Measuring labour demand volumes across the UK using Textkernel data user guide

This user guide provides information on the data and methods used to compile ONS measures of labour demand by employers, using online job adverts provided by Textkernel.

Contact:
Skills and Human Development Team
Economic.Wellbeing@ons.gov.uk
+44 1633 456265

Release date:
5 November 2024

Next release:
To be announced

Table of contents

1 . Overview

This user guide provides information on the data and methods used to compile Office for National Statistics (ONS) measures of labour demand by employers, using online job adverts provided by Textkernel. There are several statistics published using this data, but unless otherwise specified, these will be referred to throughout as “labour demand” signifying this is demand for workers that employers are trying to fill externally.

Online job adverts datasets can provide a wealth of information relating to the labour market. Due to large volumes of online job adverts posted in the UK, it is possible to derive statistics at a granular level, for example, Local Authority District and occupations simultaneously.

Online job adverts data can also be used in near real-time and provide very high coverage of employers’ advertising patterns.

There is a range of regular statistics we publish to meet different user requirements. Users interested in current demand should look at the tables using the snapshot statistics, while those interested in emerging demand should look at the tables using the new adverts statistics, as they provide a better measure of changes in short-term demand. The different tables are outlined as follows:

- users interested in the overall demand in a particular geographic area across the UK should look at tables 1 to 6, and 19 to 24 of our Labour demand by Standard Occupation Classification (SOC) associated data tables

- users interested in demand for different types of jobs (Occupations), as well as comparing how local areas differ, should start with tables 8 to 12 and 26 to 30 after exploring the earlier tables mentioned; these show demand at the highest occupation level, at two-digit Standard Occupation Classification (SOC 2020) occupations, such as for Health Professionals

- users interested in more granular demand by occupation and local area can then explore tables 13 to 15 and 31 to 33, which show further occupational detail, such as for Nursing professionals

- users purely interested in the demand of a certain occupation can go to the most granular occupational detail in tables 7, 16 to 18, 25, and 34 to 36 where the most detailed occupations are shown, such as for Registered nurse practitioners

- users interested in the ability of the local workforce to meet current demand, or if there may be specific local occupational shortages, should use the skills shortages associated data tables

More information on how shortages statistics are derived can be found in the measuring skills and qualifications suitability user guide.

2 . Methods

To measure activity in online job advertising we produce two metrics – new adverts and snapshots of all live adverts. The general process we follow is:

- bring data in from Textkernel, an external supplier, which collects adverts by searching many websites daily, using a technique known as web scraping

- remove adverts that are not in our target population, such as adverts for jobs not based in the UK

- assign each advert a local authority, based on the 2023 boundaries, which can be found on the Open Geography Portal

- remove duplicate postings to get a truer picture of demand as employers often advertise jobs on multiple sites

- estimate the occupation (Standard Occupation Classification 2020, SOC2020) for each advert using a machine learning model that matches each advert's job title to the closest corresponding job title, and its associated occupation code

- we then aggregate across the two metrics to different periodicities, grouping by multiple geographies and SOC 2020

Each of these steps is described in further detail in this section.

Data sources and preparation

For the accompanying statistics, we are using Textkernel as the source of our online job adverts data. Textkernel data is collected using comprehensive web-scraping software which downloads job advert information from approximately 90,000 job boards and recruitment pages. The scraped data includes job titles, descriptions, posting dates and expiration dates. Textkernel remove some adverts when they are low quality such as if they are missing information or have incomplete sentences, and then apply in-house methods to derive additional data including location, salary, skills requirements, industry and others.

It is worth noting that Textkernel is a source of online job adverts rather than vacancies. There are several conceptual differences between vacancies and job adverts. A vacancy is a position for which an employer is actively seeking recruits from outside of their business or organization. This search can be through any means, for example, not just through online job sites but also by word of mouth, posting in a shop window or newspaper etc. On the other hand, an online job advert would capture a vacancy posted online but may also be a speculative opportunity to identify local talent, may contain more than one position or even not be an advert for a job but for training to provide an opportunity for a job later.

Our Textkernel-based statistics are official statistics in development. Read more in the guide to official statistics in development. The Office for National Statistics' accredited official source of vacancies is the Vacancy Survey (see the latest Vacancies and jobs in the UK bulletin) We have also used Adzuna data – which is a different data source of online job adverts – for several analyses, which are also official statistics in development, and our use of it is described in Using Adzuna data to derive an indicator of weekly vacancies: Experimental Statistics.

The data goes through a standard set of data cleaning prior to other processing, which is:

- removing all job adverts identified as non-UK based posts (7.8% of total adverts)

- removing all job adverts identified as not being written in English (0.2% of total adverts)

Geographical assignment

Each advert is assigned a local authority based on Textkernel's methodology. The location of the posting is identified from the full text and then matched to its central coordinate. Then the coordinate is mapped to the Local Authority District with 2023 updated boundaries.

For those adverts not able to be assigned one, such as due to the advert not specifying a location, those adverts have an "Unknown" geographic assignment which we publish. These account for 4.7% of remaining adverts after data cleaning. Adverts in London typically just specify "London" as the location, so labour demand is not shown down to Local Authority level there.

More aggregate levels of geography, such as International Territorial Level 2, Mayoral Combined Authority, Local Skills Improvement Plan area, Local Enterprise Partnership area, and Country, are all mapped from the Local Authority, rather than being separately identified.

Deduplication

There are multiple sources of advertisements being brought together, so the same job may be posted on different sites and/or across different time periods. We refer to these as duplicate adverts. Removing duplicate adverts is therefore necessary to get to a more accurate measure of true demand. We call this process deduplication.

Textkernel provides an identifier to mark where they have identified a duplicate advert. They apply natural language processing algorithms to identify text close enough to be considered as duplicate adverts, posted in nearby locations at similar times. We use this identifier and filter out all but one of the duplicate adverts to report deduplicated counts, selecting the job advert with the earliest posting date.

Standard Occupation Classification (SOC 2020) assignment

Each advert is assigned a four-digit standard occupation classification code (SOC2020) using a bespoke ONS method. If an advert cannot be assigned an occupation, then it is assigned as "Unknown".

In summary, to assign an occupation to each advert, we standardise the job title of the advert by cleaning the text field and match it against the closest corresponding job title from an index of many standardised and regularly updated job titles, each of which has an assigned occupation. This is the SOC 2020 Volume 2 coding index.

To successfully match job titles, we convert them into numeric representations using a method called term frequency – inverse document frequency (TF-IDF) and apply a similarity metric to find the closest corresponding record from the index. Further detail on the strengths and limitations of this method can be found in the strengths and limitations section.

Aggregation metrics

We show two labour demand metrics. These are the number of adverts being posted over a period of time (referred to as "new adverts"), and the number of live adverts available to apply for at a given point in time (referred to as "snapshot").

New adverts represent the inflow of adverts that have gone online during a time period. This metric is calculated by counting the total number of adverts that appear for the first time across the reported period, whether a month, quarter or other frequency published.

This metric is recommended for picking up short-term changes in the labour market, for example, if lots of jobs are being created in a month in a specific city.

Snapshots represent the stock, as an average point-in-time number of live adverts, averaged across days in a month. This metric is calculated by counting the number of adverts that were live on each Sunday each week. The counts are then averaged across a calendar month.

This metric is recommended for understanding total demand. It can be more meaningfully compared with other data such as employment data and is why it is our headline metric. Note that quarterly and annual snapshots are derived as the maximum value of the monthly figures.

Links to other labour market statistics

In the accompanying article released on 5 November 2024, “Which skills are employers seeking in your area?” the labour demand statistics were combined with other labour market data to provide some further insights and context.

First, we divided monthly snapshot volumes in an area by the size of its working age population in that area to give an indication of relatively higher or lower demand. This removes the effect that more populated areas will generally have a higher level of business activity and hence higher number of adverts than less populated areas.

In addition, snapshot volumes were used alongside data developed and published on Understanding skills and qualification suitability to develop experimental occupational shortages statistics. These are further explained in the accompanying user guide Measuring skills and qualification suitability.

We also focused on specific healthcare and social care occupations. This list is outlined in the associated data tables.

3 . Labour demand and shortages data

Labour demand volumes by Standard Occupation Classification (SOC 2020), UK
Dataset | Released 5 November 2024
These tables contain the number of online job adverts, split by different geographic areas and occupation (SOC 2020).

Quality metrics for Labour demand volumes by Standard Occupation Classification (SOC 2020)
Dataset | Released 5 November 2024
Reference tables containing quality metrics for online job adverts metrics from Textkernel.

Occupational Shortages by International Territorial Level 2 (ITL2) across the UK, 2023
Dataset | Released 5 November 2024
Estimates comparing demand for workers (online adverts) to supply of the workforce (skills of the working-age population).

Skills supply estimates: 2012 to 2023
Dataset | Released 9 August 2024
These reference tables contain skills supply estimates for the UK between 2012 and 2023. These are official statistics in development.

4 . Glossary

Demand

The demand for workers in the labour market from employers, either as a whole or for specific occupations or for those with specific skills.

Snapshot

A snapshot is a method of representing demand, showing the average point in time number of live adverts, averaged across a period (in our published data this is a month).

New adverts

New adverts are a method of representing demand, showing the total number of adverts that have gone online in the month.

Deduplication

The process whereby duplicated adverts (the same advert posted across multiple sites and time frames) are reduced down to one advert. See deduplication section for more details.

Vacancies

Vacancies are defined as positions for which employers are actively seeking recruits from outside their business or organisation. This is not equivalent to online job adverts, as the former are actively being recruited for and each one represents one post, while sometimes adverts can be multiple posts.

Shortages

The amount of demand that cannot be filled by the current local workforce. This can be for workers as a whole, for specific occupations or skills. We derive shortages through a few metrics, described in the skills suitability user guide.

5 . Strengths and limitations

In this section we outline information on how the previously described methods may influence interpretation and use of the data. We cover how the data is revised, compare coverage to official sources and how deduplication affects this, discuss data consistency over time and time-specific issues, elaborate on the method of assigning an occupation and geography, and highlight future improvements to improve our overall use of this data.

Revisions

The dataset is revised each time we receive weekly updates on new adverts and when open adverts expire. This means the series are revised, particularly affecting the last few months of the snapshot metric. This is because Textkernel closes adverts that have been live for more than a year, or if the majority of the duplicate adverts have been closed and the remaining ones are open for more than six months. For this reason, we always recommend using the latest published series.

In addition, revisions back to 2020 come from an imputation process we have applied. Adverts with original expiration dates between the periods: 1 April 2020 to 7 December 2021 had to have their expiration dates imputed because Textkernel informed us that there was a temporary issue with the web scraping algorithms they used. The software, at that time, reported the expiration of many adverts that were, in fact, still live. Hence, we imputed new expiration dates throughout the erroneous period. This imputation mostly affected the snapshot metric.

The revisions occur when we re-impute, as the imputed expiry dates sample durations from the unaffected time period to calculate new expiry dates. Each time sampling occurred, the levels of demand for occupations during the imputed window and beyond may be affected. The sampling was applied for each Textkernel profession to match to each advert as different occupations are filled more slowly or quickly.

For example, if the sampled adverts for a particular occupation had longer durations than the previous sampling for a publication, this would increase the snapshot metric volumes of that occupation, and that increased level would persist beyond December 2021.

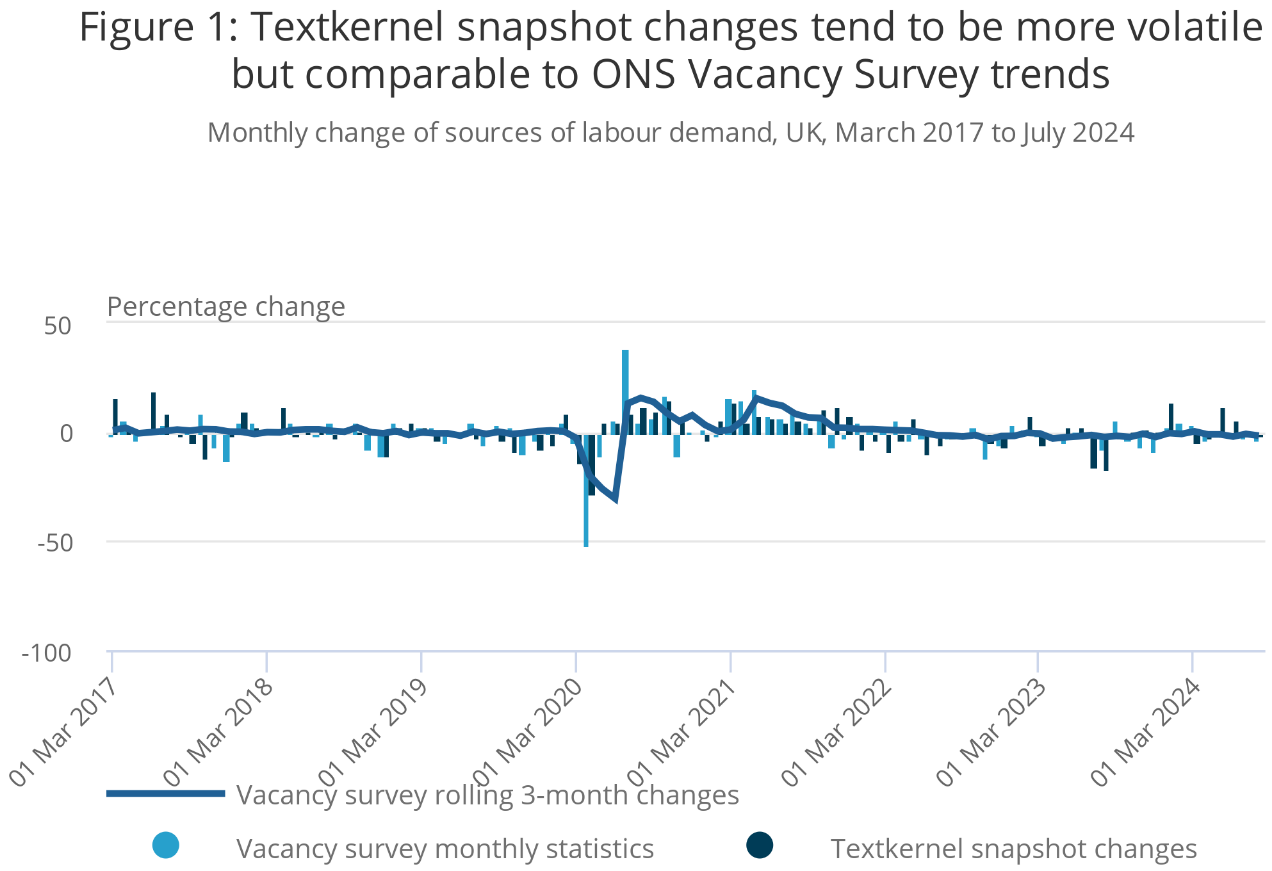
In future, we will only revise the imputation when improvements can be made and communicate this, to allow users to track a stable series of demand.

Coverage

The broad trend of demand from the snapshot metric is comparable to the ONS' accredited official source of vacancies through the vacancy survey, as seen in Figure 1.

Figure 1: Textkernel snapshot changes tend to be more volatile but comparable to ONS Vacancy Survey trends

Monthly change of sources of labour demand, UK, March 2017 to July 2024



Source: ONS Vacancy Survey, ONS analysis of Textkernel data

Generally, long-term trends tend to be consistent, but both the monthly vacancy survey statistics, which are not the official accredited source, and the developmental Textkernel statistics, tend to be more volatile than the official three-month average vacancy survey estimates. Between the end of 2021 and 2023, Textkernel snapshots tended to decline more sharply. However, despite these being deduplicated, the levels of demand are larger than official estimates, with 1.43 million snapshot adverts in August 2024, 66% higher than the 860,000 monthly volumes of vacancies for the same period. Some of these reasons are discussed further in the following subsection.

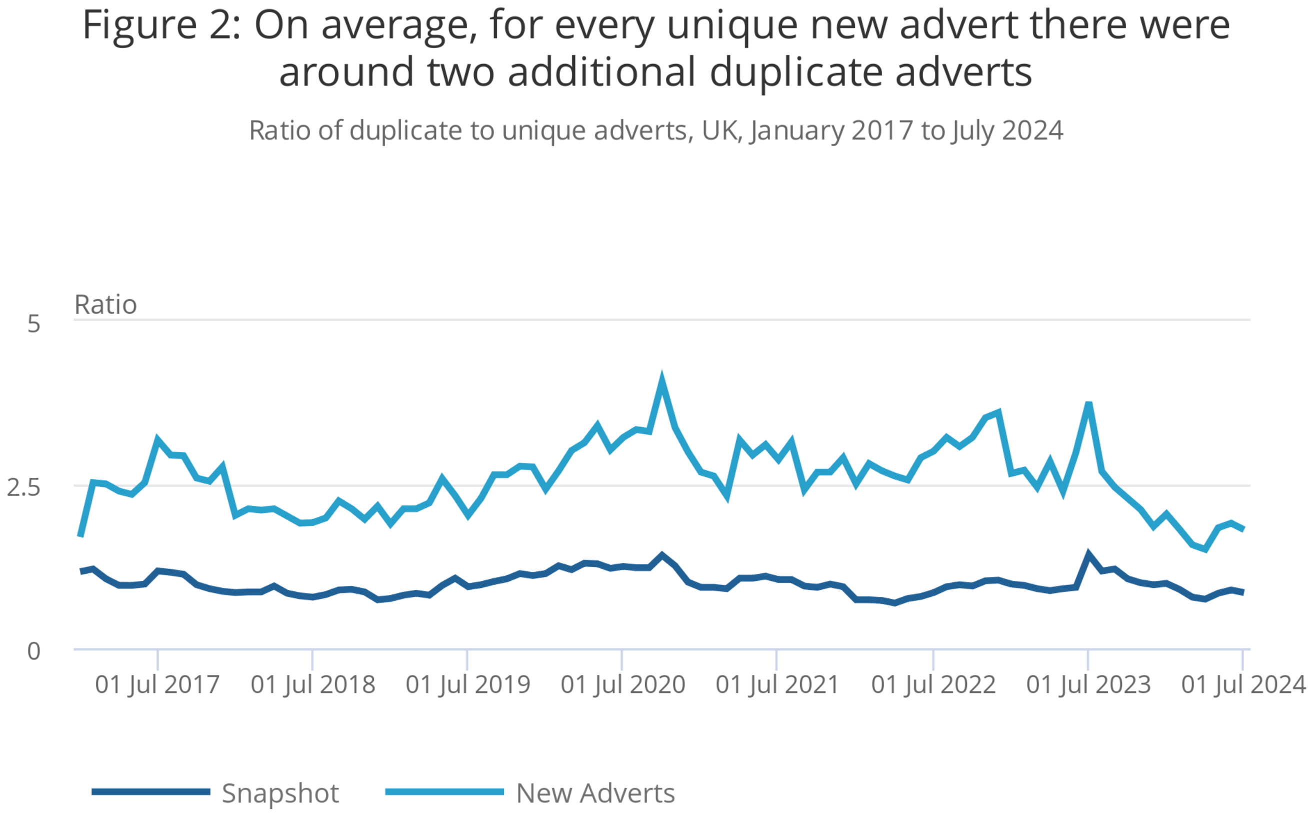
Deduplication

Textkernel captures adverts from multiple sources, so the issue of removing duplicate adverts to track genuine economic trends is critical, and we have applied a method based on Textkernel's to start to remove these. Broadly speaking, Textkernel's method uses shingling, min-wise permutation hashing and inverted indexing, as well as classification methods to find job adverts textually similar to other adverts. For more information, see this Textkernel blog "Online job postings have many duplicates. But how can you detect them if they are not exact copies of each other?".

Additionally, we record adverts marked as duplicate, but which had no crossover in the time when they were live, as additional new adverts. There are instances when for certain adverts, the job is not being advertised on any website before it reappears on another website again. We have decided to consider all such separate instances of adverts as distinct entries, if the period when the job is not on any website is longer than two days. This means there will be a new advert counted when it is posted again, if the posted date is more than two days after a period that a previous posting for the job has expired. The main benefit to this combined method is that it identifies adverts as being the same even when the text is not completely identical, and it may be more likely to pick up adverts that have not been filled and reposted.

Figure 2: On average, for every unique new advert there were around two additional duplicate adverts

Ratio of duplicate to unique adverts, UK, January 2017 to July 2024



Source: ONS, Textkernel

The majority of adverts captured in the dataset have duplicates. On average for a new unique advert coming into the dataset, there were two others identified as duplicate. For the snapshot metric this ratio is lower as adverts stay live beyond a single month.

The proportion of duplicates generally have declined when there were larger volumes of demand. This may be since adverts stay open longer when there is less demand, as employers may have already covered the costs to advertising. This relationship has changed since the second quarter of 2023, which may relate to some large website sources dropping off.

Total snapshot volumes after removing duplicates are still higher than official vacancy numbers. We know there are false positives — further duplicate adverts remaining in the dataset that future improvements may identify, but there are also some adverts currently identified as duplicates which should not be treated as such.

The issue of identifying duplicate adverts requires making assumptions about employer behaviour. For example, it is sometimes unclear whether employers are genuinely advertising for multiple posts across the country or sharing the same advert to multiple locations to be seen by employees. Currently such examples would typically be captured as duplicates and so removed, in our dataset. If employers modify the advert to try and reach different types of talent, it is also ambiguous at what point an advert is referring to a different post, and as we cannot ask the business, this may or may not be identified as a duplicate in our current method.

Future improvements in coverage

Future improvements should be able to identify a wider set of duplicate adverts, for example by matching on having the same salary in addition to having similar job descriptions. This will likely also be improved with further cleaning of appropriate adverts, though with a trade-off of potentially capturing more false negatives identified as duplicate adverts. We will also aim to provide further cleaning of adverts to approximate the conceptual similarity to vacancies. This would be looking into removing adverts for training courses as well as prospective adverts, additionally splitting out adverts where we see it is for multiple positions, and perhaps benchmarking to other official sources of business activity.

Data consistency

Adverts can drop off if the website they are on goes down, ceases to advertise jobs or is temporarily inaccessible, or the aggregator source considers the adverts to not be good quality for analytical use. These would all show declines in the snapshot statistics.

Similarly, a new website of adverts may be added to the aggregator source so such demand would only appear from that point on, showing a rise both in snapshots and new adverts statistics for any unique adverts from the new source.

Generally, no single source coming in or out of the dataset has a visible impact on the volumes, given there are many other sources compensating and there are many duplicates that appear across sources. In our time series up to August 2024, there are some exceptions users should be aware of:

- in May 2024, a large new source of adverts was brought in that increased both snapshot and new advert metrics

- in January 2024, a large source of adverts was reinstated after not receiving new adverts from it for the second half of 2023

- in July 2023, the vast majority of adverts from CareerBuilder were closed by Textkernel, so our snapshot volumes across July and August 2023 show a steep decline

- the Independent used to advertise jobs but over time there were less adverts on there and after May 2023, no new adverts were posted

We will continue to track such volatility to communicate to users monitoring short-term changes, including how this affects sub-national data.

Finally, we were made aware of an issue with the automated web scraping of adverts from websites with small numbers of postings. This issue occurred from November 2023 to July 2024 and resulted in adverts being shown as having been posted a few days later. In some cases, some adverts from such smaller sources will have been missed out, if they were posted and closed in a matter of several days. This may disproportionately affect such occupations that are filled extremely quickly. This effect should be minor but may show slightly different trends at the most granular level of detail we show, which is volumes for local authorities simultaneously with four-digit occupations.

SOC allocation

Matching

To assign an occupation code to each job advert, we standardise the advert’s job title by cleaning it, through the following steps:

- remove unnecessary words such as references to locations, hours and numbers (such as salaries) that do not reflect information about the occupation
- tidy up the text such as through the removal of punctuation and unnecessary whitespace, and converting words to lower case

We then look for the closest match in the Standard Occupation Classification (SOC 2020) coding index. The SOC coding index provides many job titles for each occupation. By way of example, for the occupation Secondary education teaching professionals, multiple standardised job titles are available to match on, for example, ” Secondary school teacher“, “Sociology teacher sixth form college” , “English teacher secondary school”, etc. The SOC coding index contains over 30,000 known job titles, each of which is matched to a corresponding occupation in the classification, and the whole index is regularly updated. See more information on the SOC 2020 classification. This provides a very good basis for matching the job titles found in job adverts.

To convert the job titles into a numeric representation, we apply the TF-IDF technique to groups of three letters (trigrams), rather than the full title. This provides some flexibility around spelling and word alternatives, and more meaningful terms within a longer job title can still find a good match to the SOC coding index. At the same time, there may still be misallocations where there is no close match in the index, or if shorter words are matched to longer words that contain them.

Method performance

To evaluate the overall performance of the SOC allocation algorithm, we assessed its accuracy on a manually labelled set of adverts. We took a random sample of 1,000 adverts, then three in-house coders assigned each advert in the sample an occupation. Not all adverts could be consistently matched across the separate manual assessment, something that has been seen in other approaches in academia. For performance assessment, we compare the occupations assigned from the model with those of at least one coder.

Overall, our assignment method was shown to have 65% precision when assigning an occupation at four-digit level from the manually labelled sample, while if adverts are grouped at the two-digit level the accuracy increased to 74%. Other approaches we have seen and evaluated generally show worse levels of precision for the granularity we produce.

The allocation method also provides a useful metric, which is the cosine-similarity to the corresponding nearest job title from the index. This metric shows the similarity of two pieces of text that have been converted into a numerical representation, on a scale between 0 and 1 where 1 is a complete match, and 0 is no similarity. To show the kind of matches our method made between job titles in online job adverts and those found in the SOC coding index, you can find table 3 in our associated data tables, which shows the most common clean job advert titles that are assigned to each four-digit occupation.

For example, adverts with the standardized job title “care assistant” match most closely to the job title “care assistant” from the 30,000 titles in the SOC coding index. In fact, this is an exact match and the cosine similarity metric is a 1. This job title from the coding index corresponds to the occupation care workers and home carers, so those adverts are assigned to that occupation. Another common job advert title was “Care Assistant – Bank – Care Home”, which also matched to the title “care assistant” from the index and hence the same occupation, with a lower cosine similarity of 0.740. This reflects the description has more combinations of letters that are not present in the matched text.

At the same time, we know there are some matches which can be improved. For example, with the occupation “waste disposal and environmental services managers”, the most common matched job title is “General Dentist”, which has matched on the SOC coding index entry “General dealer” which is in that occupation. This has a lower cosine similarity (0.624) but is still our closest match to the SOC coding index so it is our estimate. This highlights that our current method does not directly account for the importance of certain words in the overall assignment, and more complex methods that account for semantic meaning may improve our accuracy, and this is described in the future improvements section.

Similarity metric for understanding accuracy

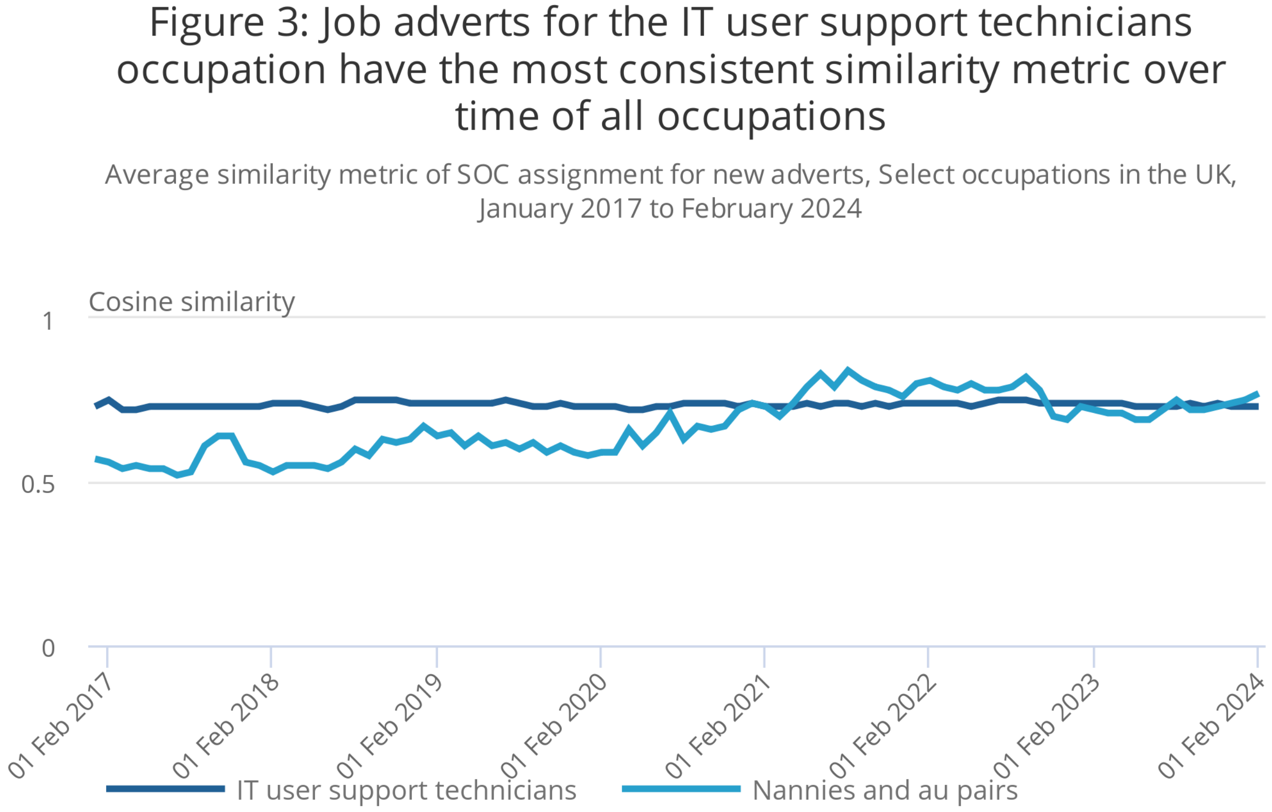
In accompanying Table 4 of the Quality metrics for Labour demand volumes by Standard Occupation Classification (SOC 2020) dataset we provide information on the average cosine similarity metric by occupation code and across time, so we can assess how the volatility of occupational demand may be influenced by the employers' wording of job advert titles.

Over time, there has not been much overall variability of adverts' average occupational assignment, with the lowest average similarity metric around 0.68 for assigned occupations, and the maximum around 0.71. In addition, over 99% of adverts have an assigned occupation, and the rate of assignment has improved further since early 2019.

However, there is more variability within occupations. Figure 3 shows the average similarity metric for IT user support technicians and Nannies and au pairs, which show the smallest and largest variation in the similarity metric over time respectively.

Figure 3: Job adverts for the IT user support technicians occupation have the most consistent similarity metric over time of all occupations

Average similarity metric of SOC assignment for new adverts, Select occupations in the UK, January 2017 to February 2024



Source: ONS, Textkernel

Occupations have different average similarity metrics, and their similarity metrics can be more or less stable over time. Adverts assigned to the nannies and au pairs occupation vary more over time, suggesting the assigned job titles vary more over time, than more stable occupations, such as for IT user support technicians.

For adverts which are assigned to occupations with a higher score, on average they should more accurately reflect those occupations. This means for a higher score in an occupation we can be more confident on average that there are at least that many in demand, as the metric tells us about adverts assigned to that occupation. However, this does not tell us of the direct accuracy of the total demand for the occupation. This is because there may be adverts which are being assigned to the wrong occupations, because they have a closer match to those occupations. We will track the volatility of this metric to help understand how it is coordinated with trends of labour demand, to report when changes may not be driven by economic trends but trends in employers' choice of job titling.

Future improvements of occupation assignment

We aim to iteratively improve this method in the future, ensuring the data quality continues to be improved. We welcome feedback from users to inform this.

One of the areas our method does not perform well is when the job title of the advert is not very informative. For this reason, we plan on using information from the full text to help us assign occupations. As the data we use would be a lot larger, we will need to apply methods to condense this information. Such methods may be able to capture the semantic meaning behind the text, which helps using more relevant information. Another area we could improve upon is the matching process – more redundant information could be removed prior to matching with further research, as well as exact matches being used when they are possible.

It is likely that improvements to our method will be employed in separate models, (for example, one model that uses the full text), and then we will use an ensemble model (a collection of all the separate models) to get a to a more informed overall assessment for the occupation assigned to each advert.

Geography allocation

Sometimes, interpreting the location of an advert can be a challenge, which is due to several reasons. Adverts may reference a location that is not to do with the location of the vacancy, for example, the headquarters of the company may be referenced. On occasion Textkernel's geography assignment method picks this up as the location of the advert. We also see instances where multiple locations are possible to work from, in which case the assignment method would pick one of them. If the job is specified as a working from home contract, it is similarly likely that Textkernel's method could pick up locations not associated with the location of the job.

Finally, many jobs that are based in London do not provide any specific location information, that is they specify "London" as the location. For this reason, we publish adverts for London as a whole rather than to local authority district level.

We will continue to identify ways we can improve the accuracy of the geography assigned to the advert.

6 . Related links

Measuring skill and qualification suitability in the UK labour market: user guide
User guide | Updated 5 November 2024
Supporting information for skill and qualification suitability estimates in the UK labour market. Methods used, data strengths, limitations, uses and users.

Human capital stocks estimates in the UK: 2004 to 2022
Bulletin | Released 19 March 2024
National and regional estimates of human capital stock in the UK from 2004 to 2022. Includes full and employed human capital estimates for each year.

7 . Cite this user guide

Office for National Statistics (ONS), released [5 November 2024], ONS website, user guide, Measuring labour demand volumes across the UK using Textkernel data: user guide