University of Oxford

Engineering Science

4YP Report

The Future of Work

Author

Terence Tan

Supervisor

Professor Michael A Osborne

March 19, 2023

Abstract

I would like to express my deepest gratitude to Professor Michael Osborne for supervising me throughout this project. His insights and guidance are very much appreciated. This project would not have been possible without the dataset provided by him.

Acknowledgements

I would like to express my deepest gratitude to Professor Michael Osborne for supervising me throughout this project. His insights and guidance are very much appreciated. This project would not have been possible without the dataset provided by him.

1 Introduction

Technological advancement is widely believed to be the primary driving force behind economic growth (**RePEc:ssa:lembks:dosietal-1988**). At the same time, there is potential for technological displacement of labour. This concept of 'technological unemployment' was first introduced by David Ricardo in the 19th century (**WoirolGregoryR1997Ttua**), who wrote that he had become "convinced that the substitution of machinery for human labour, is often very injurious to the interests of the class of labourers" (**10.1257/jep.33.2.229**). This idea was further explored by John Maynard Keynes, who blamed "our discovery of means of economising the use of labour outrunning the pace at which we can find new uses for labour" for potential widespread technological unemployment (**Keynes2010**). However, there are those that are cautiously optimistic of the impact of technological advances on labour. Prominent member of the United Automobile Workers union Walter Reuther and his colleague had hopes that automation could eliminate the drudgery of industrial work and ultimately allow workers to pursue leisurely interests. Yet, even they shared fears that automation could lead to widespread structural unemployment if not managed properly (**SteigerwaldDavid2010WRtU**). Alas, history seems to have validated their fears; the proportion of manufacturing employment in the US among non-agricultural workers decreased from 32% in 1955 to 8% in 2019 (**rose*2021**). That being said, there is no consensus on the impact of technological advances on the decline of the manufacturing sector relative to other factors such as globalisation and offshoring (**RoseElizabethL.2021TDoU**) (**krugman2019globalization**).

At the same time, ageing population is a social issue that has become increasingly relevant all over the world (**2002Wpa1**).

1.1 Ageing Population

The world population is ageing over the next few decades (**science**). The rising elderly to working age population ratio is increasing and will continue to do so (**WHO**). This trend is known as an ageing population, and will strain the public and social services of many countries around the world (**publicservicesstrain**). 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 (**futureofemployment**).

Related Works

Research into the social implications of an ageing population had been carried out as far back as 2002. **tinker2002social** outlined the demographic trends around the turn of the millennium that pointed to the future of an ageing world, and highlighted the falling potential support ratio (ratio of

people aged 15-64 to people aged 65 and above) around the world, which will affect the distribution of resources, such as in the case of pensions, of a country. This study also talked about the relative power between the young and old, and pointed out that the older generation will have larger share of votes, potentially wielding more political power. Indeed, contemporary authors, such as **Munger+2022**, have noted this power struggle between the older and younger generations. The US government has also been described as a gerontocracy (**noah˙2019**) (**thompson˙2020**) in modern times. The numbers back this up, with the average age of a US senator being 64 in 2021 (**manning˙2022**).

Cheng2020 conducted a global analysis of population ageing and mortality between 1990 and 2017, and concluded that there was a pattern of higher disease-related deaths due to population ageing around the world within that time period. The study recommended policies aimed at encouraging healthy ageing.

Cross-country comparison

In the US, there is evidence to suggest that the country's healthcare system is not prepared to meed the increasing demands of the ageing population (**foley˙retooling˙2020**).

While this paper will be focusing on the US, there are numerous other studies looking into ageing population of the US as well as other countries and regions of the world. China is expected to age rapidly over the next few decades (**BeardsonTimothy2021Ag:C**), and studies such as **LuoYanan2021TaCf** examined this phenomenon in the context of the country.

1.2 Project Overview

In this project, we aim to examine the relationship between the age distribution within occupations and the degree of automation (**futureofemployment**) of those occupations. Although similar work has been done on this topic (**twinthreats**), 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 (**GaussianProcess**); this model was used in previous work (**futureofemployment**), and so, would be a good model to start with. We will test and validate against different models and pick the best performing ones.

We will be using the proportion of elderly people as a metric for measuring the 'age' of occupations. We acknowledge that the definition of elderly age varies across different countries and cultures (**ageingculture**), and that the ageing process varies for different people depending on a variety of factors (**levine2013modeling**)(**hayflick2007biological**). In fact, there are studies looking into moving beyond using chronological age to define an elderly person (**KotterGrhn2015**)(**SOTOPEREZDECELIS2018e**

[1] https://scikit-learn.org/stable/

However, there are numerous issues surrounding this approach (**jylhava2017biological**), such as validation of such results are often difficult (**biologicalagedifficult**), resulting in little consensus on an alternative metric to chronological age. For this reason, we will stick to the conventional definition of 65 years and older as 'elderly' (**who2010definition**)(**orimo2006reviewing**)(**oecddata**). This definition is also convenient for us since the oldest age group featured in our datasets are 65 years and older. Hence, it would make sense to have the metric for measuring the 'age' of an occupation be the ratio of people aged 65 years and older to the total number of people in that particular occupation. We will refer to this metric as the Elderly Proportion (EP). However, we also recognise that this definition of 'elderly' coincides with the retirement age of the US (**MunnellAliciaH2013SSRR**). Hence, using this metric alone might lead to misleading results since we would expect people in that age group to leave the workforce. It might be better to use the proportion of 55 years and older as a metric, i.e. the ratio of people aged 55 years and older to the total number of people, since this figure might be more robust to the effects of retirement on labour force participation. We shall refer to this metric as the Old Proportion (OP). Having said that, the Elderly Proportion would still prove to be an interesting metric to investigate. For example, if there is a general trend of increasing EP over the years, that would be a sign of an ageing labour force in spite of the effects of retirement. Therefore, we shall examine both of these metric in this paper.

Related Works

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 **futureofemployment**. 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. 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

[2]https://www.bls.gov
[3]https://www.bls.gov/soc/2018/soc_structure_2018.pdf
[4]https://www.bls.gov/soc/2018/home.htm

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 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'

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

(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.

2.2 Automatability Dataset

This dataset (which we shall refer to as Automatability dataset) was obtained from **futureofemployment**, and features 702 Detailed Occupations. For each of these occupations, a Probability of Computerisation[6] had been calculated. We shall refer to this probability as PCom. Other variables are included as well, such as the skills associated with each occupation and a Category Label. These were used to calculate the PCom, but we will just focus on the PCom in this paper.

[6]Defined as 'job automation by means of computer-controlled equipment' by **osborne2017future**

3 Preliminary Findings

In order to make sense of how well the BLS dataset represents the US labour force, we plot the ratio of the total labour numbers provided by the BLS dataset for each year to the total US civilian labour force[7] for that year. We do that for *Major Group*, *Minor Group*, *Broad Group*, and *Detailed Occupation* separately. The resulting plots can be seen in Figure 2. Clearly, *Major Group* occupations are most representative of the US civilian labour force, with *Minor Group* being the least. Looking through the BLS dataset, this makes sense since the BLS tended to mostly collect high-level data (*Major Group*) and low-level data (*Broad Group* and *Detailed Occupation*). Another thing to note is that many *Broad Group* occupations only contain a single *Detailed Occupation* which also shares the same occupation name (for example, *Chief Executives*: 11-1010 and 11-1011). Hence, it is not surprising that the ratios for *Broad Group* and *Detailed Occupation* are so similar.

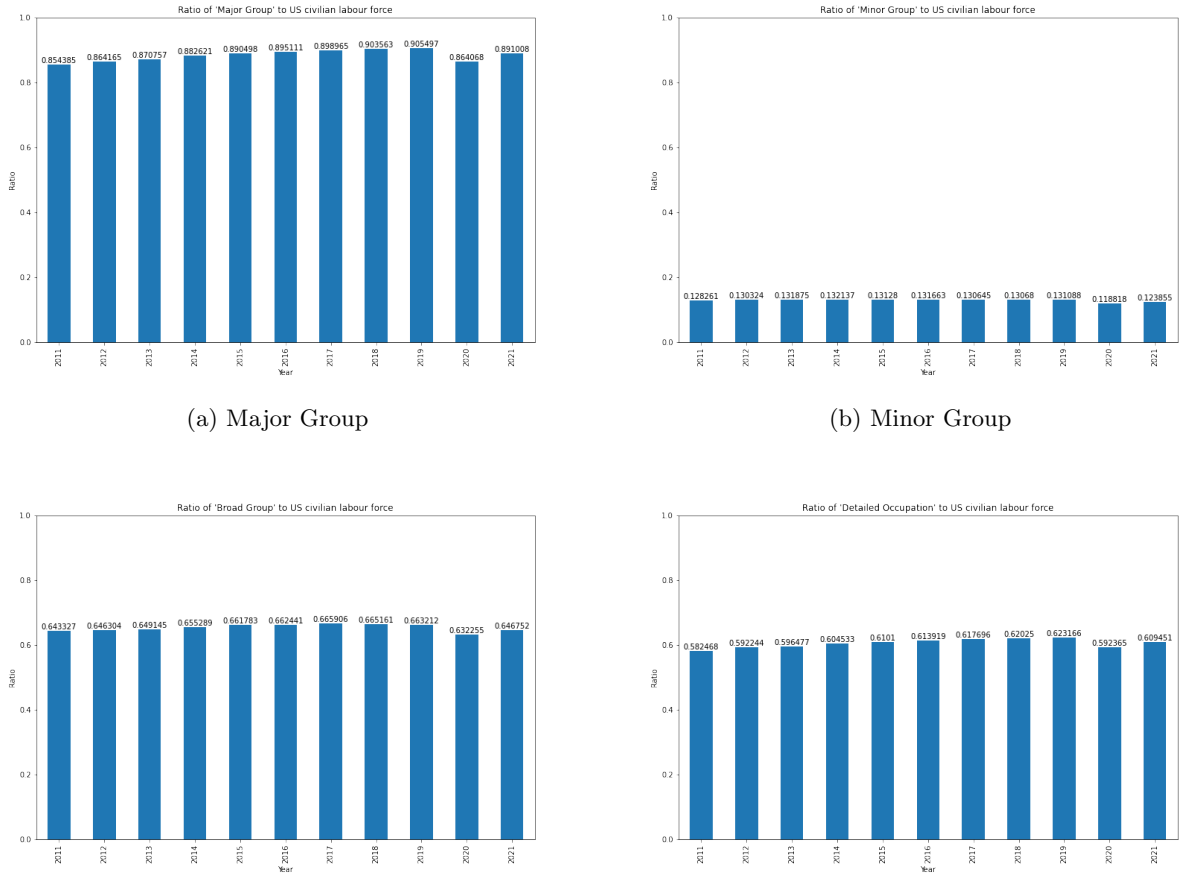
It is important to consider the fact that the US civilian labour force includes both the employed and the unemployed. In years with unusual levels of unemployment rate, the ratios will be distorted and paint a misleading picture. Indeed, we see in Figure 2 that there is a considerable dip in the ratios in 2020 relative to other years, coinciding with the onset of the COVID-19 pandemic (**covid2020unemployment**)(**congresscovidunemployment**). Plotting the ratios relative to the employed portion of the US civilian labour force accounts for this; as seen in Figure 3, the dip in 2020 is no longer present. We can also see that the unemployment rate did not distort the ratios significantly. Hence, our conclusion from before still holds: the *Major Group* is most representative of the US civilian labour force. With that in mind, we will try to use *Major Group* data as much as possible and exercise caution when using *Detailed Occupation* and *Broad Group* data. As for *Minor Group* data, we will neglect it given its low representation of the labour force.

3.1 General Trends

We average the EP (refer to Chapter 1.2 for definition) over the Major Groups for each year, and plot the values against the years to obtain Figure 4a. We do the same for OP, and get the plot in Figure 4b. We see a steady increase over the years for both plots, which is not surprising given the ageing population of the US as discussed in Chapter 1.1. We can examine these plots in further detail by plotting the EP/OP for each of the 21 Major Groups against the years to obtain Figure 5; we can see that there is generally an increase across the Major Groups. We do not break down the plots into further detail (for example looking at individual Detailed Occupations) since that will result in plots that are too messy to give us any useful insights.

That being said, it is rather tricky to make sense of Figure 5 given that there are 21 individual

[7] https://www.bls.gov/cps/cpsaat01.htm



(a) Major Group

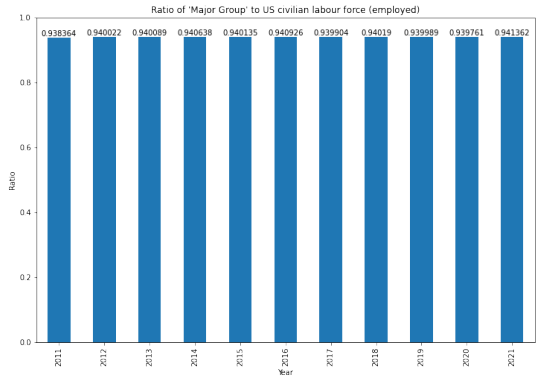
(b) Minor Group

(c) Broad Group

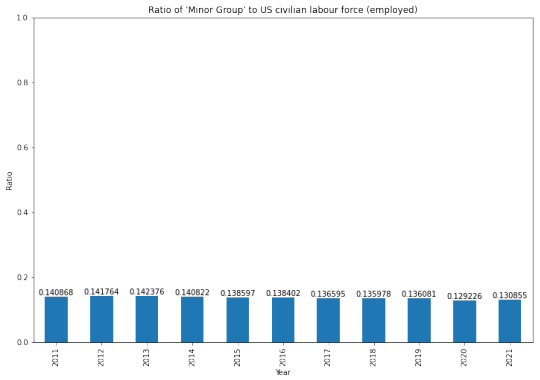
(d) Detailed Occupation

Figure 2: Ratio of the various SOC categories to the US civilian labour force

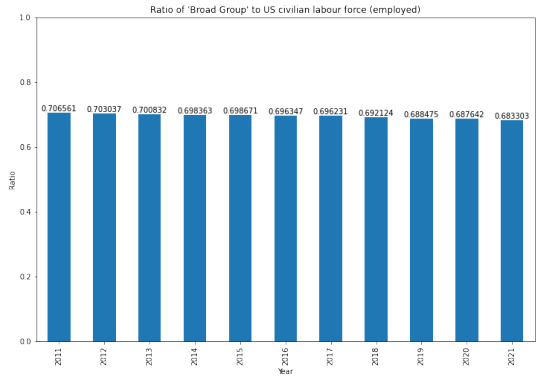
plots within each of the two figures. Hence, we can calculate the relative change in EP/OP each Major Group occupation from 2011 to 2021 to obtain Figure 6. The 'construction and extraction occupations' has the highest relative increase while the 'personal care and service occupations' features the least.



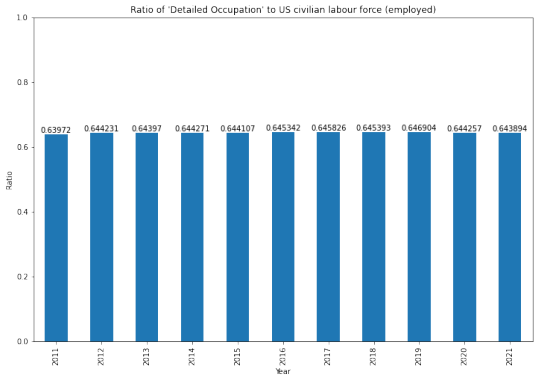
(a) Major Group



(b) Minor Group

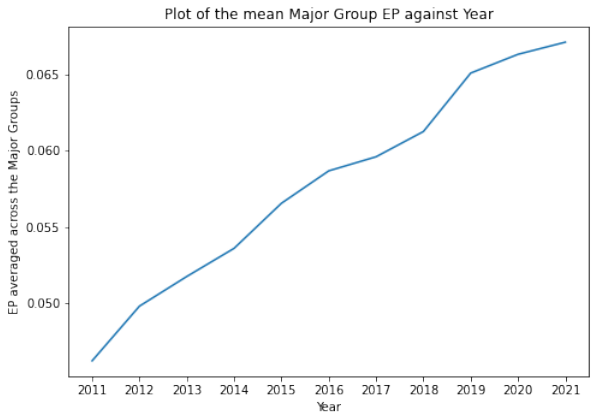


(c) Broad Group

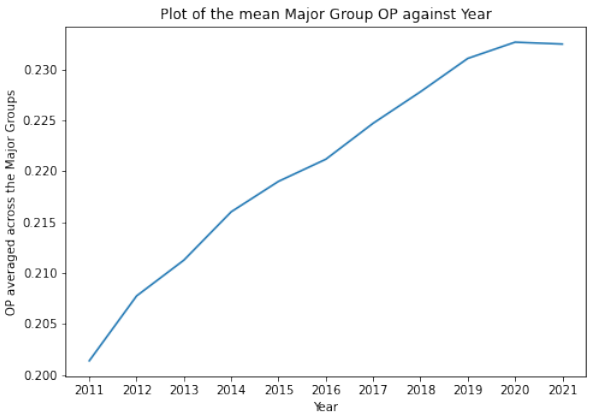


(d) Detailed Occupation

Figure 3: Ratio of the various SOC categories to the US civilian labour force (employed)



(a) Elderly Proportion

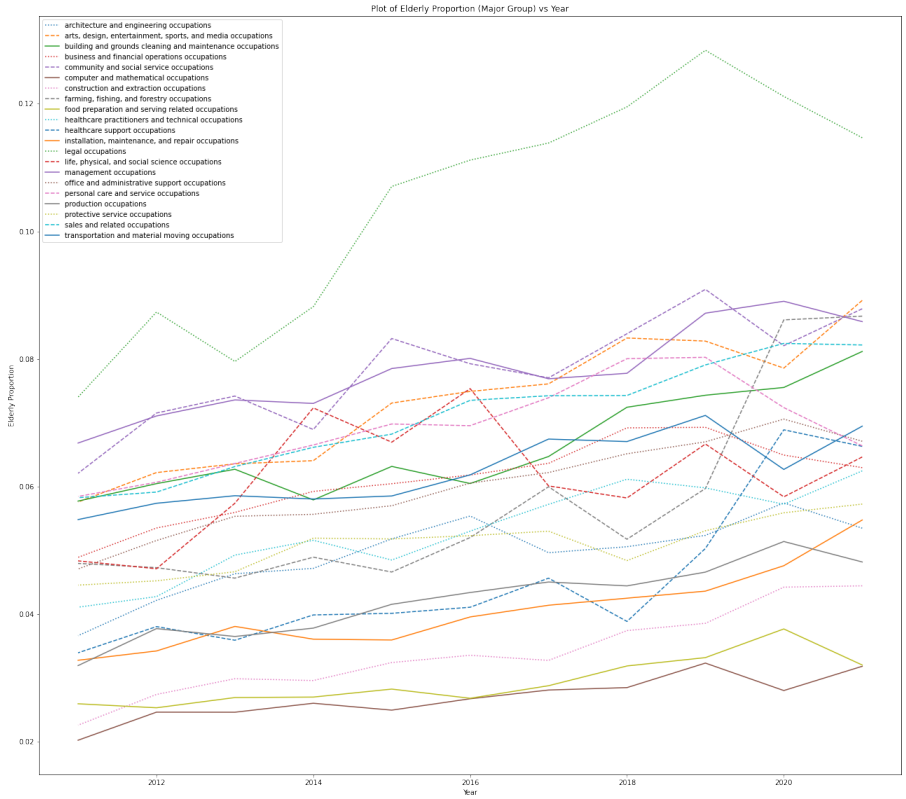


(b) Old Proportion

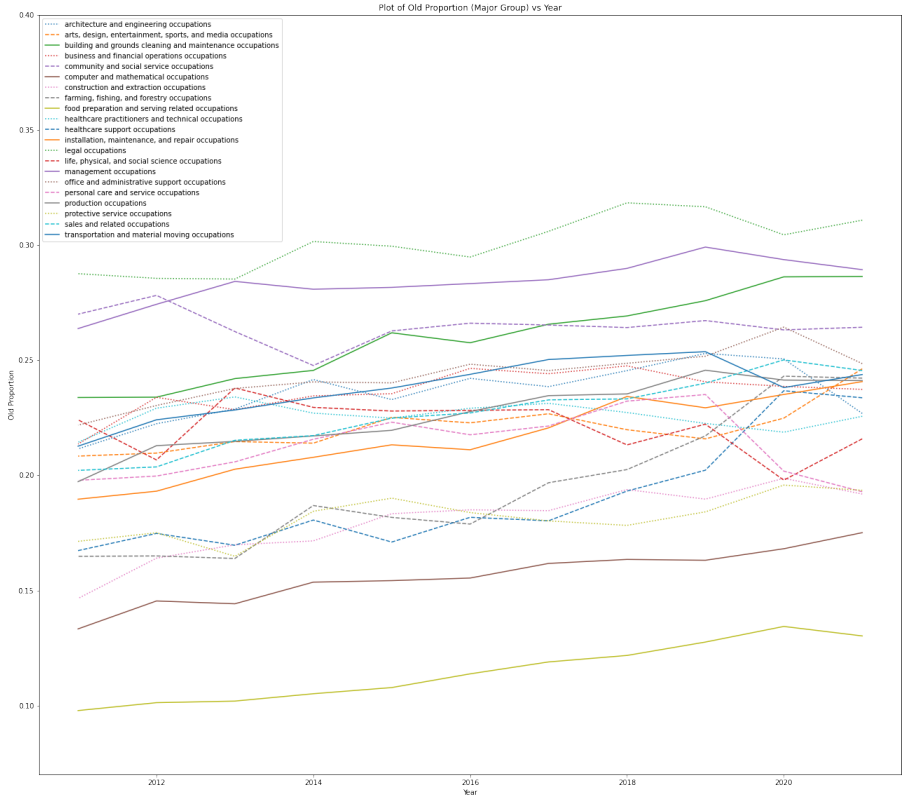
Figure 4: Plot of EP/OP (averaged over the Major Groups) against Year

4 Data Analysis

Now that we have our processed BLS dataset (with the calculated OP and EP), we can use *Pandas.merge* to join it with the automatability dataset (which includes the Probability of Computerisation) from **futureofemployment** based on the Detailed Occupation. This gives us a joint data

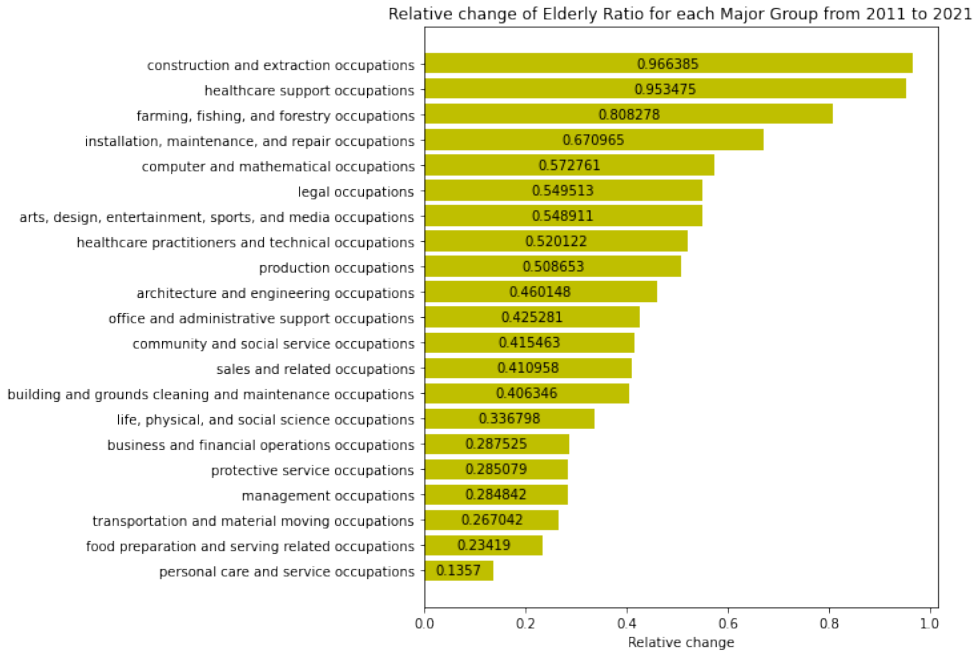


(a) Elderly Proportion

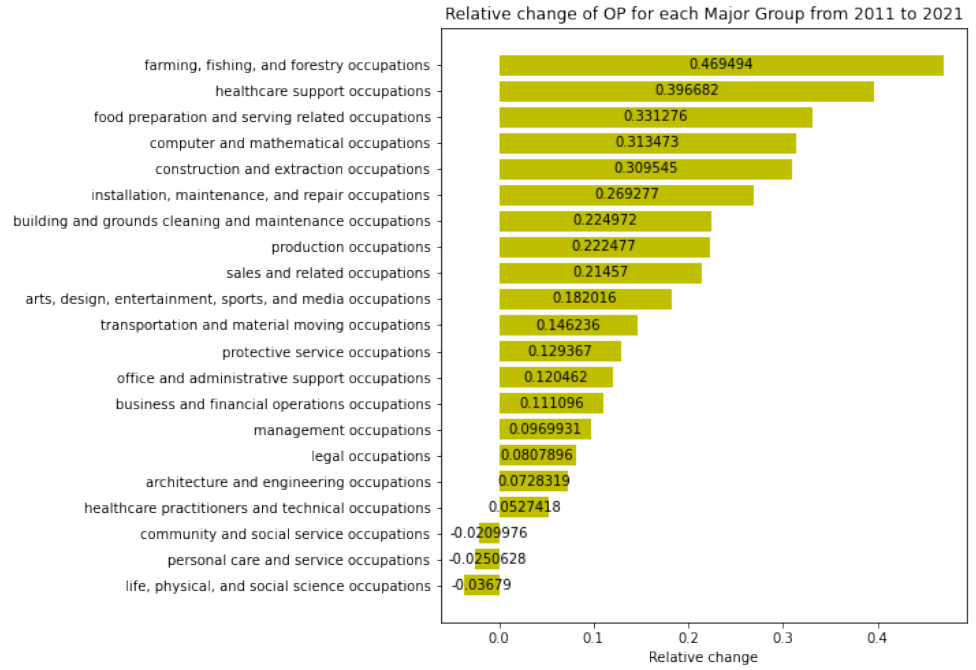


(b) Old Proportion

Figure 5: Plot of EP/OP (for each Major Groups) against Year. There is a general increase across most of the Major Group occupations, which is expected given the ageing population in the United States.



(a) Elderly Proportion



(b) Old Proportion

Figure 6: Relative change of EP/OP for each Major Group from 2011 to 2021. This shows us which occupations experienced the most and least rapid ageing during the time period of 2011 to 2021.

set that we will refer to as *joint_auto*. Unfortunately, the BLS dataset does not feature a full list of all the Detailed Occupations, so we end up with a reduced set of Detailed Occupations in the *joint_auto* dataset. As can be seen in Figure 7, *joint_auto* only represents about 40% of the total US civilian labour force, which is still a significant amount. However, there is the question of whether *joint_auto* is a representative sample of the population, i.e. the US civilian labour force. We shall analyse this question by treating it as a population sampling problem.

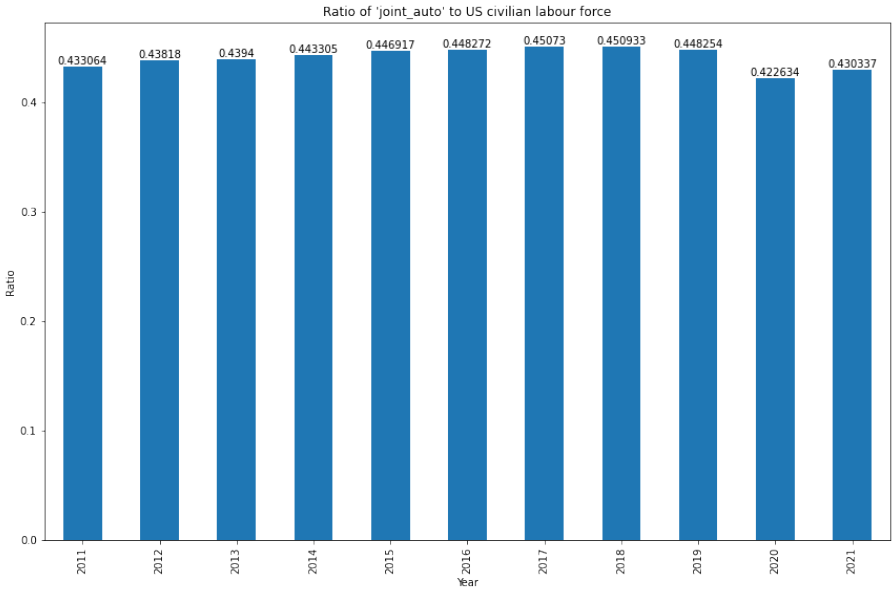


Figure 7: Plot of ratio of *joint_auto* to US civilian labour force against Year. We can see that our (joint_auto) dataset represents a significant proportion of the total labour force, but not enough for us to generalise any trends we may find in our dataset. We will have to investigate how well our dataset represents the actual US civilian labour force.

	Population		**Sample**	
Variable	Mean	Variance	Mean	Variance
EP	0.0578	0.000394	0.0516	0.00334
OP	0.220	0.00190	0.223	0.0122
PCom	0.536	0.136	0.508	0.143

Table 1: Population/Sample mean and variance. We can see that the means and variances are very similar between the sample and the population. However, we note that the EP/OP values for the sample have higher variances compared to that of the population.

4.1 Checking biasness

As mentioned earlier, we shall treat *joint_auto* as the sample and the US civilian labour force as the population. We shall use the Major Group occupations from our processed BLS dataset to represent the US civilian labour force since we have established its high representation in Chapter 3.

We start off with a simple mean and variance comparison. We find the mean and variance of the OP, EP, and PCom values for both the sample and population, which are shown in Table 1. We can see that the means and variances match quite well for the PCom variable. For OP and EP, the means match quite closely, but the variances are off by an order of magnitude; the sample have a higher variance than the population for both OP and EP. This means that the sample have more spread-out values for OP and EP, so we must exercise caution when generalising any findings from the sample to the population. Otherwise, the sample seems fairly representative of the population. That being said, the mean and variance analysis is too simplistic to give us any concrete conclusions. Hence, we need a more sophisticated measure of the sample's representativeness of the population.

We can look at the proportion of the total number of people employed within each Major Group with respect to the total US civilian labour force for each year. For example, Management Occupations represent 13.2% of the US civilian labour force in 2021, 13.4% in 2020 and so on. Transportation and Material Moving Occupations represent 6.34% in 2011, 6.38% in 2012 and so on. We put all of this information into one vector, which would represent the population proportion. We then map all the occupations (which are Detailed Occupations) in *joint_auto* back to their respective Major Groups, and sum up the number of people employed in each of those Detailed Occupations within the Major Groups for each year. These numbers are then divided by the total number of people employed in *joint_auto* for each year. This would tell us the proportion of each of the 21 Major Groups within *joint_auto* for each year; this sample proportion information would be placed in another vector. We want to see how well the sample proportion vector matches the population proportion vector, so we subtract the former from the latter and plot the result in Figure 8. We can see that the values along the y-axis are all fairly small. Just to illustrate this point further, we plot the sample proportion against the population proportion in Figure 9. Ideally, we would like the plot to be a straight line with a gradient of 1 and a y-intercept of 0, and we can see that our plot does closely resemble such a line. Indeed, fitting a best-fit line to the plot confirms this as well, with the best-fit line having a gradient of 1.18 and a y-intercept of -0.00862. Hence, the sample has a similar makeup to the population in terms of the relative weightage of each of the Major Groups.

Given our above biasness tests, we can make the reasonable assumption that our sample is fairly representative of the population, and any findings obtained from the sample can be generalised to the population to a certain extent.

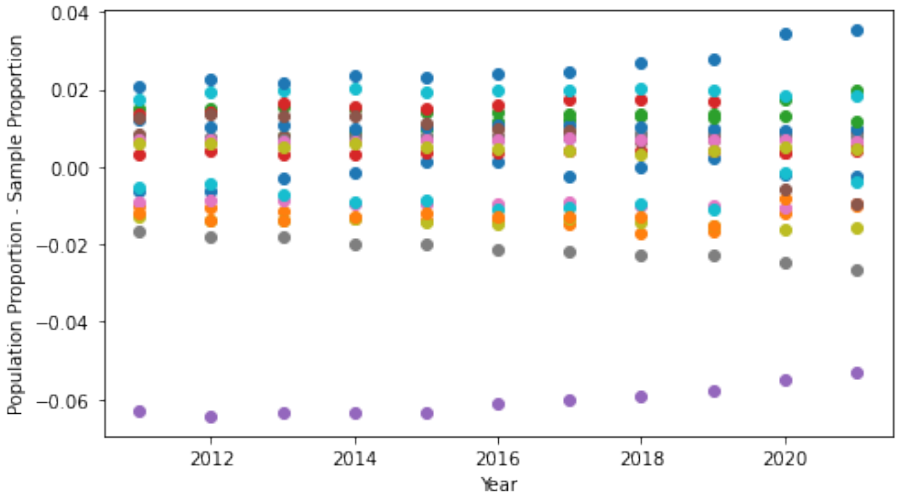


Figure 8: Plot of the difference between the population proportion and the sample proportion against Year. The different colours represent different Major Groups. The low values along the y-axis shows that the relative weightage of each of the Major Groups within the sample is similar to that of the population. Hence, the sample has a similar makeup to the population in terms of the Major Groups.

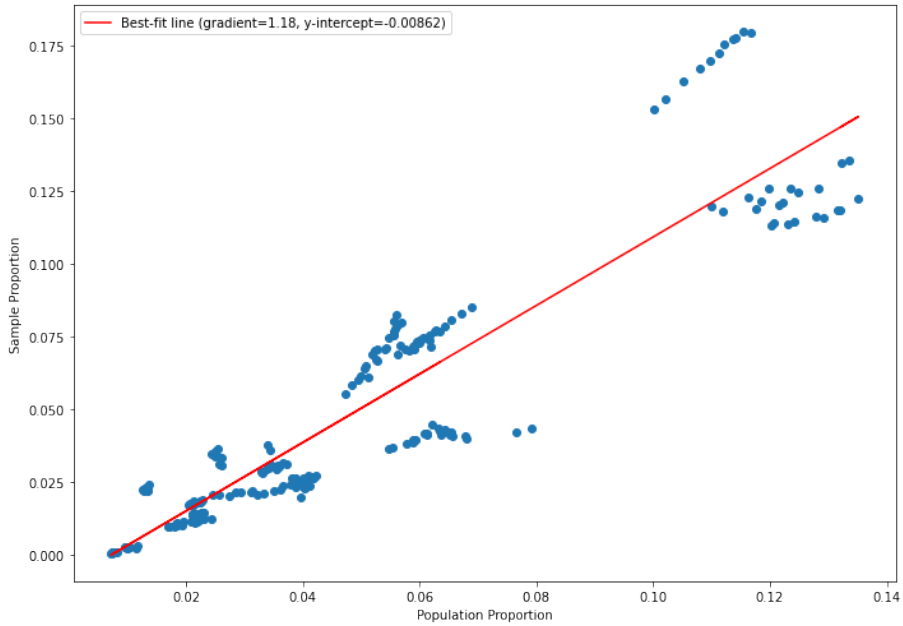


Figure 9: Plot of the sample proportion against population proportion (with a best-fit line). We can see that our plot roughly follows a straight line with a gradient of 1 and y-intercept of 0.

Type of correlation	**EP**	**OP**
Pearson	-0.00363	-0.0321
Spearman	-0.0151	-0.0661
Kendall	-0.00881	-0.0429

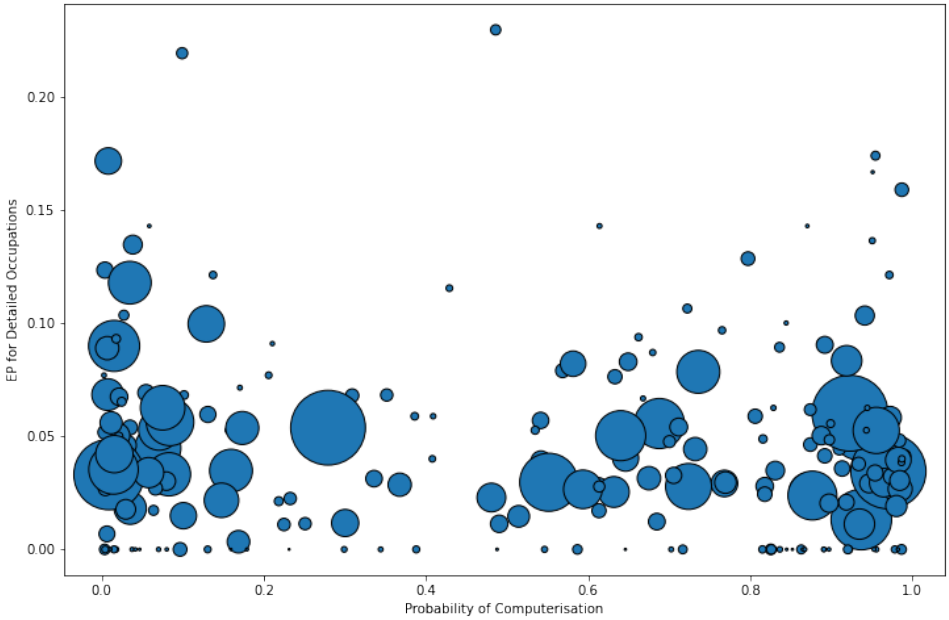
Table 2: Correlation coefficients for Figure 10

4.2 Exploration of Data

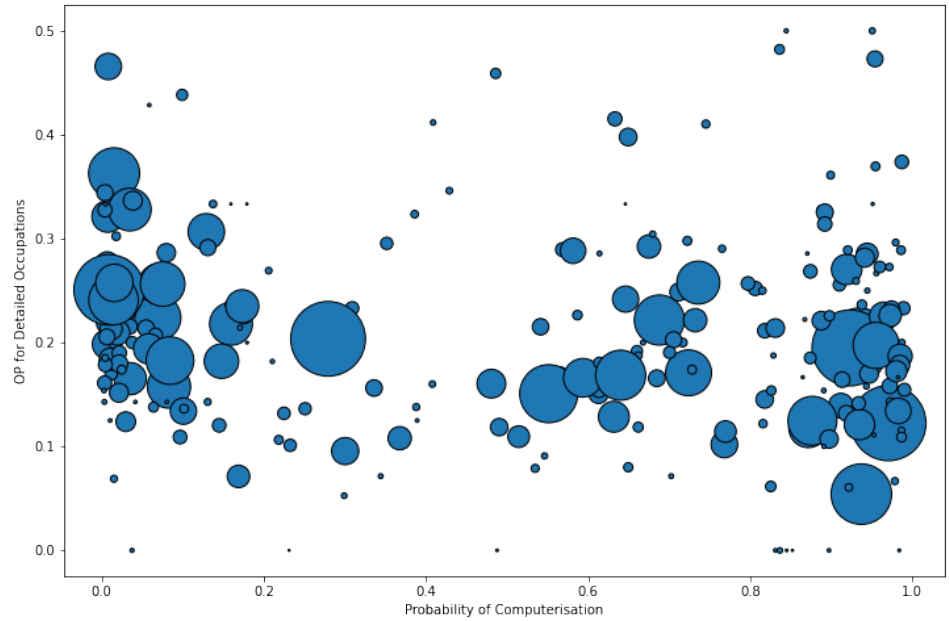
Now that we have established the representativeness of the sample, we can go about exploring the dataset.

EP/OP against PCom

The first thing we do is to create a scatterplot of the Detailed Occupations within *joint_auto*, with the EP/OP values (for 2012 only since the PCom values were calculated based on 2012 data) on the y-axis and the PCom values on the x-axis. We also scale each data point relative to the total number of people employed within that occupation, i.e. occupations with more people will be bigger on the plot. This gives us Figure 10, which we can see has no obvious trends. Zooming in on specific regions of the scatterplot (e.g. the region of high PCom) also yields no significant patterns. We also calculated the correlation coefficients, which can be found in Table 2. We do note that the majority of employed workers are concentrated in either the low PCom region or the high PCom region, with relatively few in between. This is consistent with the findings from **osborne2017future**, and is another piece of evidence that our assumption that *joint_auto* is fairly representative of the US civilian labour force is reasonable as discussed in Chapter 4.1.



(a) Elderly Proportion



(b) Old Proportion

Figure 10: Plot of EP/OP (for each Detailed Occupation) against PCom. There seems to be no obvious patterns or trends present.

Our next step of exploration is to take the base 10 logarithmic[8] of EP/OP and PCom, and plot

[8]as can be seen in Figure 10, the occupations with zero values for EP/OP have relatively few people employed within them, and we can remove them beforehand to avoid negative infinite values without affecting the representativeness of our dataset too much

Type of correlation	EP	OP
Pearson	0.00579	-0.113
Spearman	-0.0151	-0.0661
Kendall	-0.00881	-0.0429

Table 3: Correlation coefficients for Figure 11

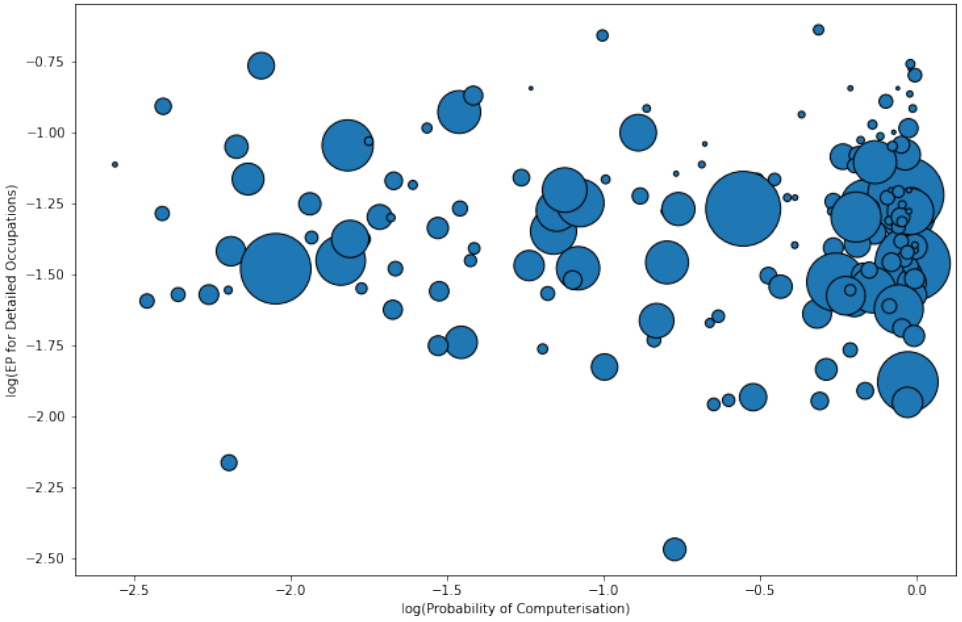
the former against the latter in a scatterplot, as shown in Figure 11. Interestingly, the $\log_{10}$(OP) plot in Figure 11b seems to show a slight downward trend as $\log_{10}$(PCom) increases; plotting the scatterplot on a logarithmic scale seems to have revealed a previously hard-to-spot trend. We also calculated the correlation coefficients for this particular plot (along with the corresponding plot for EP), which is given in Table 3. This trend suggests that 'older' occupations tend to be less likely to be computerised. This makes intuitive sense for occupations such as management; **osborne2017future** notes that management occupations tend to be at low risk of computerisation due to the high degree of social intelligence required for them, and people in management positions would also tend to be older since such occupations (especially the ones like Chief Executive Officer) would be biased towards those with more work experience. There is also some evidence that high skilled workers, who would generally be less likely to have their jobs computerised, tend to retire later than low skilled workers (**HimmelreicherRalfK.2009Saao**).

Plotting the dataset grouped by Category Labels or Major Groups do not seem to yield any interesting results either. In fact, there are too few datapoints for some of the Labels and Major Groups for us to draw any meaningful conclusions.

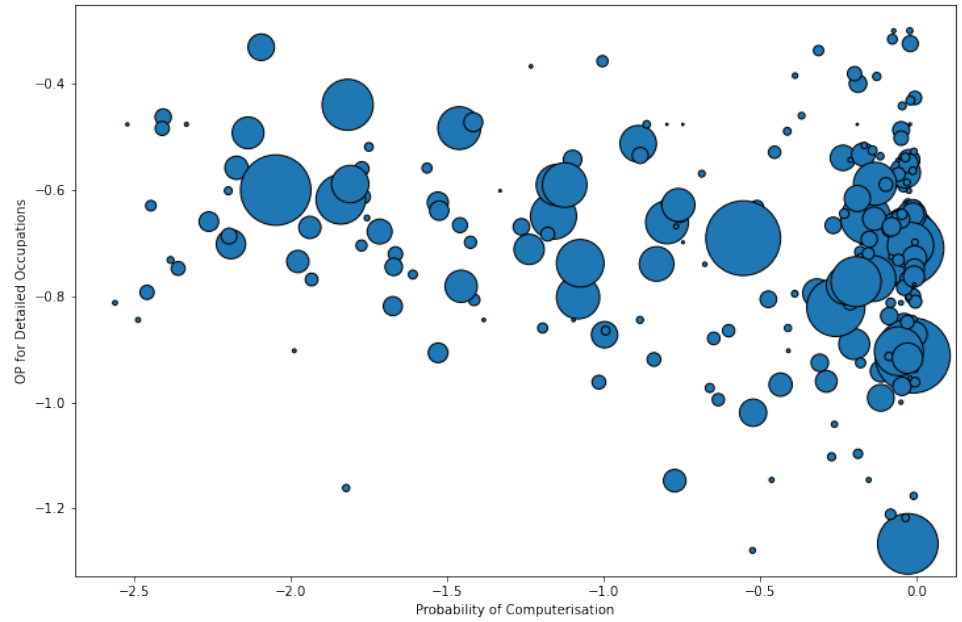
Relative change of EP/OP against PCom

In the previous section, we plotted EP/OP values for 2012 against PCom on a scatterplot. It seems like a waste to not use more of the EP/OP values for the other years given all our efforts to clean the dataset in Chapter 2. We can use the EP/OP values for 2011 and 2021 to calculate the relative change of EP/OP for the Detailed Occupations from 2011 to 2021, similar to what we did in Figure 6. Plotting the relative change values against PCom in a scatter plot gives us Figure 12. We also calculated the correlation coefficients, which can be found in Table 4. Although there seems to be no obvious trends in Figure 12, the Pearson correlation coefficient for the relative change of OP with PCom is relatively high.

Due to the fact that the relative change of EP/OP could be negative, taking the logarithmic would prove to be tricky. It is hard to justify simply removing negative values from the dataset because we would be ignoring all occupations which became 'younger' from 2011 to 2021 and they are an essential part of the dataset. Hence, we will not plot a logarithmic scale scatterplot as we had done in the previous section.



(a) Elderly Proportion

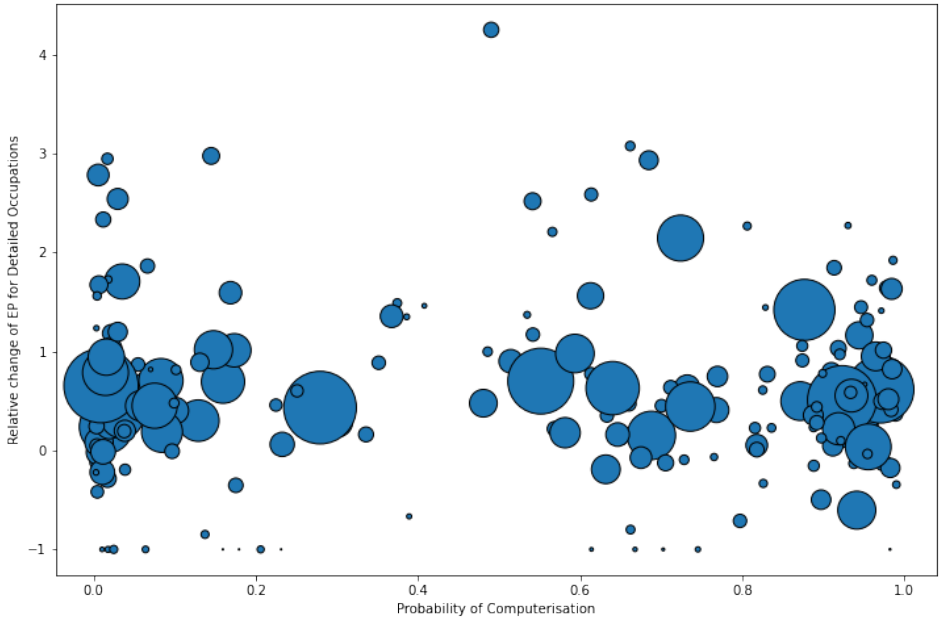


(b) Old Proportion

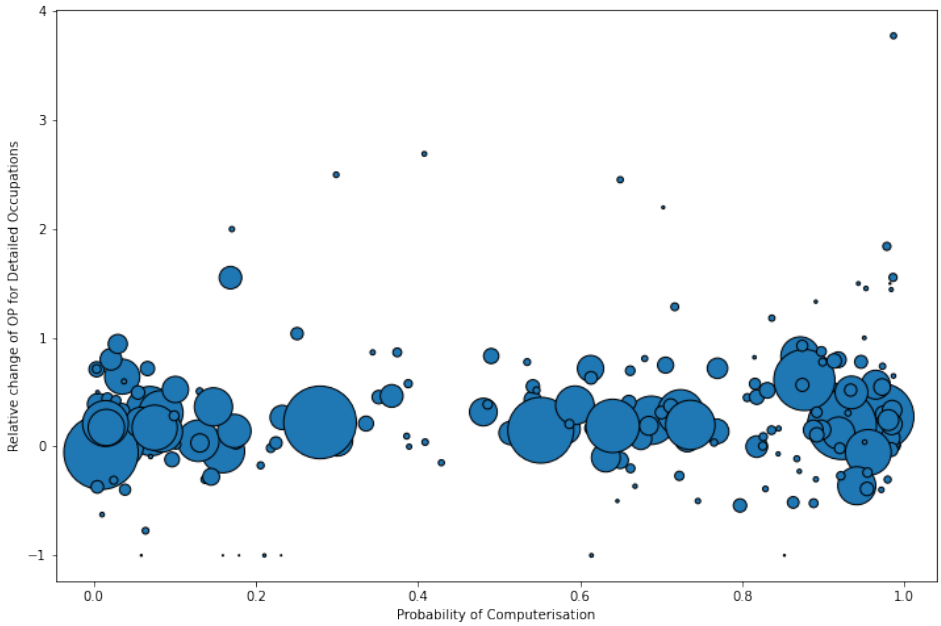
Figure 11: Plot of $\log_{10}(\text{EP})/\log_{10}(\text{OP})$ (for each Detailed Occupation) against $\log_{10}(\text{PCom})$. There seems to be a slight downward trend for the $\log_{10}(\text{OP})$ plot. This suggests that occupations that are less susceptible to computerisation tend to have a higher proportion of old people.

4.3 Conclusion

We did not find any strong trends despite using different metrics and plot scales. The most interesting results arose from the OP values, specifically the relationships of $\log_{10}(\text{OP})$ and relative



(a) Elderly Proportion



(b) Old Proportion

Figure 12: Plot of relative change of EP/OP from 2011 to 2021 (for each Detailed Occupation) against PCom. There does not appear to be any noticeable trends.

change of OP with respect to $\log_{10}$(PCom) and PCom respectively. The latter has a particularly high (relative to the others) Pearson correlation coefficient, while the former has a more obvious trend in the scatterplot. We should note that the each datapoint in the scatterplots had its size scaled relative to

Type of correlation	EP	OP
Pearson	0.0143	0.180
Spearman	0.0247	0.190
Kendall	0.0191	0.128

Table 4: Correlation coefficients for Figure 12

Variables	Weighted Pearson Correlation
Relative change of OP vs PCom	0.198
$\log_{10}$(OP) vs $\log_{10}$(PCom)	-0.434

Table 5: Weighted Pearson Correlation showing that the strongest trend we have found so far is the relationship between the $\log_{10}$(OP) for 2012 and $\log_{10}$(PCom).

the number of people employed within that particular occupation. On the other hand, our correlation coefficients were calculated without regard for this scaling. Hence, it might be worth investigating the weighted correlation coefficients. Given that both of the interesting relationships seem to be somewhat linear in nature, we shall focus on the weighted Pearson correlation. To calculate the weighted Pearson correlation between two vectors x and y with weight vector w (all three vectors must have the same dimensions), we can use the following equations (**bailey2018weighted**):

$$\begin{aligned} \text{Weighted mean} : m(x;w) &= \frac{\sum_i w_i x_i}{\sum_i w_i}. \\ \text{Weighted covariance} : cov(x,y;w) &= \frac{\sum_i w_i (x_i - m(x;w))(y_i - m(y;w))}{\sum_i w_i}. \\ \text{Weighted correlation} : corr(x,y;w) &= \frac{cov(x,y;w)}{\sqrt{cov(x,x;w)cov(y,y;w)}}. \end{aligned} \tag{1}$$

Applying these equations on our data would give us the values in Table 5. The weighted correlation for $\log_{10}$(OP) and $\log_{10}$(PCom) is much larger (in absolute terms) than the unweighted correlation, and better matches what we see in Figure 11b. It's absolute value is also larger than the weighted correlation for relative change of OP and PCom.

Hence, the strongest trend so far is the relationship between OP and PCom. In addition, it appears to be a somewhat linear trend. Hence, we will be focus on applying linear regression to this set of variables in the next section.

4.4 Linear Regression

In this section, we will be applying a number of linear regression techniques on the OP and PCom values, specifically weighted linear regression techniques given that we got the highest correlation

when weights were taken into account. That being said, all of the following techniques can be applied on the other data values; we only picked the $\log_{10}$(OP) vs $\log_{10}$(PCom) relationship because that is the one that displayed the strongest trend so far. Another caveat is that this particular trend is not that strong, and further analysis should be conducted to determine whether it is actually statistically significant. We will further elaborate on this point in the **FUTURE WORK SECTION**.

Weighted Least Squares

Conclusion

Having tried various linear regression methods, we obtain very similar results for all of them. These results could be used to predict the OP values of the occupations missing from our *joint_auto* dataset. However, we note that there is a high level of variance around our linear regression lines. Hence, any predicted values we get might not be that useful. Instead, it might be better if we could define a minimum region which contains most if not all of our known datapoints, and has a low chance of not encompassing any unknown datapoints. We would ideally also like to quantify our confidence in such a region. We shall explore this idea in the next section.

4.5 Probably Approximately Correct (PAC) Learning

Suppose we have an unknown target set T from which we obtained independent and identically distributed (i.i.d.) samples $\delta_1, ..., \delta_m$. Using the m samples, we want to construct a hypothesis set H_m that approximates T. The framework we use to learn H_m is known as PAC Learning.

Of course, we want H_m to approximate T as close as possible, such that the probability of a new sample δ from T not belonging in H_m is less than or equal to an arbitrary threshold probability ϵ, i.e. $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$. Since H_m depends on the set of i.d.d. random samples $S = \{\delta_1, ..., \delta_m\}$, it is also random. This means that $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$ is itself a random variable, allowing us to quantify a confidence for it:

$$\mathbb{P}^m\{\delta_1, ..., \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon\} \geq 1 - q(m, \epsilon), \tag{2}$$

where $1 - q(m, \epsilon)$ is a lower bound to our confidence that the probability of a new sample not belong in H_m is less than or equal to ϵ. We refer the reader to (**paclearning1**) for a more comprehensive introduction to this concept. In fact, this entire section on PAC Learning is heavily based on the work done by (**paclearning1**), (**romao2021tight**), and (**9750913**).

Furthermore, consider the following convex scenario program:

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^T x \\ \text{s.t.} \quad & g(x, \delta_i) \leq 0, \text{for all i = 1,...,m.} \end{aligned} \tag{3}$$

Suppose δ_i belongs to an uncertainty space Δ, i.e. $\delta_i \in \Delta$ for $i = 1, ..., m$, and we have obtained the

optimal solution x_m^* to the scenario program in Equation 3. If we then want to find out if a new $\delta \in \Delta$ will violate the constraint $g(x^*, \delta) \leq 0$, we can use Equation 2 to quantify the probability of such a constraint violation happening. Let $T = \Delta$, $H_m = (\delta \in \Delta : g(x_m^*, \delta) \leq 0)$, i.e. the set of samples for which x_m^* remains feasible, and we get the following:

$$\mathbb{P}^m\{\delta_1, ..., \delta_m : \mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0) \leq \epsilon\} \geq 1 - q(m, \epsilon), \tag{4}$$

where $\mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0)$ is the probability that a new sample $\delta \in \Delta$ violates the constraint $(g(x_m^*, \delta) \leq 0)$. Similar to Equation 2, the probability of such a constraint violation should ideally to less than or equal to an arbitrary value ϵ, and our confidence of that happening is at least $1 - q(m, \epsilon)$. Note that we did not specify a probability distribution for $T = \Delta$. This means that PAC Learning is a distribution-free technique, and we do not need to make any prior assumptions about the distribution of Δ.

Before further elaborating on the $q(m, \epsilon)$ function, we want to first establish a few definitions.

Definition 1 *A constraint is considered a support constraint if its removal changes the optimiser x_m^*. The support set of x_m^*, denoted by supp(x_m^*), is the collection of support constraints for x_m^*.*

Definition 2 *We say that the convex scenario program 3 is fully-supported if, for any S with $|S| = m$ and $m > n_x$, $|supp(x_m^*)| = n_x$.*

Definition 3 *We say that the convex scenario program 3 is non-degenerate if, solving the program with only the support constraints in the support set supp(x_m^*) gives us the same optimiser x_m^* as when solving the program with all constraints (constructed with all the samples in S).*

Assuming x_m^* exists and is unique, we can define $q(m, \epsilon)$ as:

$$q(m, \epsilon) = \sum_{i=0}^{n_x - 1} \binom{m}{i} \epsilon^i (1 - \epsilon)^{m-i}. \tag{5}$$

Note that this holds with equality for fully-supported programs.

Now that we have established the basics of PAC Learning, we can return to our original problem of defining a minimum region for our datapoints. We shall denote the $i^{th}\log_{10}(PCom)$ and $i^{th}\log_{10}(OP)$ values as u_i and y_i respectively. Each set of our datapoints (u_i, y_i) can be considered a sample δ_i for $i = 1, ..., m$. Hence, out entire dataset can be considered as the set S. We are trying to define a region H_m that approximates our unknown space $T = \Delta$ (unknown since we do not have a complete set of data nor do we have the data for any new occupations that may exist in the future). Since the data exhibits a fairly linear relationship, it makes sense to have a minimum vertical width strip as the

region. To define such a region, we need variables x_2, x_3 to encode the median line (as the gradient and y-intercept respectively), and a variable x_1 to denote the semi-width length. To ensure all of our datapoints are contained within this region H_m, we have to set up constraints for all m datapoints. Finally, we minimise the semi-width length x_1 to give a minimum region. Putting everything together gives us the following optimisation problem.

$$\begin{aligned} \min_{x_1 \in \mathbb{R}, x_2 \in \mathbb{R}, x_3 \in \mathbb{R}} \quad & x_1 \\ \text{s.t. } y_i - x_2 u_i - x_3 \quad & \leq x_1, \text{for all i} = 1,...,\text{m} \\ y_i - x_2 u_i - x_3 \quad & \geq -x_1, \text{for all i} = 1,...,\text{m}. \end{aligned} \tag{6}$$

On first glance, this might seem different from the convex scenario program 3. However, by placing x_1, x_2, x_3 into a vector $x \in \mathbb{R}^3$, and rearranging all the constraints into a matrix inequality constraint, we can see that both programs are of the same form. Additionally, $\mathbb{R}^3$ is a convex set, and our objective function and constraints are all convex (since they are just linear). Hence, this problem is actually a convex scenario program, making it equivalent to scenario program 3. Solving this program and plotting our minimum vertical width strip H_m gives Figure 13.

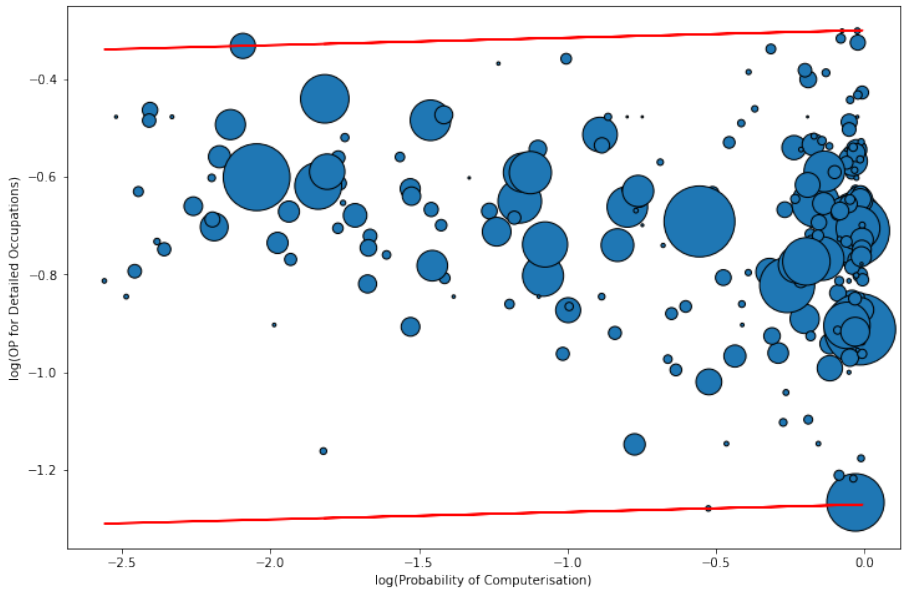


Figure 13: Plot of $\log_{10}(\text{EP})/\log_{10}(\text{OP})$ (for each Detailed Occupation) against $\log_{10}(\text{PCom})$ with PAC Learning applied. The red lines represents the boundaries of H_m. We can see that H_m contains all of our datapoints while ensuring that the vertical width is minimised.

While H_m contains all of our datapoints while keeping x_1 minimised, it is not very useful since it ignores the underlying downward trend. In fact, H_m would suggest a slight upward trend. This is because we have included the outlier datapoints (relative to the downward trend) in our set S. Hence, it would be useful to be able to apply PAC Learning to a dataset with discarded samples. Fortunately,

9750913 details such an approach.