University of Oxford

Engineering Science

4YP Report

The Future of Work

Author

Terence Tan

Supervisor

Dr. Michael A Osborne

January 5, 2023

1 Introduction

The world population is ageing over the next few decades (Gerland et al., 2014). The rising elderly to working age population ratio is increasing and will continue to do so (World Health Organization, 2022). This trend is known as an ageing population, and will strain the public and social services of many countries around the world (Wiener and Tilly, 2002). As one of the key social challenges facing the world for the next few decades, it would be interesting to examine how an ageing population will affect the economy, and in particular, the job market and the interplay with automation in the workplace. As more workers age out of the workforce, automation is expected to make up for it (Frey and Osborne, 2013).

In this project, we aim to examine the relationship between the age distribution within occupations and the degree of automation (Frey and Osborne, 2013) of those occupations. Although similar work has been done on this topic (Basu et al., 2018), the study only looked at broad categories of employment. In this project, we will zoom in to look at specific occupations. We might also look into any correlations with the skills/knowledge required for those occupations. This will all be done using the scikit-learn library[1] in Python. Specifically, we will look at a Bayesian non-parametric machine learning technique known as Gaussian Process (Ghahramani, 2013); this model was used in previous work (Frey and Osborne, 2013), and so, would be a good model to start with. We will test and validate against different models and pick the best performing ones.

2 Dataset

We used two main metrics for this project: the automatability of occupations, and the age distribution within occupations. The dataset for the former is provided in an earlier work by Frey and Osborne, 2013. The latter can be found in datasets provided by the US Bureau of Labour Statistics[2] (BLS); there is one dataset for each year from 2011 to 2021. All the datasets mentioned above use the Standard Occupational Classification (SOC) to classify the occupations, which means that we can map from one dataset to the another using the SOC codes[3]. However, it is necessary to perform some data wrangling before we can proceed with the mapping. Additionally, changes were made to the SOC in 2018, so we would have to standardise all the datasets. In the following sections, we shall examine the datasets and the required data wrangling in more detail.

2.1 BLS Dataset

As mentioned in Chapter 2, the BLS provides one dataset for each year from 2011 to 2021. The datasets from 2011 to 2019 follow the old SOC while the 2020 and 2021 ones follow the updated version.

[1] https://scikit-learn.org/stable/
[2] https://www.bls.gov
[3] https://www.bls.gov/soc/2018/soc_structure_2018.pdf

We want to standardise everything according to the updated SOC. We first label each dataset with the respective year and concatenate all of them along the row axis; we shall refer to this concatenated dataset as the BLS dataset for the rest of the paper. A section of the BLS dataset can be seen in Figure 1. Note that the numbers under the *Total* and age group columns are in thousands. Furthermore, the median age is not provided for all occupations, which makes it less useful as a metric. Hence, we will not be using it in this paper.

	Occupation	Total	16-19	20-24	25-34	35-44	45-54	55-64	65<=	Median age	Year
0	management, professional, and related occupations	64744.0	420.0	3267.0	15222.0	15625.0	14238.0	11394.0	4579.0	43.8	2021
1	management, business, and financial operations...	27864.0	100.0	1052.0	5726.0	6783.0	6603.0	5411.0	2189.0	45.5	2021
2	management occupations	18986.0	74.0	573.0	3413.0	4728.0	4704.0	3863.0	1630.0	46.5	2021
3	chief executives	1664.0	1.0	4.0	157.0	388.0	446.0	464.0	204.0	51.6	2021
4	general and operations managers	1085.0	2.0	30.0	258.0	303.0	272.0	173.0	47.0	43.4	2021
...	...	...	...	...	...	...	...	...	...	...	...
6259	pumping station operators	21.0	0.0	3.0	6.0	4.0	3.0	5.0	0.0	-	2011
6260	refuse and recyclable material collectors	92.0	2.0	12.0	22.0	16.0	24.0	12.0	4.0	41.3	2011
6261	mine shuttle car operators	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	-	2011
6262	tank car, truck, and ship loaders	3.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	-	2011
6263	material moving workers, all other	62.0	3.0	7.0	6.0	19.0	12.0	13.0	2.0	43.1	2011

6264 rows × 11 columns

Figure 1: BLS dataset (before processing)

While the BLS did provide documents[4] outlining and explaining the changes to the SOC, it is generally too vague to be anything more than a rough guide. Furthermore, some of the changes made to the SOC are fairly complex. In addition to that, the BLS collected data differently for some occupations after 2019. For example, both 'Marketing Managers' (SOC code: 11-2021) and 'Sales Managers' (SOC code: 11-2022) are classified under 'Marketing and Sales Managers' (SOC code: 11-2020). From 2011 to 2019, the BLS only collected data for 'Marketing and Sales Managers' while they collected data for 'Marketing Managers' and 'Sales Managers' separately in 2020 and 2021. While this represents more detailed data, it is inconsistent with data collected in previous years.

In order to list out all the changes and inconsistencies, we use the *pandas.DataFrame.join* function to join an old SOC dataset (from 2011 to 2019) with an updated SOC dataset (from 2020 to 2021) using the *Occupation* column. We can then obtain a list of occupations from the old SOC dataset which did not join, and a corresponding list for the updated SOC dataset. We then manually go through both lists and decide on how to standardise the BLS dataset. While this process is tedious, it is reasonably doable since each list only contains about a hundred rows. The changes and rationale for them are listed alongside the occupations in both lists. All of these are placed in an Excel file[5].

The list of actions required are as follows: -, Delete, Change, Combine, Combine but keep. The dash indicates that no action is required. 'Delete' means to delete the occupation; this is usually because the particular occupation no longer exists under the new SOC. 'Change' indicates a name

[4] https://www.bls.gov/soc/2018/home.htm

[5] https://github.com/terencetan-c/4YP-The-Future-of-Work/blob/main/Data%20cleaning/Changes.xlsx

change. 'Combine' indicates that two or more occupations should be combined into the overarching occupation. For example, the two occupations mentioned before, 'Marketing Managers' and 'Sales Managers', will be combined into 'Marketing and Sales Managers' to ensure consistency in the BLS dataset across the years. This will basically be an element-wise addition of the rows, involving only the *Total* and age group columns. This is another reason why we dropped the *Median age* column since we have no way of combining median values for the BLS dataset. Lastly, the 'Combine but keep' action is used in cases where we have to combine to maintain consistency but are still able to preserve some granularity by keeping the original rows. For example, the old SOC classifies the four occupations 'Home Health Aides' (31-1011), 'Psychiatric Aides' (31-1013), 'Nursing Assistants' (31-1014), and 'Orderlies' (31-1015) under 'Nursing, Psychiatric, and Home Health Aides' (31-1000 and 31-1010). Additionally, the old SOC also has 'Personal Care Aides' (39-9020 and 39-9021) classified separately. The new SOC renamed 'Nursing, Psychiatric, and Home Health Aides' to 'Home Health and Personal Care Aides; and Nursing Assistants, Orderlies, and Psychiatric Aides' (and changed the SOC code from 31-1000 to 31-1100) and moved 'Personal Care Aides' (now 31-1122) under this newly named occupation. Another thing to note is that the datasets following the old SOC only collected data of 'Nursing, Psychiatric, and Home Health Aides' as a whole instead of the four occupations individually. They also collected data for 'Personal Care Aides'. On the other hand, the datasets following the new SOC collected data for the four occupations, 'Home Health Aides' (now 31-1121), 'Psychiatric Aides' (now 31-1133), 'Nursing Assistants' (now 31-1131), and 'Orderlies' (now 31-1132), and the newly moved occupation, 'Personal Care Aides', separately. Note that both groups of datasets have data of 'Personal Care Aides' on its own, and we would like to keep it that way to preserve granularity of the data. For the datasets following the old SOC, we would apply 'Combine' on 'Nursing, Psychiatric, and Home Health Aides' (effectively just a name change in this case) and 'Combine but keep' on 'Personal Care Aides'. For the new SOC datasets, we apply 'Combine' on the four occupations and 'Combine but keep' on 'Personal Care Aides'. This way, we end up with data for a combined 'Nursing, Psychiatric, and Home Health Aides' to 'Home Health and Personal Care Aides; and Nursing Assistants, Orderlies, and Psychiatric Aides', while simultaneously still retaining 'Personal Care Aides'.

Having systematically gone through all the inconsistencies and indicating one of the five actions required for the inconsistencies, we then use Python to automate the standardisation process. This gives us the standardised BLS dataset.

One more thing to note is that the occupations in the BLS dataset are not labelled with their respective SOC codes. This is easily rectified once the above data wrangling is completed by joining (on *Occupation*) the BLS dataset with the list of SOC codes to map from occupation name to code.

3 Preliminary Findings

References

Basu, Meghna et al. (2018). "The twin threats of aging and automation". In: *Marsch & McLennan Companies, Mercer*.

Frey, Carl Benedikt and Michael Osborne (2013). "The future of employment". In.

Gerland, Patrick et al. (2014). "World population stabilization unlikely this century". In: *Science* 346.6206, pp. 234–237.

Ghahramani, Zoubin (2013). "Bayesian non-parametrics and the probabilistic approach to modelling". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1984, p. 20110553.

Wiener, Joshua M and Jane Tilly (2002). "Population ageing in the United States of America: implications for public programmes". In: *International journal of epidemiology* 31.4, pp. 776–781.

World Health Organization (Oct. 2022). *Ageing and Health*. URL: `https://www.who.int/news-room/fact-sheets/detail/ageing-and-health`.