Tourism Inflows? International Journal of Manpower 36(1), 103-116.

Askitas, N.; K. F. Zimmermann (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.

Barbera, P.; G. Rivero (2015): Understanding the Political Representativeness of Twitter Users. Social Sciences Computer Review, 33(6) http://journals.sagepub.com/doi/full/10.1177/0894439314558836.

Barzic, G.; M. Rose; M. Rosemain (2018): French Officials are Going to Work at Facebook for 6 Months. World Economic Forum. https://www.weforum.org/agenda/2018/11/france-to-embed-regulators-at-facebook-to-combat-hate-speech/

Blank, G. (2017): The Digital Divide Among Twitter Users and Its Implications for Social Research. Social Sciences Computer Review 35(6), 1-19. http://journals.sagepub.com/doi/full/10.1177/0894439316671698.

Broder, A. Z.; S. C. Glassman; M. S. Manasse; G. Zweig (1997): Syntactic Clustering of the Web. Computer Networks and ISDN Systems 29(8-13), 1157-1166.

Butler, D. (2013): When Google got Flu Wrong. Nature 494, 14th February 2013.

Cavaliere, G.; I. Georgiev (2007): A Note on Unit Root Testing in the Presence of Level Shifts. Statistica 66(1), 4-18.

Chala, S. A.; F. Ansari; M. Fathi (2016): A Framework for Enriching Job Vacancies and Job Descriptions Through Bidirectional Matching. In *WEBIST (2)* (pp. 219-226).

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

Cleveland, R. B.; W. S. Cleveland; J. E. McRae; I. Terpenning (1990): STL: A Seasonal-Trend Decomposition Procedure Based on Loess. Journal of Official Statistics, 6, 3-73.

Connolly, K. (2016): Angela Merkel: Internet Search Engines are "Distorting Perception". The Guardian 26 Oct 2016. https://www.theguardian.com/world/2016/oct/27/angela-merkel-internet-search-engines-are-distorting-our-perception.

Costas, L.; B. Eeckels (2011): A dynamic correlation approach of the Swiss tourism income. In Tourism Economics (pp. 127-147). Physica-Verlag HD.

De Leeuw, E. (2018): Mixed-Mode: Past, Present, and Future. Survey Research Methods 12(2), 75-89. doi:10.18148/srm/2018.v12i2.7402.

Duncan, C.; K. Jones; G. Moon (1992): Context, Composition, and Heterogeneity: Using Multilevel Models in Health Research. Social Sciences and Medicine 46, 97–117. https://www.sciencedirect.com/science/article/abs/pii/S0277953697001482.

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

European Commission (2016): Online Platforms and the Digital Single Market Opportunities and Challenges for Europe, COM(2016) 288 final). Commission's Communication on online platforms.

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

Fabo, B.; M. Beblavý; K. Lenaerts (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.

Fidrmuc, J.; I. Korhonen; I. Bátorová (2008): Dynamic Correlation Analysis of Business Cycles of the Emerging Asian Giants: The Awakening. Characteristics of Business Cycles: Have they Changed? 121.

Granger, C. W. J.; M. Hatanaka (2015): Spectral Analysis of Economic Time Series. (PSME-1). Princeton: Princeton University Press.

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

**Head, B. G.; E. Dean; T. Flanigan; J. Swicegood; M. D. Keatin** (2016): Advertising for Cognitive Interviews: A Comparison of Facebook, Craigslist, and Snowball Recruiting. Social Science Computer Review 34(3), 360-377.

**Hitzler, P.; K. Janowicz** (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.

**Iacobucci, A.** (2005): Spectral Analysis for Economic Time Series. New Tools of Economic Dynamics, 203-219.

**Jayaram, S.; I. Patnaik; A. Shah** (2009): Examining the Decoupling Hypothesis for India. Economic and Political Weekly 109-116.

**Jijkoun, V.** (2016): Online Job Postings have Many Duplicates. But how can you Detect them if they are not Exact Copies of Each Other? Retrieved March 21, 2019, https://www.textkernel.com/online-job-posting-many-duplicates-can-detect-not-exact-copies/.

**Kobayashi, V.; S. T. Mol; G. Kismihok; M. Hesterberg** (2016): Automatic Extraction of Nursing Tasks from Online Job Vacancies. In M. Fathi, M. Khobreh, & F. Ansari (Eds.), Professional Education and Training through Knowledge, Technology and Innovation (pp. 51–56). Retrieved from http://www.pro-nursing.eu/web/resources/downloads/book/Pro-Nursing_Book.pdf.

**Kureková, L. M.; M. Beblavý; A. Thum-Thysen,** (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.; B. Quenneville** (2001): Seasonal Adjustment with the X-11 Method. New York: Springer.

**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. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed 30 June 2016, and http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/.

**Lazer, D.; R. Kennedy; G. King; A. Vespignani** (2014): The Parable of Google Flu: Traps in Big Data Analysis. Science 343(6176), 1203-1205.

**Lenaerts, K.; M. Beblavý; B. Fabo** (2016): Prospects for Utilisation of Non-Vacancy Internet Data in Labour Market Analysis—An Overview. IZA Journal of Labor Economics 5(1), DOI 10.1186/s40172-016-0042-z.

**Leon, C.; B. Eeckels** (2011): A Dynamic Correlation Approach of the Swiss Tourism Income, in: Tourism Economics. Physica-Verlag HD, 127-147.

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

**Maravall, A.** (2005). "An application of the Tramo Seats automatic procedure; direct versus indirect adjustment," Working Papers 0524, Banco de España; Working Papers Homepage.

**Martin, B.** (2018): Persistent Bias on Wikipedia, Methods and Responses. Social Sciences Computer Review 36(3), 1-10. http://journals.sagepub.com/doi/full/10.1177/0894439317715434.

**Pedraza, P. de; K. Tijdens; R. Muñoz de Bustillo; S. Steinmetz** (2010): A Spanish Continuous Voluntary Web Survey: Sample Bias, Weights and Efficiency of Weights. Revista Española de Investigaciones Sociológicas Nº 131 (Julio-Septiembre 2010), 109-130. http://www.reis.cis.es/REIS/PDF/REIS_131_041277971869681.pdf.

**Pedraza, P. de; K. Tijdens; S. Visintin** (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.

**Pedraza, P. de; K. Tijdens; S. Visintin** (2018): The matching process before and after the crisis in the Netherlands. International Journal of Manpower, 39(8), 1010-1031. DOI 10.1108/IJM-10-2018-0329.

**Pfaff, B.** (2008): Analysis of Integrated and Cointegrated Time Series with R. Second Edition. Springer, New York. ISBN 0-387-27960-1 0.10-37.

**Phillips, P.; P. Perron** (1988): Testing for a unit root in time series regression. Biometrika 75.2 (1988): 335-346.

**Pissarides, C. A.** (2000): Equilibrium Unemployment Theory, 2nd edn Cambridge: MIT Press (first ed. 1990, Oxford: Blackwell).

**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.; C. A. Pissarides** (2001): Looking into the Black Box: A Survey of the Matching Function. Journal of Economic Literature XXXIX(June), 390-431.

**Rafali, P.** (2018): Nonprobability Sampling and Twitter. Strategies for Semibounded and Bounded Populations. Social Sciences Computer Review 36(2), 2018. http://journals.sagepub.com/doi/pdf/10.1177/0894439317709431.

**Sáez Martín, A.; A. Haro de Rosario; M. C. Caba Pérez** (2016): An International Analysis of the Quality of Open Government Data Portals. Social Sciences Computer Review 34(3), 2016.

**Scott, M.; Z. Young** (2018): France and Facebook Announce Partnership Against Online Hate Speech. Emmanuel Macron has Teamed up with Mark Zuckerberg to Review the Country's Regulatory Response to the Issue. Politico 11/13/2018. https://www.politico.eu/article/emmanuel-macron-mark-zuckberg-paris-hate-speech-igf/.

**Stern, M. J.; I. Bilgen; C. McClain; B. Hunsche** (2016): Effective Sampling From Social Media Sites and Search Engines for Web Surveys: Demographic and Data Quality Differences in Surveys of Google and Facebook Users. Social Sciences Computer Review 1-19. doi:10.1177/0894439316683344.

**R Core Team** (2016): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

**Revilla, M.; C. Ochoa; G. Loewe** (2017): Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior. Social Sciences Computer Review 35(4), 2017.

**Revilla, M.; A. Cornilleau; A. S. Cousteaux; S. Legleye; P. Pedraza** (2015): What is the Gain in a Probability-Based Online Panel of Providing Internet Access to Sampling Units Who Previously Had No Access? Social Sciences Computer Review 1-18 http://ssc.sagepub.com/content/early/2015/06/04/0894439315590206.full.pdf?ijkey=nNfsKd0vcQ5sRqq&keytype=finite.

**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.

**Said, E.; D. A. Dickey** (1984): "Testing for unit roots in autoregressive-moving average models of unknown order." Biometrika 71.3, 599-607.

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

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

**Taylor, L.; R. Schroeder; E. Meyer** (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.

**Trapletti, A.; K. Hornik** (2017): tseries: Time Series Analysis and Computational Finance. R package version.

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

**Zivot, E.; D. W. K. Andrews** (2002): Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis. Journal of Business & Economic Statistics 20(1), 25-44.