**Measuring the quality of commercial and big data sources for official statistics**

František Hajnovič, Office for National Statistics, *frantisek.hajnovic@ons.gov.uk*

**Abstract**

*There has been an increasing shift towards using data science techniques and accessing alternative big data sources across all sectors of society. The UK government has been exploring the use of big data within official statistics. These are neither administrative data, nor existing government data, and could include web scraped data, big data from commercial companies and social media data. The Office for National Statistics (ONS) is exploring the use of these alternative data sources as part of its "Better Statistics, Better Decisions" strategy. Using these data sources and integrating them into official statistics presents significant methodological challenges, including bias, variety of data (such as text and images) and that there is no control over data supply. As an official statistics producer, ONS is committed to ensure that any statistics derived from this data meet user needs and are of high quality standards. As an emerging field, there is little guidance on the measurement of quality within big data and data science applications for official statistics.*

*Nevertheless, there are existing dimensions of quality that offer a good framework for using data in an official capacity. Derived from the European Statistical System dimensions of quality, the UK Government Statistical Service’s (GSS) eight quality dimensions for all published statistics can be applied to big data. These include; relevance, accuracy, timeliness and coherence.*

*The presentation will explore the quality implications of big data and data science methods through example projects of the ONS Big Data team, including use of social media data and collaborations with other National Statistical Institutes. Through these, we will establish how quality could and has been measured using the eight dimensions, challenges faced and ideas for applying quality measures to future big data sources.*

**Keywords:** Data Science, Big Data, Official Statistics, Quality

**1. Introduction**

In recent years, the term “Big Data” has been more frequently used in the public sector, especially in the national statistics offices (NSI). The UK’s Office for National Statistics (ONS) has established the Big Data team in 2014 with the aim of exploring the potential use of Big Data and alternative data sources in the context of official statistics.

Other NSIs and public bodies have conducted similar investigations. Australian Bureau of Statistics has a “Big Data Flagship” with a number of work packages including population mobility using mobile phone data or deriving agricultural statistics from satellite sensor data (Tam & Clarke, 2015). Statistics Netherlands have analysed traffic intensity using data from Dutch traffic loop detectors and explored the potential of social media data. They found that the derived sentiment from Twitter and Facebook messages strongly correlates with the consumer confidence index (Buelens et al., 2014, Daas et al., 2015). Finally, ESSNet Big Data is a project jointly undertaken by 22 partners (mainly NSIs) within the European statistical system, with a primary objective being integration of Big Data in the regular production of official statistics. The data sources considered as part of the 8 work packages include online job vacancy data or mobile phone data[[1]](#footnote-1). For more general discussion on Big Data in official statistics and the role the NSIs might adopt, see (Struijs et all, 2014).

In this paper, we first present two use cases of using Big Data for official statistics from ONS : *Web-scraping for job vacancies* pilot and *Using geo-located Twitter data to infer residence and mobility*. Using the UNECE Big Data quality framework, we highlight the associated quality issues, and discuss possible ways of addressing them, as well as measuring the quality of the outputs before concluding.

**2. Two use cases**

*2.1. Web-scraping for job vacancies pilot*

This two-year project was undertaken as part of ESSNet Big Data Work Package 1[[2]](#footnote-2). The overall goal was to investigate “which approaches (techniques, methodology etc.) are most suitable to produce statistical estimates in the domain of job vacancies”.

Our efforts in UK mainly focused on investigating the idea of nowcasting job vacancy (JV) estimates using online data[[3]](#footnote-3). We used two types of data sources: daily JV counts by company from 7 different job portals[[4]](#footnote-4) (web-scraped using automatic spiders) and time series of job vacancy data from several 3rd party providers, including Burning Glass (BG) and Adzuna[[5]](#footnote-5). Figure 1 shows the large differences in the total number of JV counts by the source, including the job vacancy survey (JVS) estimates.

**Figure 1 Total JV counts in time by source. The thick dotted line corresponds to the ONS job vacancy survey estimates.**

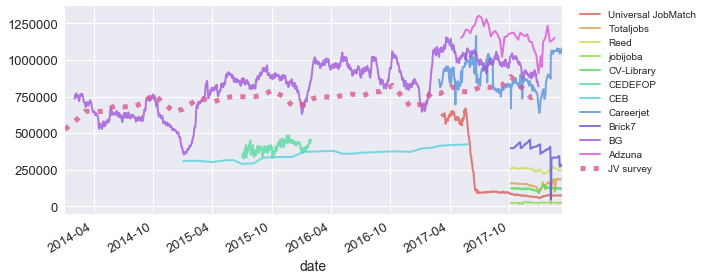
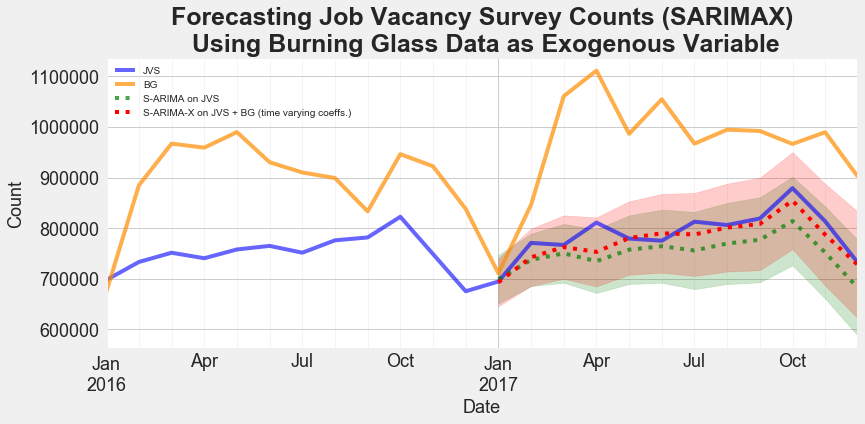


Figure 2 Time series model based on JVS only (green dotted line) and with BG data included as exogenous regressor (red dotted line), the latter following the real trend (blue line) more tightly



Burning Glass, as the source with the longest time-series, showed a strong and very significant (p-value < 0.001) correlation of ~ 0.77 with the JVS. This was exploited to build a S-ARIMA-X (seasonal ARIMA with exogenous regressor) time-series model on the JVS data with BG time-series included as exogenous regressor. Evaluated on the data from 2017, the model showed 50% improvement in the root mean squared error against a S-ARIMA model based solely on the JVS data (Figure 2).

We also investigated nowcasting at the individual company level. Data were linked on the company name using a heuristical string-matching algorithm and nowcasting models were built on the matched data, using online JV counts, as well as historic JVS counts as predictors. Here, however, only mediocre improvements were made over a “persistence” baseline model (one that nowcasts the last-seen JVS estimate). For more details, see the UK country annex of the SGA-2 final technical report[[6]](#footnote-6).

*2.1. Using geo-located Twitter data to infer residence and mobility*

One of the initial pilots in ONS was exploring the potential of using geolocated Twitter data to provide insights into population and mobility. The analysis used a dataset of all geolocated tweets sent within Great Britain during a period of 7 months in 2014. These were subsequently clustered by location, with the formed clusters further enriched with additional information providing address type (residential, commercial, other). For each user, a *dominant residential cluster* was determined based on the location with the highest number of tweets, de-facto a proxy for the user’s residence. Tracking changes of the residence over time can then be used to infer mobility of the population.

However, a number of quality issues were identified that prevent using the data at its face value to directly supplement the official statistics. These are largely related to the high selectivity of the Twitter user base (biased towards younger population with higher socio-economic status and living in urban areas), as well as bias due to focusing only on the users that posted ***geo-located*** tweets (and that did so regularly over the measured period). More information on this pilot can be found in (Swier et al, 2015).

**3. Applying UNECE big data dimensions**

The Suggested Framework for the Quality of Big Data, published by the UNECE Big Data Quality Task Team in 2014[[7]](#footnote-7) provides a useful guide when considering the quality of a given Big Data source. To save space, here we only provide the quality issues corresponding to the main 11 quality dimensions of the framework (Table 1). This list is not exhaustive, but rather indicative of the main challenges needed to be considered before using the data for producing official statistics [[8]](#footnote-8).

**Table 1 Main challenges classified by the UNECE quality dimensions for the two use cases**

|  |  |  |
| --- | --- | --- |
|  | **Web-scraping for JVs** | **Using geolocated Tweets** |
| **Institutional / Business environment** | **Unstable data supply** – possible ban of web-scraping activity/change in business model/termination of 3rd party data supply at any time | |
| **Privacy and Security** |  | **Passive agreement to T&C** – users often do not read thoroughly the T&Cs, raising ethical questions of using the data |
| **Complexity** | **Complex enterprise structure** – this creates a risk of miss-alignment when linking the company-level data to JVS units |  |
| **Completeness** | **Necessity of flow to stock conversion** – lack of information on job ad expiry date necessitated assumptions to be made when converting some data from flow of new vacancies to stock measures | **Lack of demographics** – majority of the Twitter profiles lack information on demographics, such as age, gender or occupation |
| **Usability** | **Classification non-uniformity** – 3rd party sources often use non-standard industry or occupation classifications  **Cost of building the infrastructure** – there is a substantial cost in building a web-scraping framework, cleaning and processing JV data. | **Cost of obtaining the data** – resources are required to gather and store the data |
| **Time factors** | **Presence of time-lag** – the time-series for online job vacancies often exhibit a variable time lag when compared to the JVS time-series |  |
| **Accuracy** | **Online JV selectivity** – not all of JVs are advertised online, with the sample favoured towards certain industries. There is further bias for individual sources, as not all online JVs are advertised on all job portals. | **Twitter selectivity** – it is well known that the sample of Twitter users is biased towards young adults with high socio-economic status. Also, only a fraction (1.6%) of Twitter users use geolocated tweets, with the penetration rate generally higher in urban areas |
| **Coherence** | **Challenging linkability** – due to absence of other information, the unit-level data were matched only based on the company name string, leading to both type I and type II errors. | **Consistency over time** – (Swier et al, 2015) showed there was a distinct dip in number of geo-located tweets from iPhone users after the release of iOS8, which included changes to default privacy and location settings |
| **Validity** | **Job ad vs. job vacancy** – there are definitional differences between the job ad (measured concept) and the job vacancy (target concept) | **Dominant residential cluster proxy** – the dominant residential cluster is only a (possibly inaccurate) proxy for the true residence of the given user |
| **Accessibility and Clarity** | **Black box problem** – often little information is provided about the processes used to produce the 3rd party data |  |

**4. Addressing quality issues and measuring quality**

Despite the significant amount of work and literature in this area, relatively little has been done to measure or correct for the quality issues of using Big Data in official statistics. In (Struijs et al, 2014), ways of dealing with several of these challenges are discussed and in (Buelens et al, 2014), the discussion is further developed for selectivity. Here we look at addressing selected challenges listed in the table above.

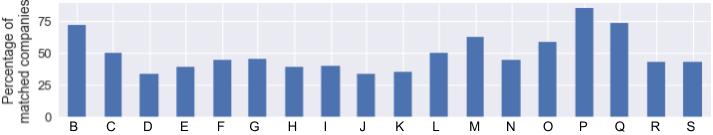
*4.1. Web-scraping for job vacancies pilot*

**Unstable data supply problem**. This problem can be mitigated by using multiple datasets and either the aggregated information, or the best available source for the purpose. Although having access to multiple datasets can be considered a luxury, in this pilot we managed to get access to 11 different data sources. For nowcasting of number of job vacancies, the online source with the highest correlation with the JVS on the period prior to the date of the nowcast was picked to serve as predictor of the nowcasted value. This means that in the case of disruption of data supply, a second-best source can be substituted in place. The amount of correlation can also provide a (highly experimental) indicator of the quality of the nowcast.

**Online JV selectivity.** Possibly the most important issue is that of selectivity of the online sources. This can be better assessed by linking the company-level data to the business register, although successfully matched pairs are more likely to be produced for certain industries, thus additional bias is introduced[[9]](#footnote-9) (Figure 3). However, once the data are linked, an assessment can be carried out on the final sample to compare its distribution against the original JVS distribution. Sample design techniques (e.g. assigning high design weights to underrepresented strata) can then be applied to correct for the bias. Due to lack of time and a relatively small sample size, attempting to do this remains a point of future research.

To measure the quality of the actual company-level nowcasts, we used standard machine learning evaluation techniques such as cross-validation, evaluating on a test dataset and comparing the results to the baseline model. This, combined with the mediocre results and the black-box nature of the used neural network models is unlikely to be sufficient for the use in official statistics. However, it is likely that having only such “experimental” metrics will become ever more a common case, as the use of alternative datasets and machine learning become more frequent in official statistics.

**Figure 3 Percentage of matched company names, by SIC industry class, for the largest enterprises in UK (with 2500 or more employees)**



*4.2. Using geo-located Twitter data to infer residence and mobility*

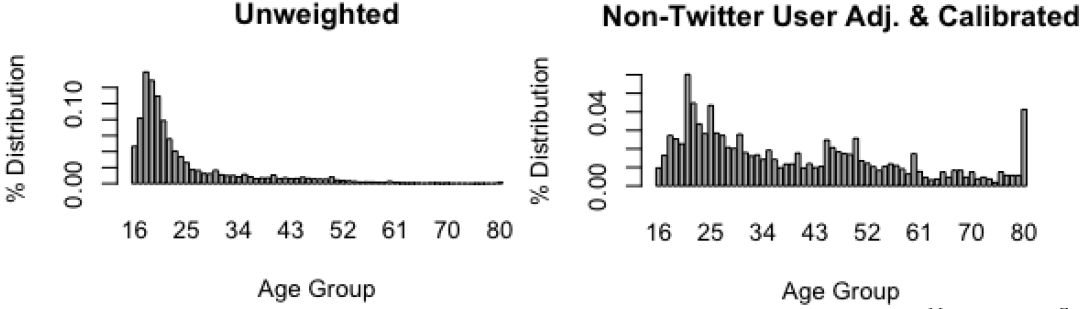
The work of (Swier et al, 2015) was further followed with investigation into some of the issues accompanying the gathered Twitter data. Although this follow-up work has never been completely finished and published, it contains some interesting ideas presented below, which address some of the main quality issues.

**Lack of demographics**. The lack of demographics metadata in the user profiles represents a challenge in understanding the sample of Twitter users. This can be mitigated by inferring some of the demographics from the profile descriptions, or using other metadata as proxies. Information on gender was derived for 88.8% users by matching a list of names from the birth registration data to the name field of the Twitter profile. Age was extracted for 6.06% of users by applying a string matching algorithm considering numbers present in the username string, demonstrating a substantial bias of the Twitter users towards young population. A similar experimental classification into students and retired people showed a similar strong bias: 29.8% of the users were identified as students and only 0.29% as retired people[[10]](#footnote-10).

A more robust and methodological way of understanding the sample of Twitter users was through including questions on the use of Twitter in the ONS Opinion and Lifestyle (OPN) survey[[11]](#footnote-11). The benefits are clear: by using a relatively small-scale survey, one gets understanding of a large, timely and rich Big Data source.[[12]](#footnote-12)

**Twitter demographics bias**. After understanding the amount of bias in the Twitter data, the next challenge is to correct for it. Here, an idea of creating design weights based on the results from the OPN survey was explored. First, a logistic regression model was built, linking the survey demographic variables to the probability of the user regularly tweeting. Age, ethnicity, occupation and whether the person is a student or not were detected as significant in terms of explaining the variance of in the model. A design weight (as a reciprocal of the mentioned probability) can then be assigned for each user with sufficient information on their demographics. Finally, the weights are further calibrated to satisfy known population totals. Figure 4 shows the age distribution before and after applying these bias correction steps.

**Figure 4 Age distribution before and after applying the bias correction steps described above**



The combined lack of relevant demographics metadata, error-prone inferring of this information from the profiles and a relatively small sample size of the OPN survey means the above-mentioned way of bias correction may be too ambitious to work in practice. More research would be desirable to build confidence in this approach.

**5. Conclusion**

This paper has explored the quality issues associated with big data sources for two use case projects we have undertaken in the UK’s Office for National Statistics. The UNECE Framework for the Quality of Big Data provided a useful tool for assessment of these quality gaps. Because the two projects and the involved data sources were of different nature, the associated quality issues and the approaches to their corrections also varied considerably. It is therefore our belief that while generic methods for measuring and ensuring quality of the statistics based on big data sources should be sought after, one mainly needs to focus on the specifics of the individual projects.

One of the considerations is whether there are multiple datasets available. This could often mitigate problems with disruption of data supply and improve quality by using aggregated, or best available source for given purpose at any given time. A second consideration is whether the data can be linked at a unit level to a known population sampling frame (e.g. companies from the online job vacancy data could be linked to the business register), as this can greatly help understand the selectivity bias. Ethical issues of doing so should be considered. Where the linking is not possible, additional metadata for individual units can be often derived from present metadata (e.g. inferring age, language or gender from Twitter user profiles). Information about the population of the Big Data source can be obtained by including questions in an omnibus survey, focused on the nature of the units within the sample (e.g. what kind of people tweet regularly). This in turn can serve as a basis for creating design weights and correcting for the bias via standard sample design techniques.

Finally, we note that with the advances in machine learning, the situation where standard machine learning metrics obtained by evaluating the model on a testing dataset provide the main indication of the accuracy may become ever more frequent. This, combined with the black-box nature of many machine learning models, e.g. neural networks, may be difficult to gain recognition in decision and policy making. Yet the results obtained this way should be carefully considered, improved upon and their contribution well documented to promote future advances in the method-logy.

**6. References**

Buelens, B., Daas, P.J.H., Burger, J., Puts, M. and Brakel, J. (2014). Selectivity of Big data. Statistics Netherlands Discussion Paper. 201411. Available at: <http://pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf> (Accessed 26.4.2018)

Daas, P., Puts, M., Buelens, B. and van den Hurk P.A.M. (2015). Big Data as a Source for Official Statistics. Journal of Official Statistics, 31(2), pp. 249-262. Available at: <https://www.degruyter.com/downloadpdf/j/jos.2015.31.issue-2/jos-2015-0016/jos-2015-0016.pdf> (Accessed at 26.4.2018)

Struijs, P. and Daas, P.J.H. (2014). Quality approaches to big data in official statistics. In European conference on Quality in Official Statistics. Available at: <http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf> (Accessed 26.4.2018)

Swier N. (2018). A quality framework for on-line job vacancy data. CEDEFOP/ESSNet validation meeting in Milan, March 2018. Available at: <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/7d/WP1_Quality_Framework_v1.1.pdf> (Accessed 26.4.2018)

Swier, N., Komarniczky B. and Clapperton B. (2015). Using geo-located Twitter traces to infer residence and mobility. GSS methodology series, no. 41. Available at: <https://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--41--using-geolocated-twitter-traces-to-infer-residence-and-mobility.pdf> (Accessed 26.4.2018)

Struijs P., Braaksma B., Daas P.J.H (2014). Official statistics and Big Data.Big Data & Society, volume 1, issue 1. Available at <http://journals.sagepub.com/doi/full/10.1177/2053951714538417> (Accessed 26.4.2018)

Tam, S.M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. International Statistical Review, *83*(3), pp.436-448. Available at: <https://onlinelibrary.wiley.com/doi/epdf/10.1111/insr.12105> (Accessed 26.4.2018)

1. For more information on the project and list of collaborating NSIs, see <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data> [↑](#footnote-ref-1)
2. For more information on this work package, involved partners and details of the work carried out in the UK, refer to the final technical report (SGA-2) <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP1_Working_area> [↑](#footnote-ref-2)
3. More specifically, we were interested if up to date information from online JV data can be used to predict (nowcast) the true current job vacancy estimates, as given by job vacancy survey [↑](#footnote-ref-3)
4. These were: Careerjet, CV-library, Universal JobMatch, Reed, Totaljobs, Jobijoba, Brick7 [↑](#footnote-ref-4)
5. Access granted for the purpose of this research [↑](#footnote-ref-5)
6. <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP1_Working_area> [↑](#footnote-ref-6)
7. Full version of the framework document can be found at <https://statswiki.unece.org/display/bigdata/Big+Data+in+Official+Statistics> [↑](#footnote-ref-7)
8. A more thorough application of the UNECE framework can be found in (Swier, 2018), where it was applied on a pan-european job vacancy web-scraping effort carried out by CEDEFOP. [↑](#footnote-ref-8)
9. SIC sectors P (education) and Q (human health) contained highest percentage of matched entries. The least proportion of matches was found in sectors J (information & communication technologies) and K (finance) where more complex enterprise structures (i.e. more difficult matching) are expected. [↑](#footnote-ref-9)
10. The census lists 4.5% of population (aged 16+) as students and 22.8% as retired [↑](#footnote-ref-10)
11. For information on methodology of the survey, see <https://www.ons.gov.uk/aboutus/whatwedo/paidservices/opinions> [↑](#footnote-ref-11)
12. The results showed that only 10% of the respondents tweet regularly and that 65% of the Twitter users do not include information on their age in their profile. [↑](#footnote-ref-12)