University of Oxford

Engineering Science

4YP Report

The Future of Work

Author

Nian Yang Terence Tan

Supervisor

Professor Michael A. Osborne

May 16, 2023

Abstract

Automation of jobs and ageing population are two key social challenges that the world will face over the next few decades. This report aims to examine the relationship between the age distribution within occupations and their automatability, and find out if automation can make up for the shrinking labour force due to an ageing population. Previous studies looked at broad categories of employment. Others examined the effects of ageing population and automation on the labour market. In this report, we directly examine the relationship between the age distribution within occupations in the United States and their automatability on a more detailed scale. The data corresponding to the age distribution within occupations was obtained from the US Bureau of Labor Statistics[1] (BLS), and spans the period 2011 to 2021. This data was standardised according to the 2018 Standard Occupational Classification (SOC) System. The data corresponding to the automatability of the occupations was obtained from Frey and Osborne, 2017; specifically, we used the Probability of Computerisation as a measure of automatability. We explored trends within the age distribution dataset before merging it with the automatability dataset. We found no correlations between the rate of ageing in occupations and their Probability of Computerisation. However, we did find a possible relationship between the proportion of people aged 55 years and older in occupations and the Probability of Computerisation. Assuming this trend is statistically significant, it suggests that occupations at higher risk of automation tend to have a lower proportion of workers who are 55 years and older. We applied a number of linear regression techniques on the joint dataset to further investigate this trend. We then used Probably Approximately Correct Learning to perform statistical analysis on the joint dataset, which acts as a proof of concept of using such a technique in combination with a scenario discarding scheme to deal with outliers in datasets. Based on our findings, we can offer no guarantees that automation can make up for the shrinking labour force due to an ageing population, and recommend governments adopt a more active approach in managing the effects of automation and an ageing population.

[1] https://www.bls.gov

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Professor Michael Osborne for supervising me throughout this project. His insights and guidance have been crucial to the successful completion of this project. This project would also not have been possible without the automatability dataset provided by him.

I would also like to extend my heartfelt thanks to Professor Kostas Margellos for sharing his knowledge on Probably Approximately Correct Learning and providing resources for further reading. His assistance has greatly enriched my understanding of the topic.

Additionally, I would like to extend my appreciation to my fellow colleagues at Oxford who reviewed my drafts and provided invaluable feedback that improved the quality of my work.

I would also like to acknowledge the US Bureau of Labor Statistics for their meticulous work in collecting and cataloguing the labour data that this project is based on.

Lastly, I am grateful to everyone who has supported and helped me in any respect throughout this journey. This project would not have been possible without your encouragement and assistance.

Contents

1 Introduction **4**

1.1 Ageing Population 4

1.2 Project Overview 6

2 Dataset **9**

2.1 BLS Dataset 9

2.2 Automatability Dataset 9

3 Preliminary Findings **11**

3.1 General Trends 11

4 Data Analysis **18**

4.1 Checking bias 18

4.2 Exploration of Data 21

4.3 Conclusion of data analysis 23

4.4 Linear Regression 26

4.5 Probably Approximately Correct (PAC) Learning 29

5 Conclusion **39**

References **41**

1 Introduction

Technological advancement is widely believed to be the primary driving force behind economic growth (Dosi et al., 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 (Woirol and Backhouse, 1997), 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” (Hollander, 2019). 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 (Keynes, 2010). However, there are those who are cautiously optimistic of the impact of technological advances on labour. Walter Reuther, a prominent member of the United Automobile Workers union, and his colleagues 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 (Steigerwald, 2010). 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, 2021b). 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 (Rose, 2021a; Krugman, 2019).

At the same time, ageing population is a social issue that has become increasingly relevant all over the world (*World population ageing, 1950-2050.* 2002). It is expected to lead to a reduction in the labour force of countries, and an increase in social spending (Marešová, Mohelská, and Kuča, 2015). Given the loss of jobs due to automation and the shrinking of the labour force in many countries, a question naturally arises: can we expect automation to make up for the lack of workers in most jobs?

1.1 Ageing Population

The world population is projected to age significantly over the next few decades (Gerland et al., 2014). The ratio of the elderly population to the working age population is increasing and will continue to do so (World Health Organization, 2022). This trend is known as an ageing population, and is projected to 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 over 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. In the following sections, we shall first examine the ageing trends in the US as well as other countries, and look into related works on ageing population.

Cross-country comparison

The United States Census Bureau provides a comprehensive report regarding ageing trends in the US in Vespa, Armstrong, Medina, et al., 2018. From 2030 onwards, the baby boomer generation will all be over the age of 65. Starting that year, population growth of the country will also be primarily driven by immigration rather than natural increase (number of births minus number of deaths). The number of people aged 65 and older in the US is expected to increase from 49 million in 2016 to 95 million in 2060. This represents an increase in the share of people aged 65 and older from 15% in 2016 to nearly 25% of the total US population in 2060. By 2030, 20% of the US population will be 65 years and older. The number of people aged 85 and older is expected to increase from 6.4 million in 2016 to 19 million in 2060. This represents an increase in the share of people aged 85 and older from about 2% in 2016 to about 4.7% of the total US population in 2060.

China is currently home to the largest population of older people in the world (Lancet, 2022), and is expected to age rapidly over the next few decades (Beardson and Fielding, 2021). The number of people aged 60 and older will increase from 254 million people in 2019 to 402 million people in 2040; this would be about 28% of the population in China ("Ageing and health - China", n.d.). "Ageing and health - China", n.d. also showed that around 75% of people aged 60 and older in China suffer from noncommunicable diseases such as diabetes, which will likely put more pressure on the country's healthcare system in the coming decades. Indeed, Luo, Su, and Zheng, 2021 recommended pushing for more flexible and innovative population health strategies. That being said, the same study also provided some evidence of declining disease burden as well as healthy ageing.

Japan currently has the highest proportion of elderly citizens in the world (Kyodo, 2019). In fact, 28% of the population is 65 and older. By 2036, this age group will represent about 33% of the population in Japan (D'Ambrogio, 2020). D'Ambrogio, 2020 also highlights the various socioeconomic impacts of an ageing population in Japan, including a worsening economy and higher expenditure on healthcare. Parsons and Gilmour, 2018 provides evidence that fertility- and migration-based policies are unlikely to reverse the ageing trend in Japan significantly. Therefore, the paper recommends improving work productivity of its population and increasing participation rate of senior citizens in the labour force.

The above three countries currently have the three largest economies in the world (n.d.). The effects of an ageing population in these countries on their economies will be felt by the world. As it stands, Japan is currently the most 'aged' society among the three, with the US being the 'youngest'. As the country most ahead of the curve in terms of an ageing society, Japan will likely be closely observed by other countries as a case study of what will transpire in the coming decades.

Related Works

Research into the social implications of an ageing population has been carried out as far back as 2002. Tinker, 2002 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).

Cheng et al., 2020 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. Indeed, there is evidence to suggest that the healthcare system in the US is not prepared to meet the increasing demands of the ageing population (Foley and Luz, 2020).

Marešová, Mohelská, and Kuča, 2015 examined the development of healthcare expenditure in the context of an ageing population. This paper pointed out the consequences of an ageing population such as a shrinking labour force, higher incidence of chronic diseases, and increased government expenditure to support the elderly. This aligns with what was discussed earlier in the two papers (Tinker, 2002; Cheng et al., 2020).

1.2 Project Overview

This project aims to examine the relationship between the age distribution within occupations and the Probability of Computerisation[2] (calculated by Frey and Osborne, 2013) of those occupations, and apply statistical techniques on any significant trends identified. All the code and relevant supplementary materials can be found in a GitHub repository[3]. This project is motivated by the question of whether automation can solve the problem of ageing population and vice versa. We will go into further detail about the data and where we got them from in Chapter 2. Although similar work has been done on this topic (Basu et al., 2018), the study only looked at broad categories of employment. In contrast, we zoomed in to look at specific occupations in this project. Other papers (Acemoglu and Restrepo, 2017; Phiromswad, Srivannaboon, and Sarajoti, 2022) examined the effects of ageing population and automation on the labour market. In particular, Phiromswad, Srivannaboon, and

[2] Defined as 'job automation by means of computer-controlled equipment' by Frey and Osborne, 2017.

[3] https://github.com/terencetan-c/4YP-The-Future-of-Work

Sarajoti, 2022 went one step further and took into account the interaction effects of ageing population and automation as well. To the best of our knowledge, our study is the first to directly examine the relationship between automation and the age distribution within occupations at such a detailed scale. Additionally, we used a number of linear regression models in Chapter 4.4, and applied a Probably Approximately Correct Learning model with a scenario discarding scheme on a set of data in Chapter 4.5.

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 (*The Oxford Dictionary of Sports Science & Medicine*, 2006), and that the ageing process varies for different people depending on a variety of factors (Levine, 2013; Hayflick, 2007). In fact, there are studies looking into moving beyond using chronological age to define an elderly person (Kotter-Grühn, Kornadt, and Stephan, 2015; Soto-Perez-de-Celis et al., 2018; Klemera and Doubal, 2006). However, there are numerous issues surrounding this approach (Jylhävä, Pedersen, and Hägg, 2017), such that validation of such results are often difficult (Cho, Park, and Lim, 2010), 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' (Organization et al., 2010; Orimo et al., 2006; "Elderly Population", n.d.). 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 R_{65+}. However, we also recognise that this definition of 'elderly' coincides roughly with the retirement age of the US (Munnell, 2013). 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 R_{55+}. Having said that, R_{65+} would still prove to be an interesting metric to investigate. For example, if there is a general trend of increasing R_{65+} over the years, that would be a sign of an ageing labour force in spite of the effects of retirement. Therefore, both of these metrics are examined in this report.

Related Works

Frey and Osborne, 2013 and Frey and Osborne, 2017 used data from the Occupational Information Network[4] (O*NET) as well as hand-labelled data by experts to calculate the Probability of Computerisation for 702 detailed occupations using a Gaussian process classifier. It was predicted that 47% of employment in the US was at risk of computerisation. The papers also provided evidence that

[4] https://www.onetonline.org

wages and educational attainment have a negative relationship with an occupation's Probability of Computerisation. Pajarinen, Rouvinen, Ekeland, et al., 2015 used the same methodology to show that about 33% of employment in Finland and Norway was at risk of computerisation. Fuei, 2017 did the same for Singapore, and found that 25% of employment in the city-state was at high risk of computerisation. Similarly, Brzeski and Burk, 2015 extended the analysis to Germany and concluded that 59% of employment was at risk of computerisation.

Acemoglu and Restrepo, 2017 showed that there is no inverse relationship between ageing population and economic growth, i.e. an ageing population does not negatively affect the economy. The paper argues that countries that are ageing more rapidly tend to embrace and faster implement automation technologies.

Basu et al., 2018 mapped data from Frey and Osborne, 2017 to occupations defined under the United Nations' International Standard Classification of Occupations[5] (ISCO), in order to estimate the risk of automation to older workers across 15 countries. The report found that older workers were at medium to high risk of being displaced by automation in several countries. Countries with higher rates of ageing also face the highest risk of automation. Ultimately, the paper encouraged discussions around the challenges faced by older workers, in order to enable them to remain in the labour force in light of the decreasing proportion of younger workers.

Phiromswad, Srivannaboon, and Sarajoti, 2022 used the Probability of Computerisation data calculated by Frey and Osborne, 2017 and data about age-related abilities from O*NET to examine the interaction effects of computerisation and ageing population on the US labour market in terms of employment growth and earnings growth. The latter two are obtained from the US Bureau of Labor Statistics (BLS), giving a complete dataset consisting of 501 occupations. The paper found that both computerisation and ageing population have significant effects on employment growth but not earnings growth. Specifically, the higher the Probability of Computerisation, the lower the employment growth. This is also supported by Graetz and Michaels, 2018 and Acemoglu and Restrepo, 2020. The paper also found that the strength of the relationship between the Probability of Computerisation and employment growth is dependent on the age-related abilities required for the particular occupation.

[5] https://www.ilo.org/public/english/bureau/stat/isco/

2 Dataset

We used two main metrics for this project: the Probability of Computerisation of occupations, and the age distribution within occupations. The dataset for the former is provided in an earlier work by Frey and Osborne, 2017. The latter can be found in datasets provided by the US Bureau of Labour Statistics; there is one dataset for each year from 2011 to 2021. All the datasets mentioned above use the Standard Occupational Classification (SOC) system to classify the occupations, which means that we can map from one dataset to another using the SOC codes[6]. 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 give a brief overview of the SOC system and examine the datasets.

Standard Occupational Classification (SOC) system

The SOC system categorises occupations using a hierarchy of aggregation. This hierarchical structure consists of four categories: Major Group, Minor Group, Broad Group, and Detailed Occupation. Each category represents an increasingly more granular level of classification. At the highest level of aggregation, there is the Major Group, which is followed by the Minor Group, then the Broad Group, and finally, at the lowest level, the Detailed Occupation. For example, 'Chief Executives' (SOC code: 11-1011) is a Detailed Occupation, which belongs to the Broad Group that is also called 'Chief Executives' (SOC code: 11-1010). This Broad Group occupation belongs to the Minor Group called 'Top Executives' (SOC code: 11-1000), which in turn belongs to the Major Group called 'Management Occupations' (SOC code: 11-0000). Different Detailed Occupations may belong to the same Broad Group, Minor Group, and/or Major Group under this system of classification.

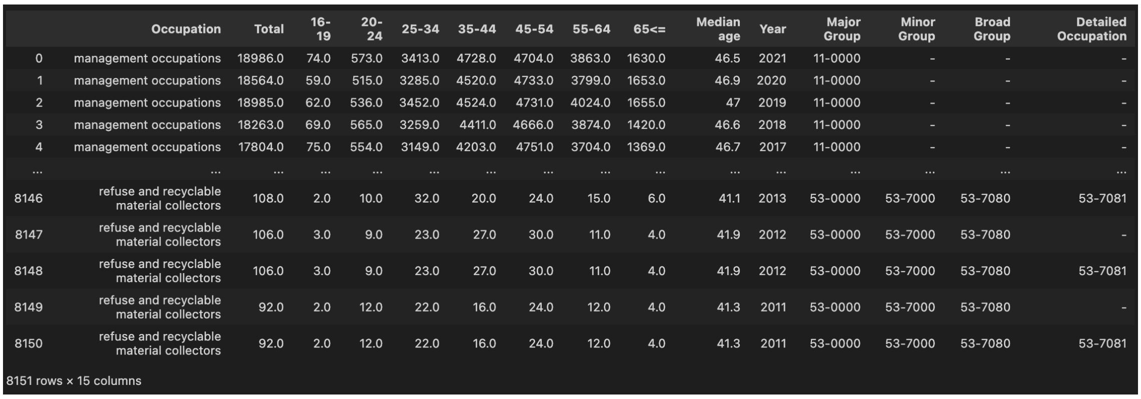
2.1 BLS Dataset

As mentioned earlier, 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. After standardising and processing the datasets according to the updated SOC, we obtain the dataset shown in Figure 1; we shall refer to this dataset as the BLS dataset for the rest of this report. Note that the values under the *Total* and age group columns are given 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 report. Further details on the standardisation process can be found in the Appendix.

2.2 Automatability Dataset

This dataset (which we shall refer to as Automatability dataset) was obtained from Frey and Osborne, 2017, and features 702 Detailed Occupations as shown in Figure 2. For each of these

[6] https://www.bls.gov/soc/2018/soc_structure_2018.pdf

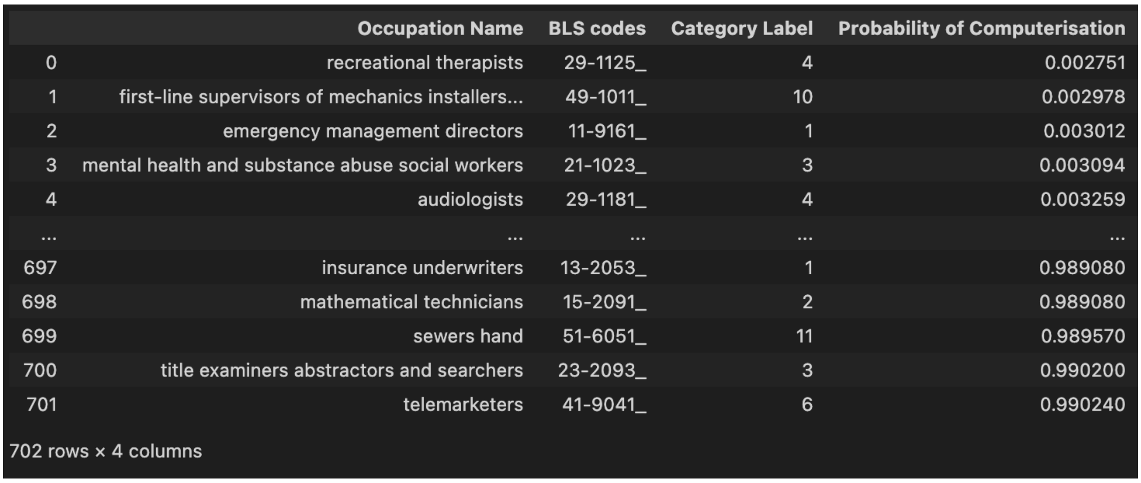


	Occupation	Total	16- 19	20- 24	25-34	35-44	45-54	55-64	65<=	Median age	Year	Major Group	Minor Group	Broad Group	Detailed Occupation
0	management occupations	18986.0	74.0	573.0	3413.0	4728.0	4704.0	3863.0	1630.0	46.5	2021	11-0000	-	-	-
1	management occupations	18564.0	59.0	515.0	3285.0	4520.0	4733.0	3799.0	1653.0	46.9	2020	11-0000	-	-	-
2	management occupations	18985.0	62.0	536.0	3452.0	4524.0	4731.0	4024.0	1655.0	47	2019	11-0000	-	-	-
3	management occupations	18263.0	69.0	565.0	3259.0	4411.0	4666.0	3874.0	1420.0	46.6	2018	11-0000	-	-	-
4	management occupations	17804.0	75.0	554.0	3149.0	4203.0	4751.0	3704.0	1369.0	46.7	2017	11-0000	-	-	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8146	refuse and recyclable material collectors	108.0	2.0	10.0	32.0	20.0	24.0	15.0	6.0	41.1	2013	53-0000	53-7000	53-7080	53-7081
8147	refuse and recyclable material collectors	106.0	3.0	9.0	23.0	27.0	30.0	11.0	4.0	41.9	2012	53-0000	53-7000	53-7080	-
8148	refuse and recyclable material collectors	106.0	3.0	9.0	23.0	27.0	30.0	11.0	4.0	41.9	2012	53-0000	53-7000	53-7080	53-7081
8149	refuse and recyclable material collectors	92.0	2.0	12.0	22.0	16.0	24.0	12.0	4.0	41.3	2011	53-0000	53-7000	53-7080	-
8150	refuse and recyclable material collectors	92.0	2.0	12.0	22.0	16.0	24.0	12.0	4.0	41.3	2011	53-0000	53-7000	53-7080	53-7081

8151 rows × 15 columns

Figure 1: BLS dataset after standardisation process. There are 11 of each occupation for each of the years from 2011 to 2021. The values under the *Total* and age group columns are given in thousands. The numbers in the last four columns are the corresponding occupations' SOC codes.

occupations, a Probability of Computerisation had been calculated. We shall refer to this probability as P_{Com}. Other variables are included as well, such as the skills associated with each occupation and a Category Label. These were used to calculate the P_{Com}, but we will just focus on the P_{Com} in this report.



	Occupation Name	BLS codes	Category Label	Probability of Computerisation
0	recreational therapists	29-1125_	4	0.002751
1	first-line supervisors of mechanics installers...	49-1011_	10	0.002978
2	emergency management directors	11-9161_	1	0.003012
3	mental health and substance abuse social workers	21-1023_	3	0.003094
4	audiologists	29-1181_	4	0.003259
...	...	...	...	...
697	insurance underwriters	13-2053_	1	0.989080
698	mathematical technicians	15-2091_	2	0.989080
699	sewers hand	51-6051_	11	0.989570
700	title examiners abstractors and searchers	23-2093_	3	0.990200
701	telemarketers	41-9041_	6	0.990240

702 rows × 4 columns

Figure 2: Automatability dataset with only the first four columns displayed. This dataset was obtained from Frey and Osborne, 2017.

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 3. Clearly, *Major Group* occupations are most representative of the US civilian labour force, with *Minor Group* as the least. Looking through the BLS dataset, this makes sense since the BLS tended to neglect information about the Minor Group. Another thing to note is that many *Broad Group* occupations only contain a single *Detailed Occupation* that also shares the same occupation name (for example, 'Chief Executives': 11-1010 and 11-1011, as shown earlier in Chapter 2). Hence, it is not surprising that the ratios for *Broad Group* and *Detailed Occupation* are 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 3 that there is a considerable dip in the ratios in 2020 relative to other years, coinciding with the onset of the COVID-19 pandemic (Kozicki and Gornikiewicz, 2020; Falk et al., 2021). Plotting the ratios relative to the employed portion of the US civilian labour force accounts for this; as seen in Figure 4, 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 R_{65+} (refer to Chapter 1.2 for definition) over the Major Groups for each year, and plot the values against the years to obtain Figure 5a. We do the same for R_{55+}, and get the plot in Figure 5b. We see a steady increase over the years for both plots, which is expected given the ageing population of the US as discussed in Chapter 1.1. We can examine these plots in further detail by plotting the R_{65+}/ R_{55+} for each of the 21 Major Groups against the years to obtain Figure 6; 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 6 given that there are 21 individual

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

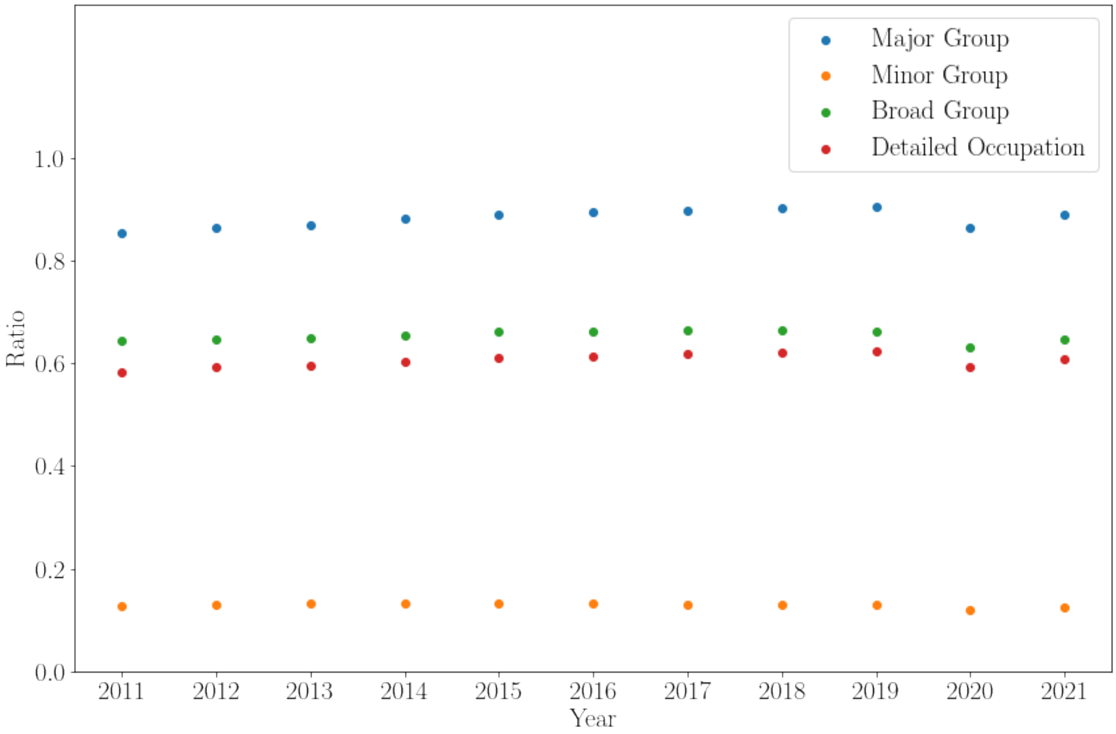


Figure 3: Ratio of the various SOC categories to the US civilian labour force. This figure represents the proportion of the civilian labour force captured within Major Group/Minor Group/Broad Group/Detailed Occupation. It can be seen that there is little change in the ratios over the years. Additionally, the total employment captured within Major Group is the largest.

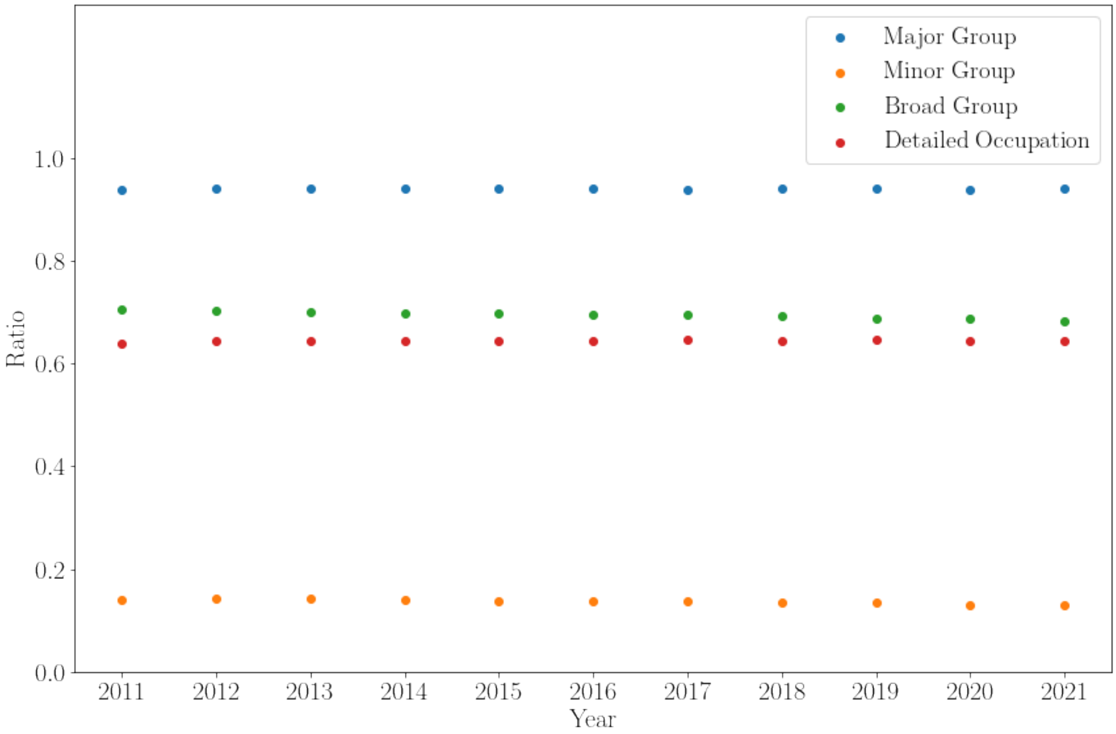
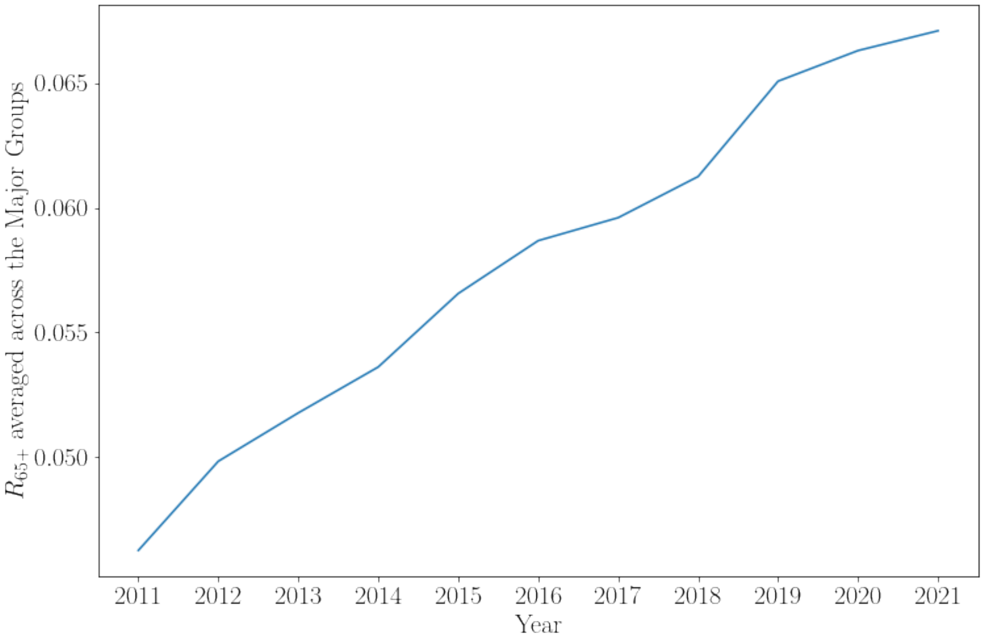
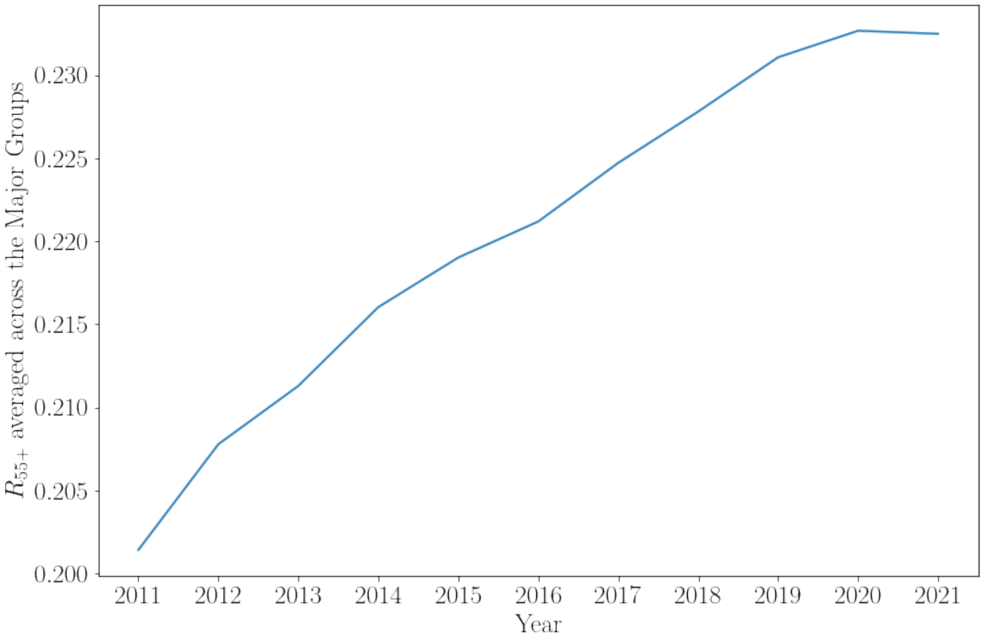


Figure 4: Ratio of the various SOC categories to the employed portion of the US civilian labour force. This figure represents the proportion of the employed civilian labour force captured within Major Group/Minor Group/Broad Group/Detailed Occupation. The plot is similar to Figure 3.



(a) R_{65+}

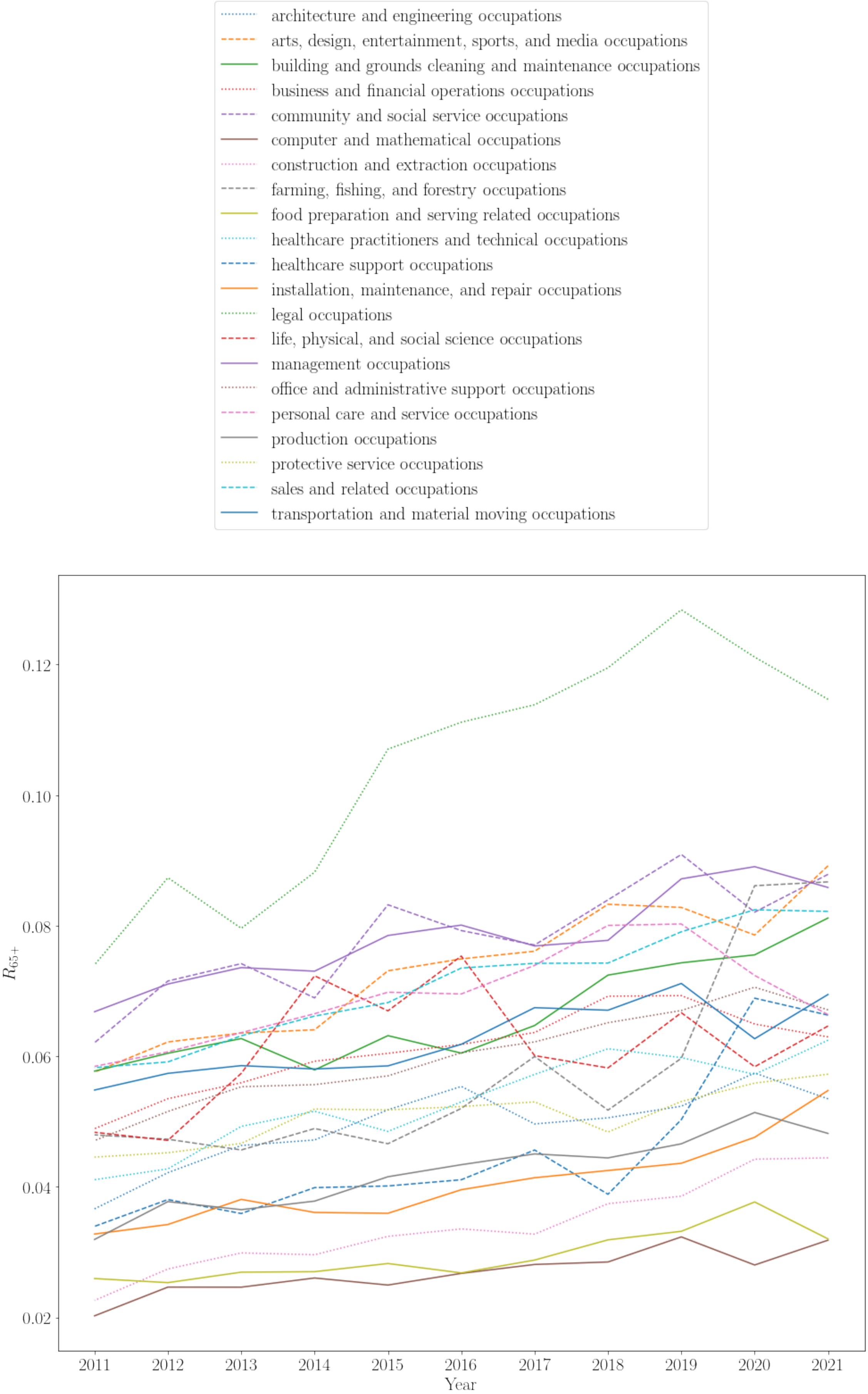


(b) R_{55+}

Figure 5: Plot of R_{65+}/ R_{55+} (averaged over the Major Groups) against Year. There is an increase in the mean values over the years.

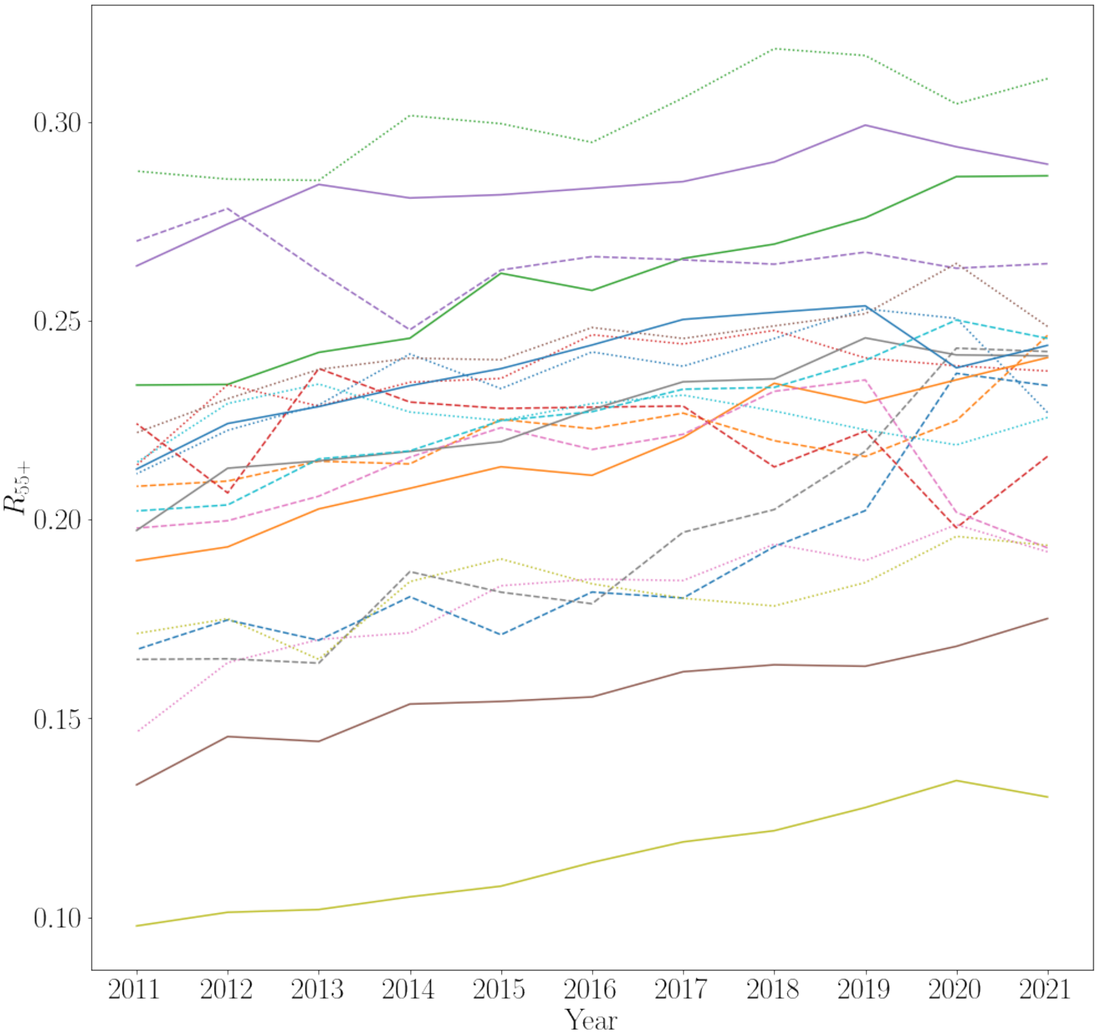
plots within each of the two figures. Hence, we can calculate the relative change in R_{65+}/ R_{55+} for each Major Group occupation from 2011 to 2021 to obtain Figure 7. With regards to R_{65+}, the 'construction and extraction occupations' had the highest relative increase while the 'personal care and service occupations' had the least. When we look at R_{55+}, 'farming, fishing, and forestry occupations' had the highest relative increase, while there are some occupations that had a relative decrease. The 'life, physical, and social science occupations' had the largest relative decrease. However, this relative decrease is still merely about 4%, so it is accurate to say that there is a general trend of ageing across the occupations.

These findings are not surprising given what was discussed in Chapter 1.1. We expect that an ageing population would lead to an ageing labour force. Naturally, these occupations will age even more over the next few decades as the trend of ageing population in the US continues.



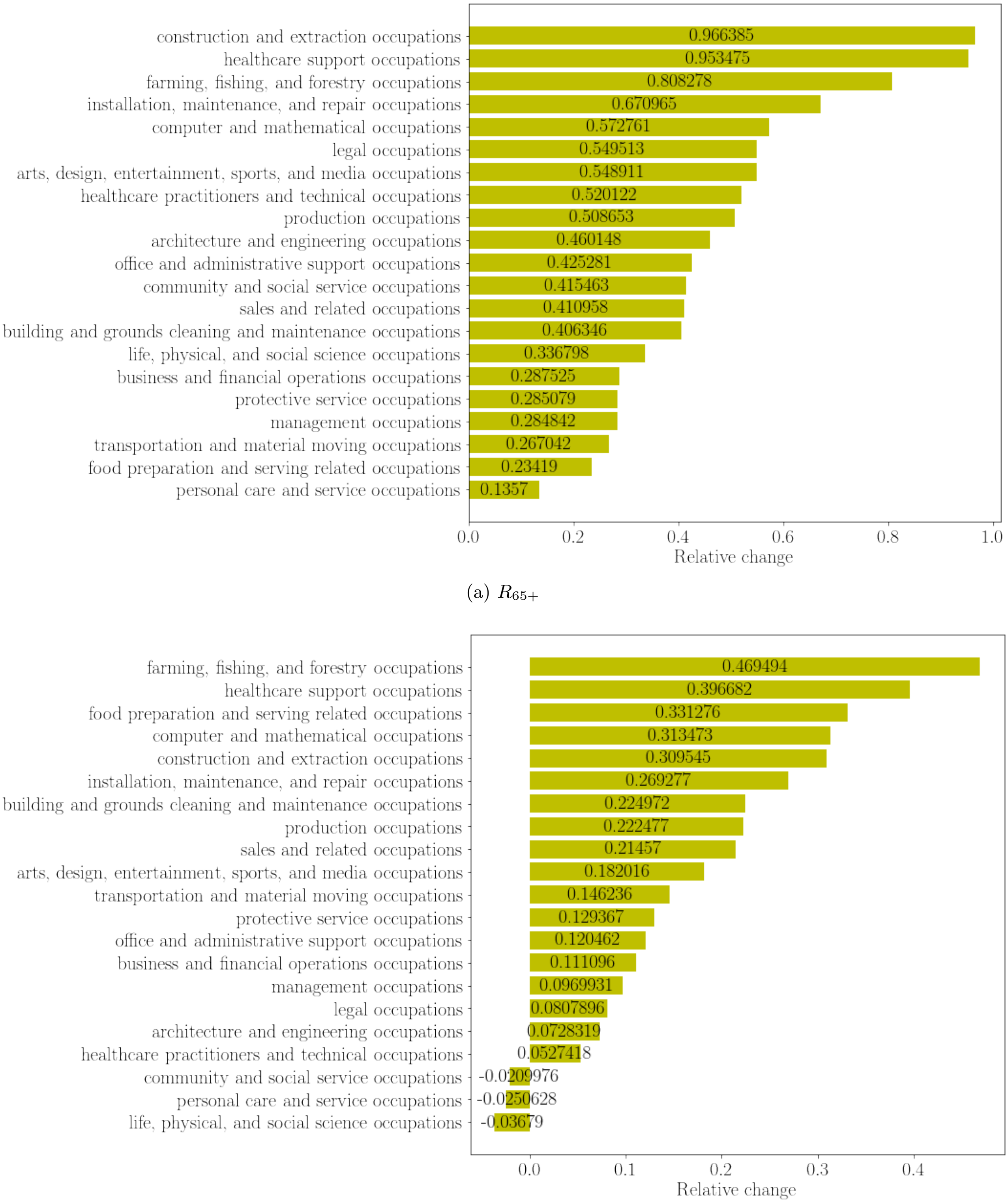
(a) R_{65+}

Figure 6



(b) R_{55+}

Figure 6: Plot of R_{65+}/ R_{55+} (for each Major Group) 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) R_{65+}

(b) R_{55+}

Figure 7: Relative change of R_{65+}/ R_{55+} 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.

4 Data Analysis

Now that we have our processed BLS dataset (with the calculated R_{55+} and R_{65+}), we can join it with the automatability dataset (which includes the Probability of Computerisation) from Frey and Osborne, 2017 based on the Detailed Occupation. This gives us a joint data 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 240 Detailed Occupations in the *joint_auto* dataset (recall that the automatability dataset contains 702 Detailed Occupations as mentioned in Chapter 2.2). As can be seen in Figure 8, *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.

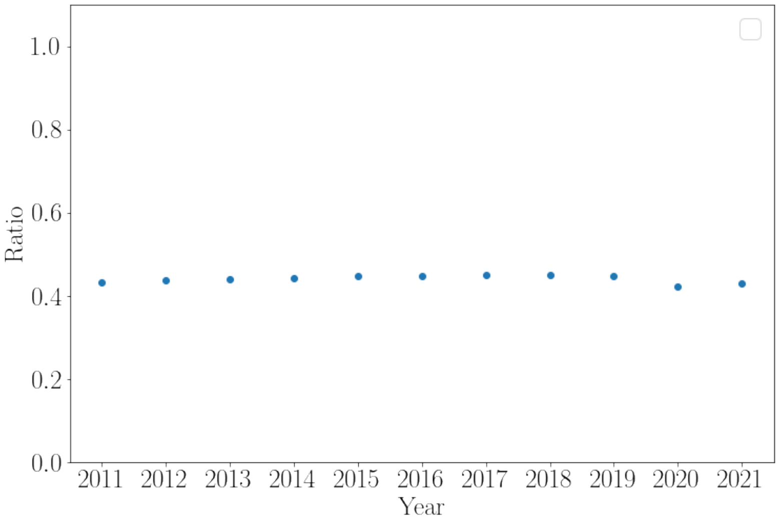


Figure 8: Plot of ratio of *joint_auto* to US civilian labour force against Year. We can see that our *joint_auto* dataset represents a significant proportion of the total labour force, but not enough for us to generalise any trends we may find in our dataset. We will have to investigate how well our dataset represents the actual US civilian labour force.

4.1 Checking bias

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 calculate the means and variances of the R_{55+}, R_{65+}, and P_{Com} values for both the sample and population, which are shown in Table 1. Note that we do not make any distinctions between the biased and unbiased estimator of the variances when we refer to the sample variance because the differences between both calculated values

	Population		**Sample**	
Variable	Mean	Variance	Mean	Variance
R_{65+}	0.0578	0.000394	0.0516	0.00334
R_{55+}	0.220	0.00190	0.223	0.0122
P_{Com}	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 R_{65+}/R_{55+} values for the sample have higher variances compared to that of the population.

are negligible. We can see that the means and variances match quite well for the P_{Com} variable. For R_{55+} and R_{65+}, the means match quite closely, but the variances are off by an order of magnitude; the sample has a higher variance than the population for both R_{55+} and R_{65+}. This means that the sample has more spread-out values for R_{55+} and R_{65+}, so we must exercise caution when generalising any findings from the sample to the population. Otherwise, the sample seems fairly representative of the population. Since the mean and variance analysis is too simplistic to give us any concrete conclusions, 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 9. 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 10. 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 least-squares line to the plot confirms this as well, with the least-squares 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 weighting of each of the Major Groups.

Given our above bias 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.

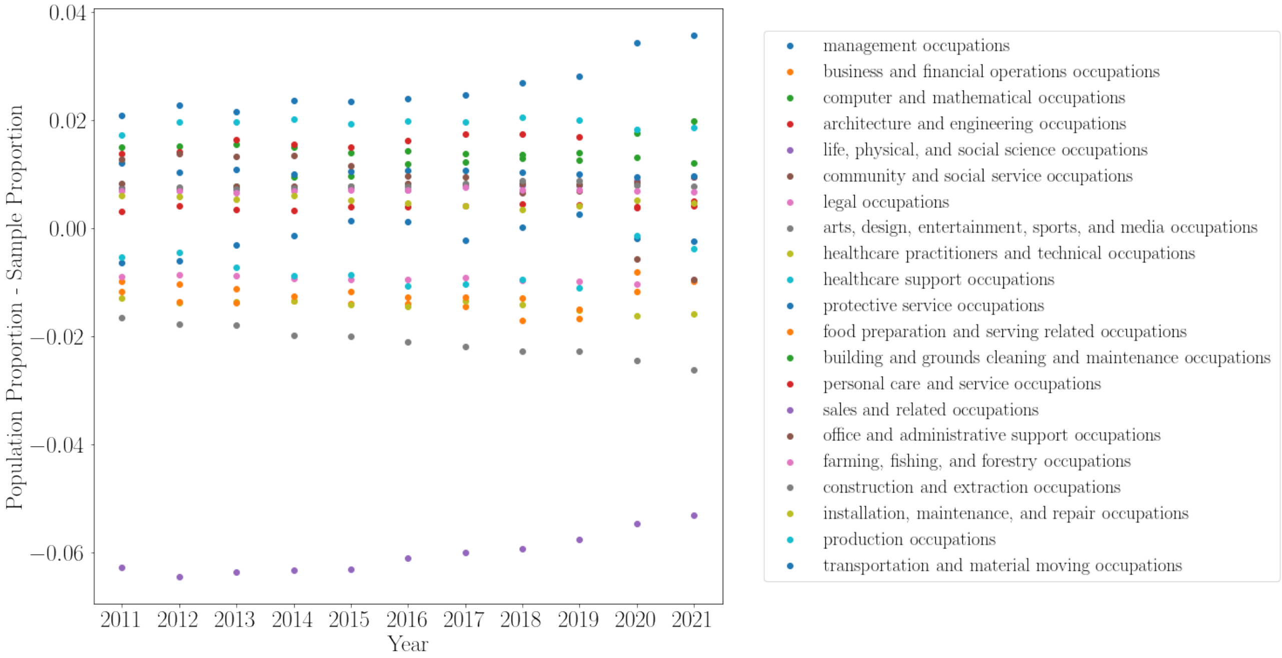


Figure 9: Plot of the difference between the population proportion and the sample proportion against Year. The different colours represent different Major Groups. The low values along the y-axis shows that the relative weighting of each of the Major Groups within the sample is similar to that of the population. Hence, the sample has a similar makeup to the population in terms of the Major Groups.

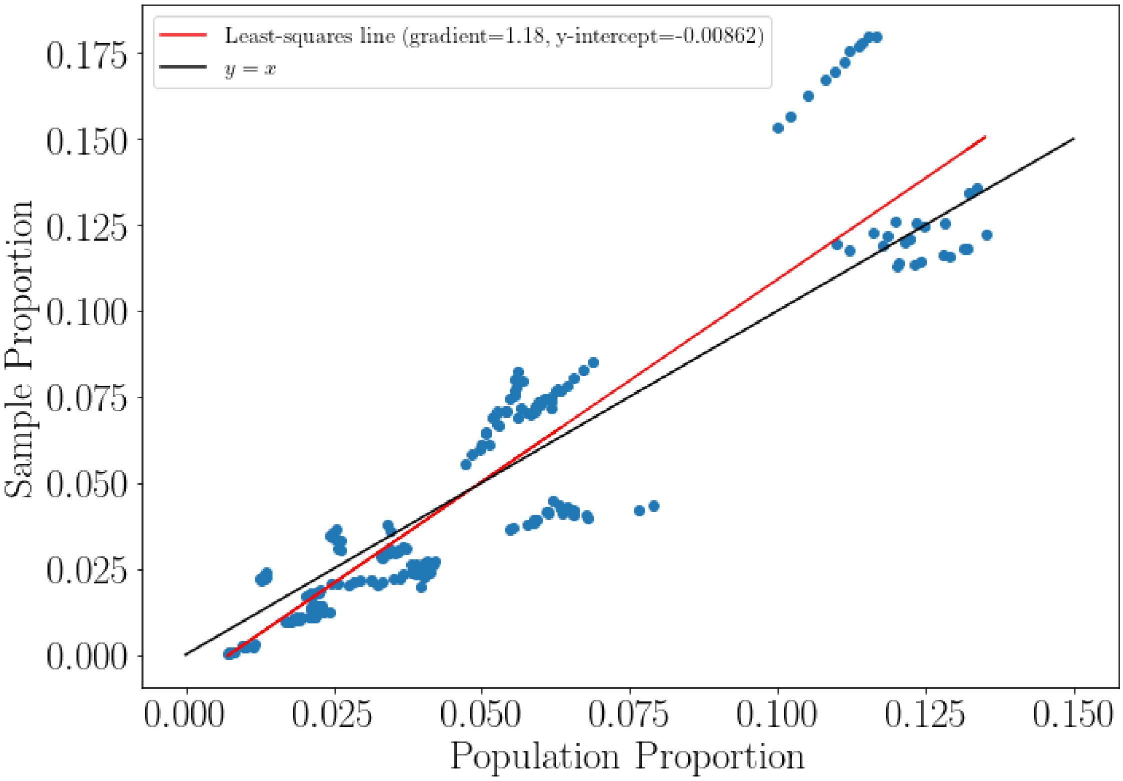


Figure 10: Plot of the sample proportion against population proportion (with a least-squares line). We can see that our plot roughly follows a straight line with a gradient of 1 and y-intercept of 0.

Type of correlation	R_{65+}	R_{55+}
Pearson	-0.00363	-0.0321
Spearman	-0.0151	-0.0661
Kendall	-0.00881	-0.0429

Table 2: Correlation coefficients for Figure 11.

4.2 Exploration of Data

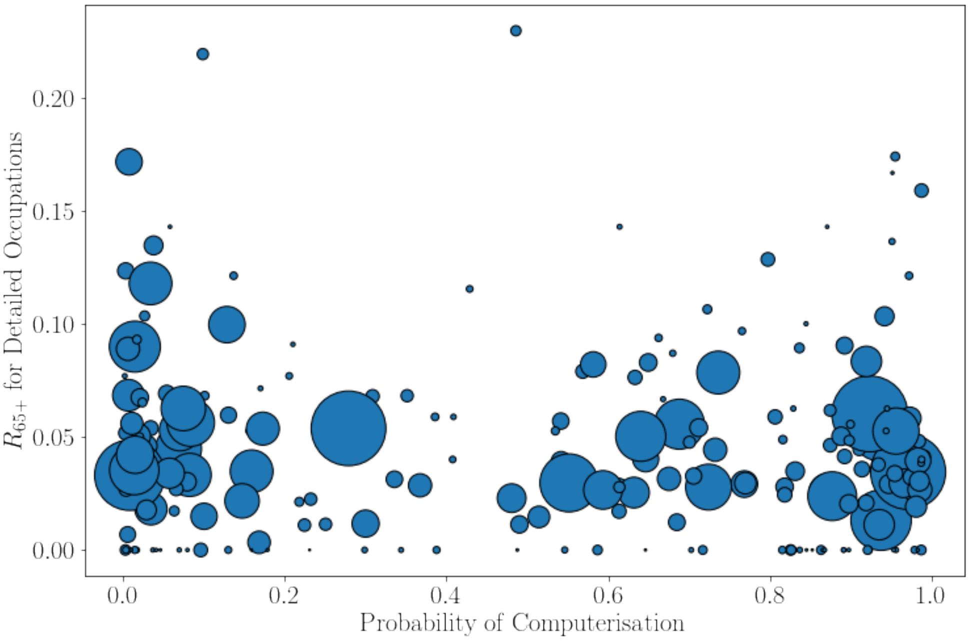
Now that we have established the representativeness of the sample, we can go about exploring the dataset.

R_{65+}/R_{55+} against $P_{\mathbf{Com}}$

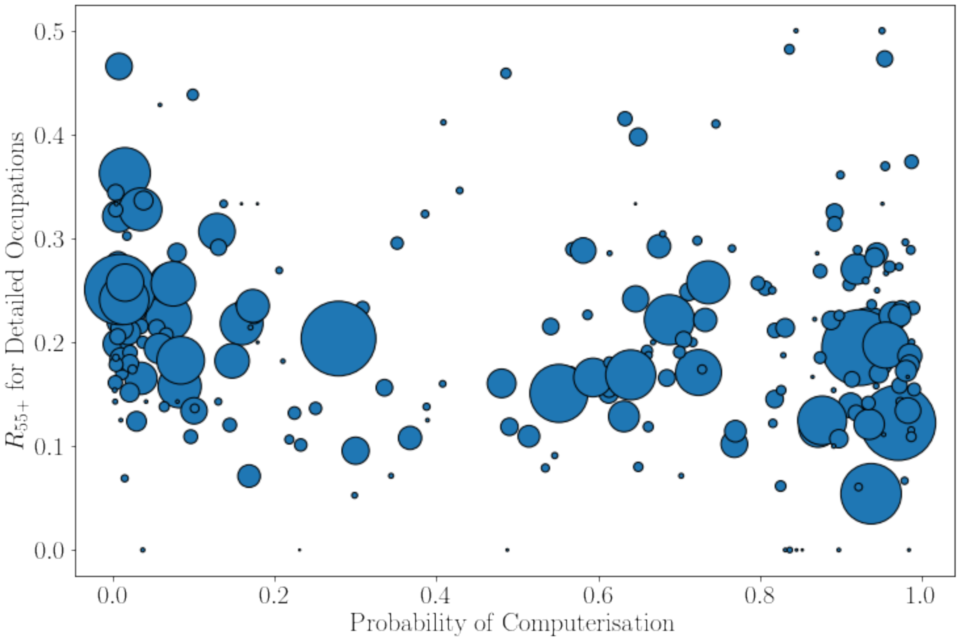
We create a scatterplot of the Detailed Occupations within *joint_auto*, with the R_{65+}/R_{55+} values (for 2012 only since the P_{Com} values were calculated based on 2012 data) on the y-axis and the P_{Com} values on the x-axis. We also scale each datapoint relative to the total number of people employed within that occupation, i.e. occupations with more people will be bigger on the plot. This gives us Figure 11, which we can see has no obvious trends. Zooming in on specific regions of the scatterplot (e.g. the region of high P_{Com}) also yields no significant patterns. We also calculate 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 P_{Com} region or the high P_{Com} region, with relatively few in between. This is consistent with the findings from Frey and Osborne, 2017, 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.

Our next step of exploration is to take the base 10 logarithmic[8] of R_{65+}/R_{55+} and P_{Com}, and plot the former against the latter in a scatterplot, as shown in Figure 12. Interestingly, the $\log_{10}(R_{55+})$ plot in Figure 12b seems to show a slight downward trend as $\log_{10}(P_{\text{Com}})$ increases; plotting the scatterplot on a logarithmic scale seems to have revealed a previously hard-to-spot trend. We also calculate the correlation coefficients for this particular plot (along with the corresponding plot for R_{65+}), 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; Frey and Osborne, 2017 noted 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

[8] As can be seen in Figure 11, the occupations with zero values for R_{65+}/R_{55+} 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.



(a) R_{65+}



(b) R_{55+}

Figure 11: Plot of R_{65+}/R_{55+} (for each Detailed Occupation) against P_{Com}. There seems to be no obvious patterns or trends present.

less likely to have their jobs computerised, tend to retire later than low skilled workers (Himmelreicher, Hagen, and Clemens, 2009).

Plotting the dataset grouped by Category Labels or Major Groups do not seem to yield any

Type of correlation	$\log_{10}(R_{65+})$	$\log_{10}(R_{55+})$
Pearson	0.00579	-0.113
Spearman	-0.0151	-0.0661
Kendall	-0.00881	-0.0429

Table 3: Correlation coefficients for Figure 12.

Type of correlation	R_{65+}	R_{55+}
Pearson	0.0143	0.180
Spearman	0.0247	0.190
Kendall	0.0191	0.128

Table 4: Correlation coefficients for Figure 13

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.

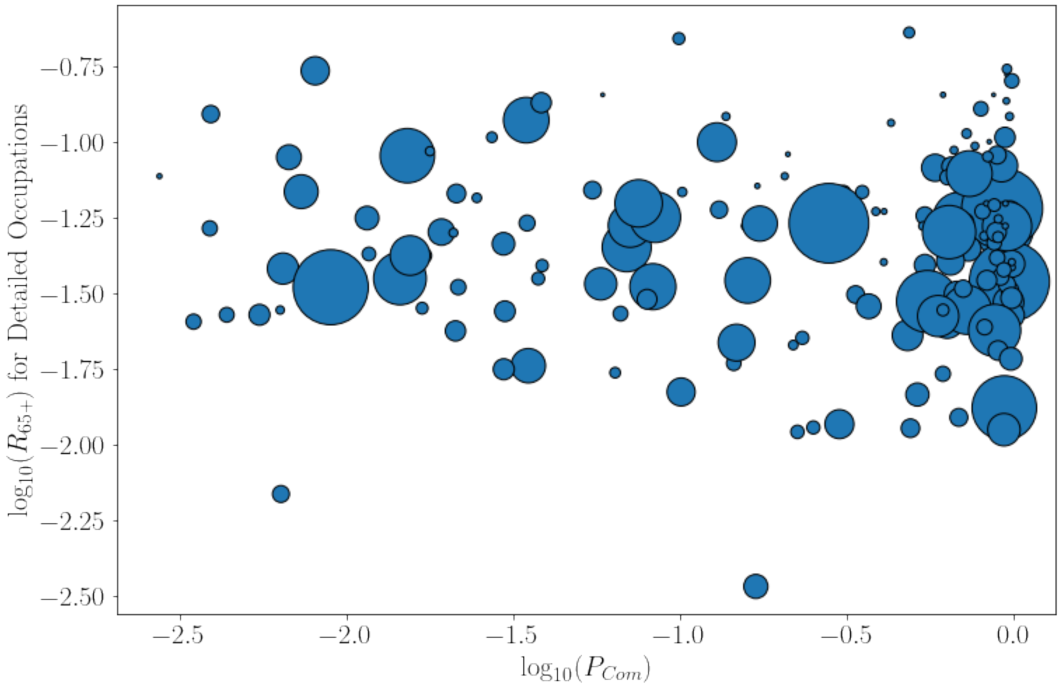
Relative change of R_{65+}/R_{55+} against P_{Com}

In the previous section, we plotted R_{65+}/R_{55+} values for 2012 against P_{Com} on a scatterplot. It seems like a waste to not use the R_{65+}/R_{55+} values for the other years given all our efforts to standardise the dataset in Chapter 2. We can use the R_{65+}/R_{55+} values for 2011 and 2021 to calculate the relative change of R_{65+}/R_{55+} for each Detailed Occupation from 2011 to 2021, similar to what we did in Figure 7. Plotting the relative change values against P_{Com} in a scatter plot gives us Figure 13. We also calculate the correlation coefficients, which can be found in Table 4. Although there seems to be no obvious trends in Figure 13, the Spearman correlation coefficient for the relative change of R_{55+} with P_{Com} is relatively high, which may suggest a monotonically increasing trend.

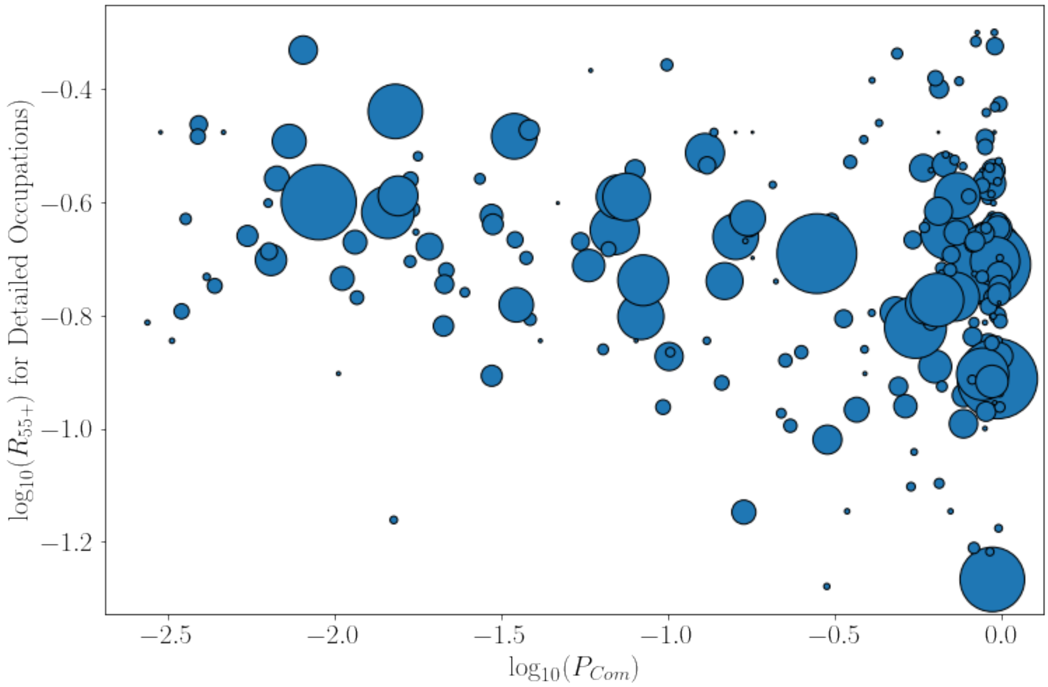
Due to the fact that the relative change of R_{65+}/R_{55+} could be negative, taking the logarithmic would prove to be tricky. It is hard to justify simply removing negative values from the dataset because we would be ignoring all occupations which became 'younger' from 2011 to 2021 when they are actually an essential part of the dataset. Hence, we will not plot a logarithmic scale scatterplot as we had done in the previous section.

4.3 Conclusion of data analysis

We did not find any strong trends despite using different metrics and plot scales. The most interesting results arose from the R_{55+} values, specifically the relationships of $\log_{10}(R_{55+})$ and relative change of R_{55+} with respect to $\log_{10}(P_{\text{Com}})$ and P_{Com} 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 the number of people employed within that particular occupation. However, our correlation coefficients were calculated without regard for this scaling. Hence, it might be worth investigating the weighted

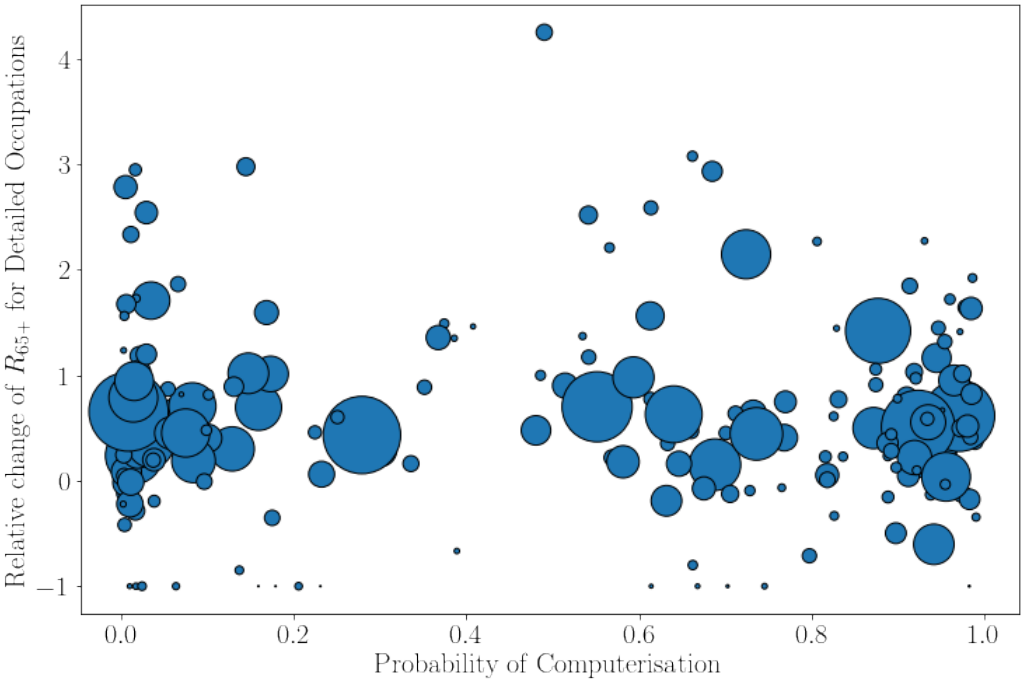


(a) $\log_{10}(R_{65+})$

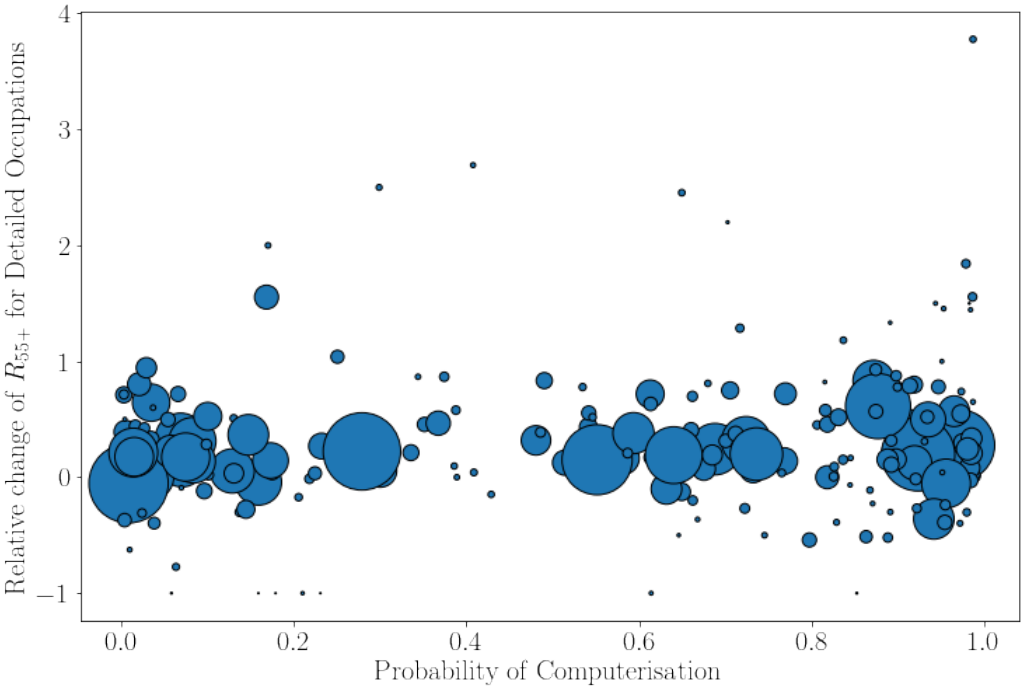


(b) $\log_{10}(R_{55+})$

Figure 12: Plot of $\log_{10}(R_{65+})/\log_{10}(R_{55+})$ (for each Detailed Occupation) against $\log_{10}(P_{\text{Com}})$. There seems to be a slight downward trend for the $\log_{10}(R_{55+})$ plot. This suggests that occupations that are less susceptible to computerisation tend to have a higher proportion of old people.



(a) R_{65+}



(b) R_{55+}

Figure 13: Plot of relative change of R_{65+}/R_{55+} from 2011 to 2021 (for each Detailed Occupation) against P_{Com}. There does not appear to be any noticeable trends.

Variables	**Weighted Pearson correlation**
Relative change of R_{55+} vs P_{Com}	0.198
$\log_{10}(R_{55+})$ vs $\log_{10}(P_{\text{Com}})$	-0.434

Table 5: Weighted Pearson Correlation showing that the strongest trend we have found so far is the relationship between $\log_{10}(R_{55+})$ for 2012 and $\log_{10}(P_{\text{Com}})$.

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 (Bailey et al., 2018):

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

Applying these equations on our data would give us the values in Table 5. The weighted correlation for $\log_{10}(R_{55+})$ and $\log_{10}(P_{\text{Com}})$ is much larger (in absolute terms) than the unweighted correlation, and better matches what we see in Figure 12b. It's absolute value is also larger than the weighted correlation for relative change of R_{55+} and P_{Com}.

Hence, the strongest trend so far is the relationship between $\log_{10}(R_{55+})$ and $\log_{10}(P_{\text{Com}})$. In addition, it appears to be a somewhat linear trend. Therefore, we will 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 three linear regression techniques (using the Scikit-learn package[9] for Python) on the $\log_{10}(R_{55+})$ and $\log_{10}(P_{\text{Com}})$ values, specifically weighted least squares, weighted Bayesian ridge, and RANdom SAmple Consensus (RANSAC). We take the weights into account for the models since we got the highest correlation when weights were taken into account. All three techniques can be applied on the other data values; we only picked the $\log_{10}(R_{55+})$ vs $\log_{10}(P_{\text{Com}})$ 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 extensions section in Chapter 5. Hence, our selection of models, which are a good mix of

[9] https://scikit-learn.org

Bayesian and non-Bayesian models, can serve as a robustness check; if the three models have similar results, this will increase our confidence in the trend.

We shall start with a quick overview of a least squares linear model. In a least squares model, the dependent variable is a linear combination of the independent variables along with an error term:

$$y_i = x_i^T \beta + \epsilon_i \quad \text{for} \quad i = 1, ..., m, \tag{2}$$

where m is the number of samples, y_i is the i^{th} observation of the dependent variable, x_i^T is the corresponding row vector of all the independent variables, and ϵ_i is the corresponding error term. In our case, x_i and β are both just 2×1 vectors since we only have the one independent variable P_{Com} (the second element of the vectors corresponds to the y-intercept of the linear model, i.e. $x_i = [x_{i,0}, 1]^T$ and $\beta = [\beta_0, \beta_1]^T$). We also define the vector $y = [y_1, ..., y_m]^T$ and matrix $X^T = [x_1, ..., x_m]$.

Weighted Least Squares

In a weighted least squares model, we aim to minimise (with respect to β) the following function:

$$S = \sum_{i=1}^{m} w_i (y_i - x_i^T \beta)^2, \tag{3}$$

where w_i is the corresponding i^{th} scalar weight.

Doing so gives us the plot in Figure 14a, which we can see fits the downward trend fairly well. Note that ordinary least squares is a special case of weighted least squares with all weights being equal to unity.

Bayesian Ridge

For a Bayesian approach, we implement linear regression using probability distributions rather than point estimates like with the weighted least squares method. We model our observation y as follows:

$$\mathcal{P}(y|\beta, X, \alpha) = \mathcal{N}(y|X\beta, \alpha), \tag{4}$$

where α is a hyper parameter.

We also model the prior for the coefficients β as follows:

$$\mathcal{P}(\beta|\lambda) = \mathcal{N}(\beta|0, \lambda^{-1} I), \tag{5}$$

where λ is another hyper parameter and I is a 2×2 identity matrix.

Additionally, we model the priors for α and λ as gamma distributions. We then write out the posterior distribution as follows:

$$\mathcal{P}(\beta|y, X, \lambda, \alpha) = \frac{\mathcal{P}(y|\beta, X, \lambda, \alpha)\mathcal{P}(\beta|X, \lambda, \alpha)}{\mathcal{P}(y|X, \lambda, \alpha)}. \tag{6}$$

The hyper parameters α and λ are estimated by maximising the log marginal likelihood (method for doing so is described in Tipping, 2001). Using these estimates and the above probability distributions, we can obtain an estimate for β by taking the mean of the posterior distribution in Equation 6. Note that the Scikit-learn implementation[10] of Bayesian Ridge involves an iterative process of updating the posterior mean and calculating the corresponding root mean square error between the predicted values of y and the actual values of y, and then updating the estimates of the hyper parameters α and λ using the calculated root mean square error. The weights of each datapoint are taken into account in this implementation by scaling each datapoint (x_i, y_i) by the square root of the corresponding weight w_i to get $(scaled\_x_i, scaled\_y_i) = (\sqrt{w_i}x_i, \sqrt{w_i}y_i)$ [11]. By doing so, datapoints with a higher weight will contribute more to the root mean square error, thereby having a bigger effect on the update of the hyper parameters and consequently the update of the posterior mean. This gives us Figure 14b, which we can see is similar to Figure 14a.

RANdom SAmple Consensus (RANSAC)

Unlike the previous two regression models, RANSAC models attempt to identify the inliers and outliers, and then completely ignore the outliers.

The RANSAC algorithm has two main steps. Firstly, a subset of the datapoints is randomly chosen to construct a model and its parameters. Secondly, the algorithm tests all datapoints against this model. Datapoints which exceed a certain error threshold would be labelled as outliers while the rest would be inliers. These two steps are then repeated until the number of inliers exceed a certain selected threshold. In our case, the model constructed is restricted to a linear model, and the error is calculated using Equation 3 (thereby allowing us to incorporate the weights into the training process of the model).

Applying RANSAC on our datapoints give us Figure 14c. We can see that the algorithm correctly identifies the outliers and gives a regression line that is very similar to the ones predicted by weighted least square and Bayesian ridge.

Conclusion of linear regression

The three linear regression models produced similar results. We applied these models on our data in order to quantify the strength of the relationship between $\log_{10}(R_{55+})$ and $\log_{10}(P_{\text{Com}})$. These models could also be used to predict the R_{55+} values of the occupations missing from our *joint_auto* dataset

[10] https://github.com/scikit-learn/scikit-learn/blob/9aaed4987/sklearn/linear_model/_bayes.py

[11] Calculating the root mean square error with this scaling is equivalent to using the weighted root mean square S defined in Equation 3.

(recall that our dataset does not feature all Detailed Occupations as shown in Figure 8). However, we note that there is a high level of variance around our linear regression lines, which makes it challenging for the models to confidently predict the R_{55+} values for the missing datapoints. Indeed, if we use the Bayesian ridge model, which inherently quantifies confidence levels, to make predictions about the 462 missing occupations, we get a very high standard deviation as can be seen in Figure 15. The high variance also makes validation of our models challenging since the error of predictions would always be somewhat high. Additionally, we note that a large portion of our known 240 occupations featured in our linear regression models have relatively few people within those occupations; in other words, a lot of the training datapoints were assigned small weights in our models, resulting in a minority of datapoints having a disproportionately larger influence on the model. Consequently, the subset of datapoints with larger weights is too limited in size to be effectively divided into separate training and testing datasets, which again makes it tricky to validate our models. It would be ideal if we have a model that could be generalised to missing values with high confidence without having to worry about validation. This would allow the model to predict the missing datapoints with high confidence. More specifically, we would like to define a region that encompasses our known datapoints (*joint_auto*) and has a low probability of not encompassing the missing datapoints, while also quantifying 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 m 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:

$$\mathbb{P}^m\{\delta_1, ..., \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon\} \geq 1 - q(m, \epsilon), \tag{7}$$

where $1 - q(m, \epsilon)$ is a lower bound to our confidence that the probability of a new sample not belonging in H_m is less than or equal to ϵ. We refer the reader to Campi and Garatti, 2008 for a more comprehensive introduction to this concept. In fact, this entire section on PAC Learning is heavily based on the work done by Campi and Garatti, 2008; Romao, Margellos, and Papachristodoulou, 2021; Romao, Papachristodoulou, and Margellos, 2022.

Furthermore, consider the following convex scenario program:

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^T x \\ \text{s.t.} \quad & g(x, \delta_i) \leq 0, \quad \text{for all i = 1,...,m.} \end{aligned} \tag{8}$$

Suppose δ_i belongs to an uncertainty space[12] 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 8. 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 7 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:

$$\mathbb{P}^m\{\delta_1, ..., \delta_m : \mathbb{P}(\delta \in \Delta : g(x^*_m, \delta) > 0) \leq \epsilon\} \geq 1 - q(m, \epsilon), \tag{9}$$

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 7, 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 8 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 8 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).*

Note that if a problem is fully-supported, it is non-degenerate, but the converse does not hold.

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

$$q(m, \epsilon) = \sum_{i=0}^{n_x - 1} \binom{m}{i} \epsilon^i (1 - \epsilon)^{m-i}. \tag{10}$$

[12] Δ denotes the set of values that δ can take with non-zero probability, and δ is distributed according to some possibly unknown probability distribution.

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} $\log_{10}(P_{\text{Com}})$ and i^{th} $\log_{10}(R_{55+})$ 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, our 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 $x_2, x_3 \in \mathbb{R}$ to encode the median line (as the gradient and y-intercept respectively), and a variable $x_1 \in \mathbb{R}^+$ (set of all non-negative real numbers) 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:

$$\begin{aligned} \min_{x_1 \in \mathbb{R}^+, x_2 \in \mathbb{R}, x_3 \in \mathbb{R}} \quad & x_1 \\ \text{s.t. } y_i - x_2 u_i - x_3 \leq \quad & x_1, \quad \text{for all i = 1,...,m,} \\ y_i - x_2 u_i - x_3 \geq \quad & -x_1, \quad \text{for all i = 1,...,m.} \end{aligned} \tag{11}$$

At first glance, this might seem different from convex scenario program 8. However, by placing x_1, x_2, x_3 into a vector $x \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}$, and rearranging all the constraints into a matrix inequality constraint, we can see that both programs are of the same form. Additionally, the domain of x 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 8. Solving this program and plotting our minimum vertical width strip H_m gives Figure 16.

Using Equations 9,10 and $\epsilon = 0.01$, we calculate that $\mathbb{P}^m\{\delta_1, ..., \delta_m : \mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0) \leq \epsilon\}$ is at least 0.407. This means that we are at least 40.7% confident that there is at most a 1% chance that a new sample falls outside of our hypothesis H_m in Figure 16. The confidence lower bounds for other values of ϵ are shown in Table 6. The lower bound shoots up to almost 100% with just an ϵ value of 0.05. This is not too surprising given that H_m covers a relatively huge area.

While H_m contains all of our datapoints and keeps 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,

ϵ	$1 - q(m, \epsilon)$
0.01	0.4070228538
0.05	0.9993799909
0.10	0.9999999905

Table 6: Confidence lower bound for corresponding ϵ values when using all samples.

Romao, Papachristodoulou, and Margellos, 2022 details such an approach with a specific scenario discarding scheme.

Using scenario program 8 as reference, let us assume that x^*_m exists and is unique, and that the scenario program is fully-supported. We then discard the support set for x^*_m, which corresponds to removing n_x constraints (recall Definition 2). We are then left with a new scenario program with $m - n_x$ number of constraints. Assuming the optimiser for this new program exists and is unique, and that this new program is also fully-supported, we can repeat the discarding process to obtain a third scenario program with $m - 2n_x$ constraints. We can repeat this k number of times, removing $r = kn_x$ constraints in total and leaving us with a scenario program with $m - kn_x$ constraints. Hence, we get the following:

$$q(m, \epsilon) = \sum_{i=0}^{r+n_x-1} \binom{m}{i} \epsilon^i (1 - \epsilon)^{m-i}. \tag{12}$$

Note that we assumed above that all k new scenario programs obtained through this discarding scheme are fully-supported. Even if this assumption does not hold, we can introduce a lexicographic order as a way to select which constraints to discard, in which case we can still use the above result provided that all the k new scenario programs as well as the original program are non-degenerate (recall that fully-supported implies non-degenerate, but not the other way around).

By assuming x^*_{m-j} for our scenario program 11 is unique for $j = n_x, 2n_x, ..., kn_x$, and that the scenario program, along with all subsequent programs with reduced constraints, are at least non-degenerate, we can apply the above scenario discarding scheme as a way of excluding outlier datapoints. Figure 17 shows the plots for various values of r.

For $r = 60$, the confidence region is constrained to the collection of datapoints that reflect the downward trend. 60 discarded datapoints may seem like a lot, but they only represent 12.9% of the total number of employed people present in *joint_auto* in 2012. This means that a large proportion of the employed people represented in our dataset are captured within H_{m-60}. Using Equations 9,12, we can calculate the lower bounds for our confidence for various values of ϵ when $r = 60$ (Table 7).

While we are at least about 0% confident that at most 1% of new datapoints will lie outside H_{m-60} (which is not very helpful), we can be at least 50% sure that at most 10% of new datapoints will lie

ϵ	$1 - q(m, \epsilon)$
0.01	2.331468×10^{-15}
0.05	0.001427
0.10	0.540832
0.15	0.990461
0.20	0.999988

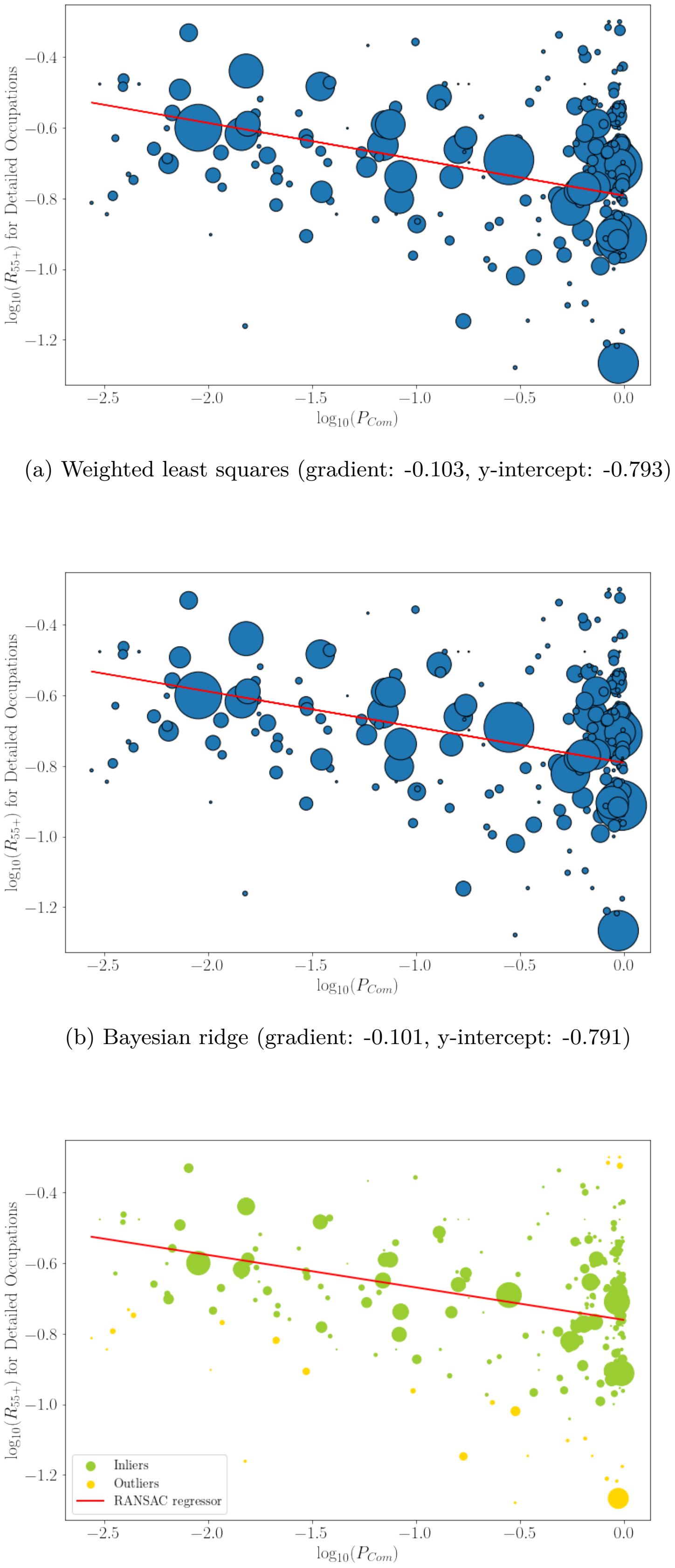
Table 7: Confidence lower bound for corresponding ϵ values when $r = 60$.

outside H_{m-60} and pretty much 100% confident that at most 15% of new datapoint will lie outside H_{m-60}.

However, the gradient ($x_2 = -0.0304$) of the median line of H_{m-60} is smaller in absolute terms than the gradients of the linear regression lines calculated earlier in Figure 14. This limitation in capturing the underlying trend suggests that the PAC Learning model's predictive power may be inferior to that of the linear regression models. Recall our earlier discussion on predicting the R_{55+} values of the occupations missing from the *joint_auto* dataset. It might be better to use the linear regression models in conjunction with the PAC Learning model; specifically, we can use the former to obtain point estimates and the latter can provide a measure of confidence around those estimates. Referring to the predicted values in Figure 15, we can be almost 100% confident that at least 85% of them (which amounts to at least 392 occupations out of the 462 missing occupations) will lie within H_{m-60}. This is illustrated in Figure 18. Crucially, this claim is made solely based on the predictions and theoretical guarantees of our models, rather than validation of the models on a test set.

Conclusion of PAC Learning

In this section, we discussed the basic theory behind PAC Learning, and applied it to our dataset. We then noted that the presence of potential outliers were affecting the performance of our PAC Learning model and introduced a scenario discarding scheme that was devised for control problems. We used this scheme as a way of excluding outliers from our model and showed that we could still obtain decent confidence lower bounds. In doing so, we have demonstrated the viability of using these techniques to build a minimum region of confidence for any dataset (recall that we did not make any assumptions about the underlying probability distribution) provided that the dataset can be considered i.i.d samples of the population. The assumptions of the existence of a unique optimiser and non-degeneracy must also hold. Finally, we combined the predictive power of the Bayesian ridge model with the confidence guarantees of the PAC Learning model to ensure more robust and reliable results.



(a) Weighted least squares (gradient: -0.103, y-intercept: -0.793)

(b) Bayesian ridge (gradient: -0.101, y-intercept: -0.791)

(c) RANSAC (gradient: -0.0919, y-intercept: -0.761)

Figure 14: Plot of $\log_{10}(R_{55+})$ against $\log_{10}(P_{\text{Com}})$. The red lines represent the respective linear regression lines. All three models seem to agree on the strength of the relationship. Additionally, all three models would give similar predictions for unknown datapoints.

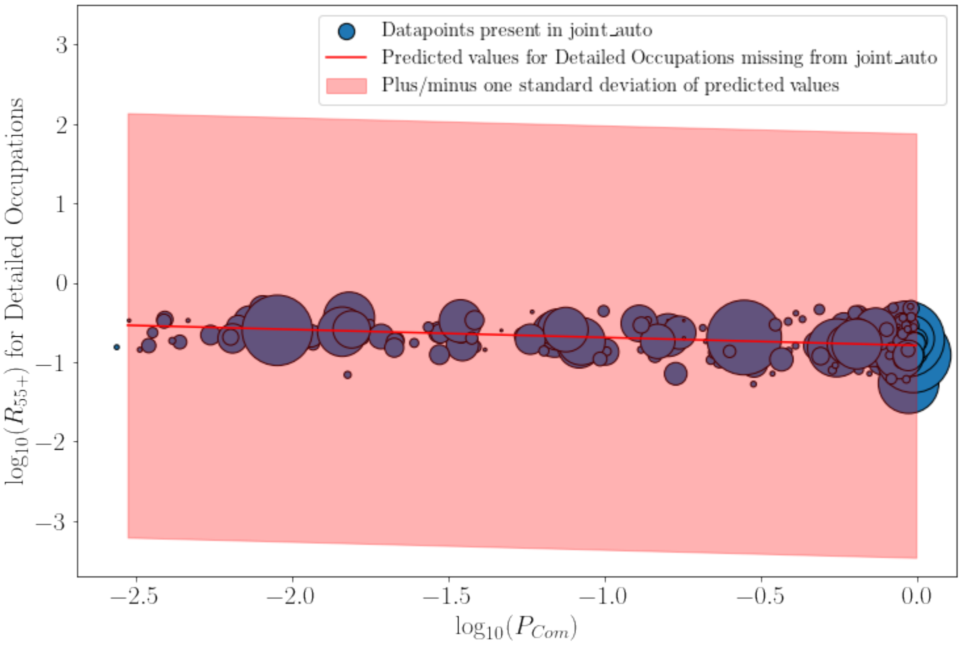


Figure 15: Plot of $\log_{10}(R_{55+})$ against $\log_{10}(P_{\text{Com}})$. The predicted values for the Detailed Occupations present in the automatability dataset but missing from the *joint_auto* dataset match the regression line in Figure 14b. However, the standard deviation of the predicted values is very large (about 2.67), which means the Bayesian ridge model is not confident about its predictions.

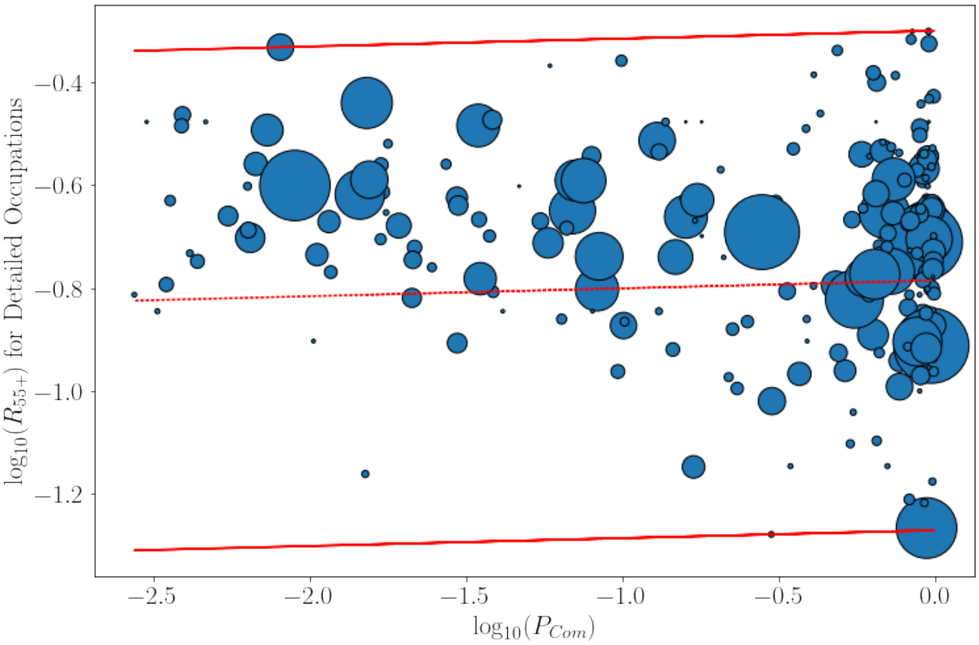
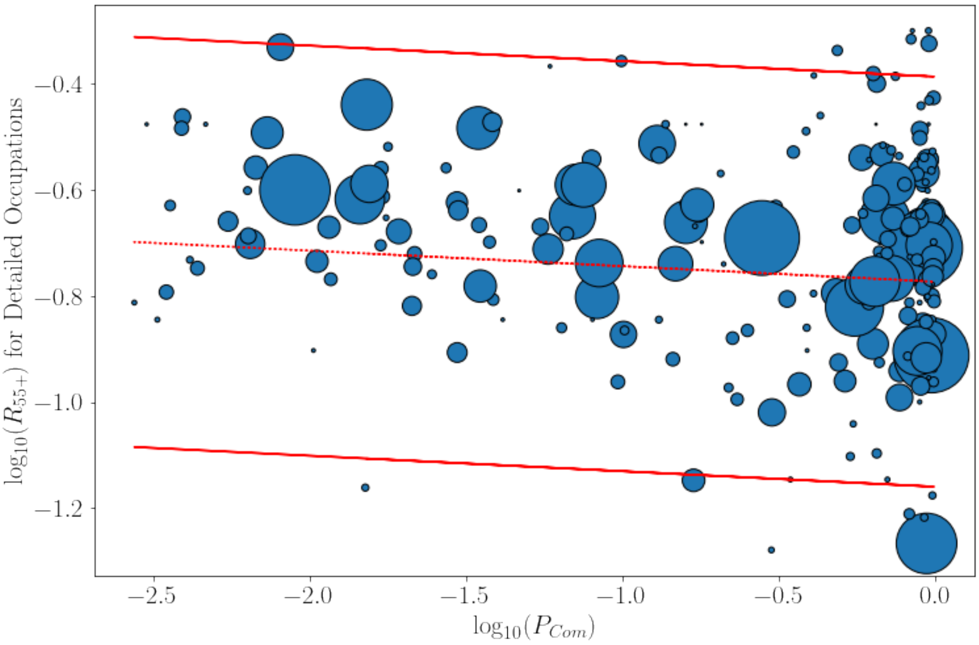
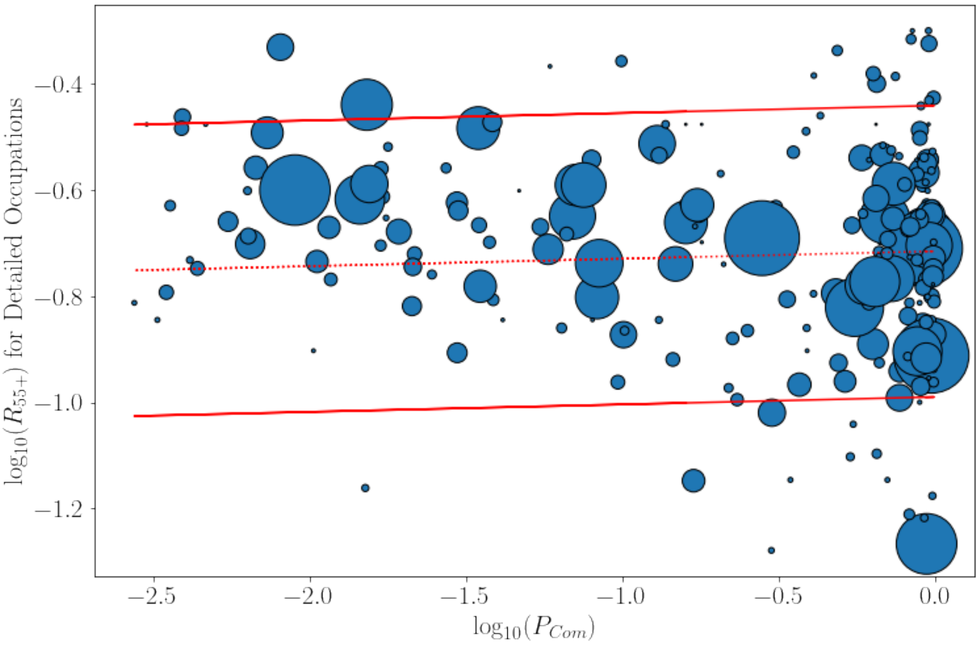


Figure 16: Plot of $\log_{10}(R_{55+})$ (for each Detailed Occupation) against $\log_{10}(P_{\text{Com}})$ with PAC Learning applied ($x_1 = 0.485, x_2 = 0.0153, x_3 = -0.785$). The red solid lines represent the boundaries of H_m while the red dotted line is the median line. We can see that H_m contains all of our datapoints while ensuring that the vertical width is minimised.

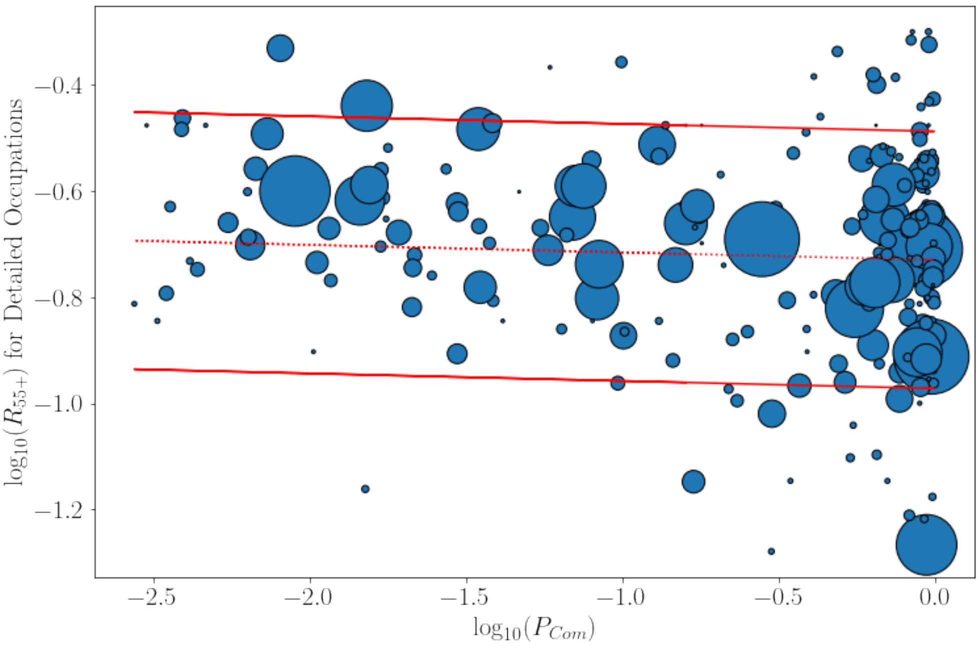


(a) $r = 15$ $(x_1 = 0.386, x_2 = -0.0292, x_3 = -0.774)$

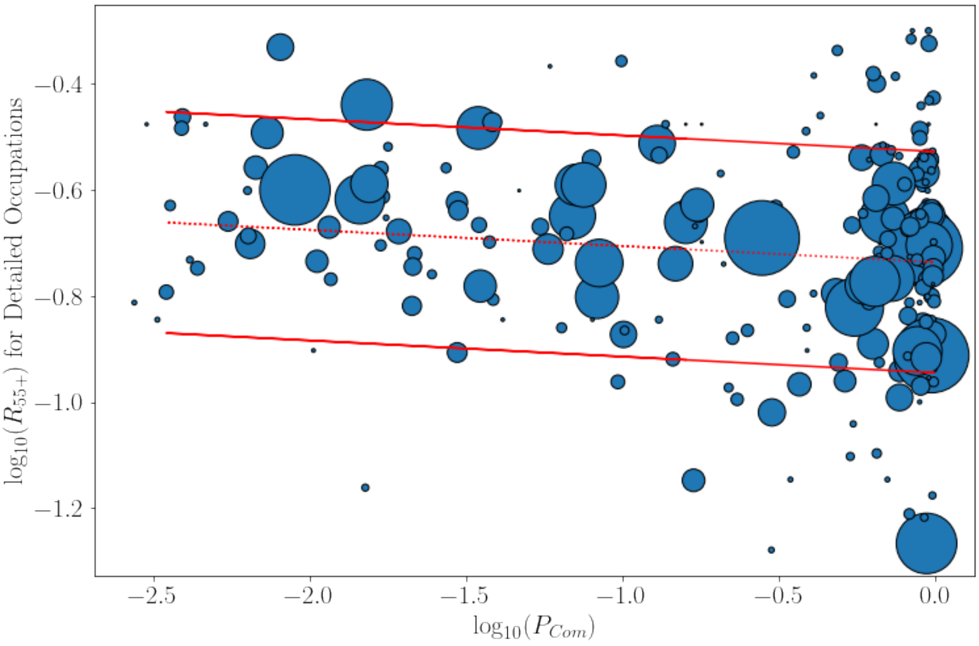


(b) $r = 30$ $(x_1 = 0.274, x_2 = 0.0140, x_3 = -0.716)$

Figure 17



(c) $r = 45$ ($x_1 = 0.242, x_2 = -0.0142, x_3 = -0.730$)



(d) $r = 60$ ($x_1 = 0.208, x_2 = -0.0304, x_3 = -0.737$)

Figure 17: Plot of $\log_{10}(R_{55+})$ (for each Detailed Occupation) against $\log_{10}(P_{\text{Com}})$ with PAC Learning and scenario discarding scheme applied. The red solid lines represent the boundaries of H_{m-r} while the red dotted line is the median line. We can see H_{m-r} starts to reflect the downward trend that we identified in the data. x_1 also gets smaller as r increases, which corresponds to the decreasing vertical distance between the two red boundary lines.

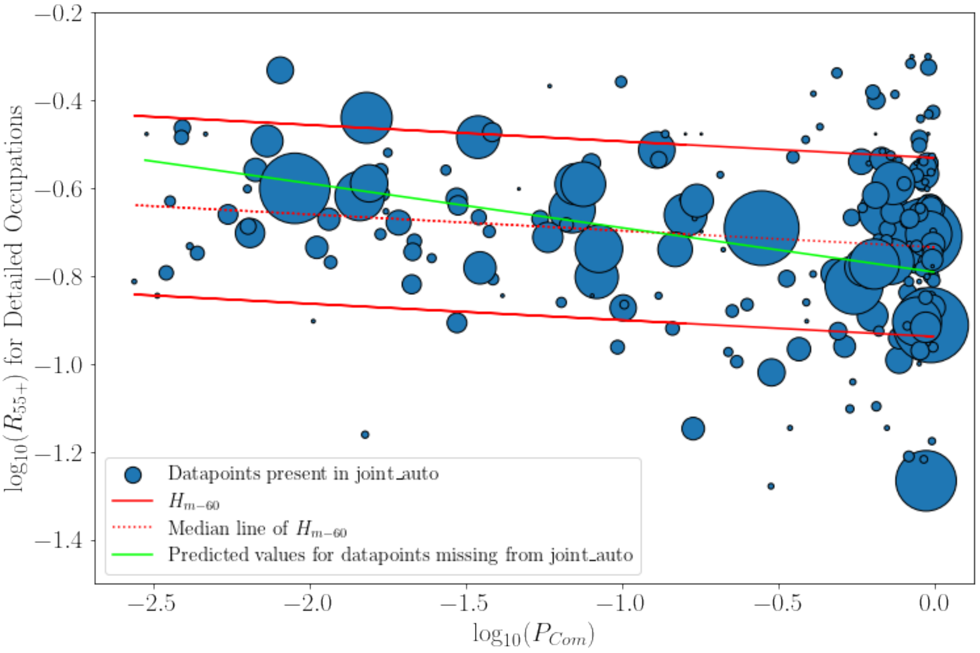


Figure 18: Plot of $\log_{10}(R_{55+})$ (for each Detailed Occupation) against $\log_{10}(P_{\text{Com}})$. The line representing the predicted values is obtained from Figure 15 (Bayesian ridge). We can be nearly 100% confident that at least 85% of the 462 predicted values will actually lie within H_{m-60}.

5 Conclusion

In this report, we attempt to examine the relationship between automation and age distribution within occupations. To do so, we took age distribution data from the US Bureau of Labor Statistics and standardised all of it according to the 2018 Standard Occupational Classification system, and mapped it to the dataset containing Probability of Computerisation values from Frey and Osborne, 2017. This gave us a joint standardised dataset covering the years 2011 to 2021, which may be useful to others doing related work. We also found that there was a general trend of ageing across most occupations. We then explored trends within this joint dataset and discovered a possible trend: an inverse relationship between the proportion of workers aged 55 years and older within an occupation and the Probability of Computerisation of that occupation. We applied weighted least squares, Bayesian ridge, and Random Sample Consensus regression to this set of variables. Given the high variance around the regression lines, we wondered if we could instead construct a region that we have high confidence would contain a high percentage of new datapoints. To do so, we turned to Probably Approximately Correct Learning, and used it in conjunction with a scenario discarding scheme originally devised for optimisation programs within control problems. This allowed us to construct the region while excluding some outliers. We concluded that it was possible to construct a region for which we are almost 100% confident would contain at least 85% of new datapoints. This also acts as a proof of concept of using such a technique on datasets with high variance.

However, we cannot offer guarantees of this trend's statistical significance without further tests and research. Hence, we tentatively state that we have found no significant relationships between age distribution within occupations and their automatability for the most part, and conclude that governments should not expect that the loss of jobs from automation will automatically balance out the labour force shortages caused by an ageing population. Thus, we recommend that they take more active steps towards addressing these issues. Indeed, job losses are likely to accelerate over the next few years given the recent major advances in Artificial Intelligence (AI) which gave us AI models such as ChatGPT[13] and MidJourney[14]. Goldman Sachs predicts that 300 million jobs are at risk of automation due to generative AI (Hatzius et al., 2023), while the World Economic Forum estimates that 85 million jobs are at risk due to AI (World Economic Forum, 2020). This will likely result in a huge number of people needing to be retrained or reskilled. In fact, an IBM study predicts that 120 million people may need retraining over the next three years (LaPrade et al., 2019), and this will just be due to automation without taking into account the effects of ageing populations. That being said, there is cause for some optimism. While AI will lead to loss of jobs, it is also expected to create

[13] https://openai.com/blog/chatgpt

[14] https://www.midjourney.com

new jobs (Hawksworth, Berriman, and Goel, n.d.; World Economic Forum, 2020). This technology is also predicted to contribute substantially to economic growth by increasing productivity (Hatzius et al., 2023; LaPrade et al., 2019; Rao and Verweij, n.d.). Governments around the world have the opportunity to strategically leverage the economic benefits derived from AI to support their respective countries' ageing population while simultaneously investing in the retraining and reskilling of older and/or at-risk workers. Such a two-pronged approach could mitigate the negative effects of both an ageing population and automation. If successfully managed, governments can not only prevent the fears expressed by Ricardo, Keynes, and Reuther (Chapter 1) from becoming reality, but also create inclusive, diverse, and forward-looking societies that embrace technological advancements.

Future extensions

In Chapter 2.1, we processed the BLS data to get a standardised dataset that conforms to the 2018 SOC standards. This processed dataset may prove useful to others doing research on ageing population.

In the future, more work can be done on expanding on the bias test in Chapter 4.1 in order to get a better sense of how well the *joint_auto* dataset represents the US labour force. This will allow us to more appropriately generalise any trends found in the dataset to the wider country.

As mentioned in Chapter 4.4, we focused on the $\log_{10}(R_{55+})$ vs $\log_{10}(P_{\text{Com}})$ relationship because it is the strongest relationship we have found so far. We do not offer any guarantees that the $\log_{10}(R_{55+})$ vs $\log_{10}(P_{\text{Com}})$ relationship is statistically significant. Hence, future extensions should also focus on testing the statistical significance of this particular relationship. This would validate our results from our linear regression and PAC Learning models. That being said, our work on PAC Learning in Chapter 4.5 is more of a proof-of-concept that is just using the aforementioned relationship as an example. It can be applied to any dataset as long as the assumptions of i.i.d, existence of unique optimiser, and non-degeneracy hold.

Following on the previous point about PAC Learning, it may be possible to find even tighter bounds on our level of confidence $1 - q(m, \epsilon)$ to improve the performance of our PAC Learning models. Unfortunately, this is currently beyond the capabilities of this author due to the lack of experience and expertise in that particular area of research. That being said, we can keep ourselves updated on works pertaining to that area of research and apply any relevant new findings to our models in future extensions to this project.

Additionally, more exploratory work can be done on the *joint_auto* dataset to find other statistically significant relationships. Specifically, we can include more variables (such as the ones from the original automatability dataset from Frey and Osborne, 2017) in our models.

References

(N.d.). URL: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?most_recent_value_desc=true&amp;year_high_desc=true.

Acemoglu, Daron and Pascual Restrepo (May 2017). "Secular Stagnation? The Effect of Aging on Economic Growth in the Age of Automation". In: *American Economic Review* 107.5, pp. 174–79. DOI: 10.1257/aer.p20171101. URL: https://www.aeaweb.org/articles?id=10.1257/aer.p20171101.

— (2020). "Robots and jobs: Evidence from US labor markets". In: *Journal of political economy* 128.6, pp. 2188–2244.

Ageing and health - China (n.d.). URL: https://www.who.int/china/health-topics/ageing.

Bailey, Paul et al. (2018). "Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas. AIR–NAEP Working Paper No. 2018-01. NCES Data R Project Series# 02." In: *American Institutes for Research.*

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

Beardson, Timothy and Nick Fielding (2021). *Ageing giant : China's looming population collapse.* eng. Oxford. ISBN: 9781909930988 (hardback).

Brzeski, Carsten and Inga Burk (2015). "Die Roboter kommen". In: *Folgen der Automatisierung für den deutschen Arbeitsmarkt. INGDiBa Economic Research* 30.

Campi, M. C. and S. Garatti (2008). "The Exact Feasibility of Randomized Solutions of Uncertain Convex Programs". English. In: *SIAM Journal on Optimization* 19.3. Copyright - Copyright] © 2008 Society for Industrial and Applied Mathematics; Last updated - 2022-10-20, pp. 1211–20. URL: https://www.proquest.com/scholarly-journals/exact-feasibility-randomized-solutions-uncertain/docview/920844687/se-2.

Cheng, Xunjie et al. (June 2020). "Population ageing and mortality during 1990–2017: A global decomposition analysis". In: *PLOS Medicine* 17.6. Ed. by Sanjay Basu, e1003138. DOI: 10.1371/journal.pmed.1003138. URL: https://doi.org/10.1371/journal.pmed.1003138.

Cho, Il Haeng, Kyung S Park, and Chang Joo Lim (2010). "An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI)". In: *Mechanisms of ageing and development* 131.2, pp. 69–78.

D'Ambrogio, Enrico (2020). "Japan's ageing society". In.

Dosi, Giovanni et al., eds. (1988). *Technical Change and Economic Theory.* Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy. URL: https://EconPapers.repec.org/RePEc:ssa:lembks:dosietal-1988.

Elderly Population (n.d.). URL: https://data.oecd.org/pop/elderly-population.htm.

Falk, Gene et al. (Aug. 2021). *Unemployment Rates During the COVID-19 Pandemic.* URL: https://crsreports.congress.gov/product/pdf/R/R46554.

Foley, Kevin T and Clare C Luz (Oct. 2020). "Retooling the Health Care Workforce for an Aging America: A Current Perspective". In: *The Gerontologist* 61.4. _eprint: https://academic.oup.com/gerontologist/artic pdf/61/4/487/38445439/gnaa163.pdf, pp. 487–496. ISSN: 0016-9013. DOI: 10.1093/geront/gnaa163. URL: https://doi.org/10.1093/geront/gnaa163.

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

Frey, Carl Benedikt and Michael A Osborne (2017). "The future of employment: How susceptible are jobs to computerisation?" In: *Technological forecasting and social change* 114, pp. 254–280.

Fuei, Lee King (2017). "Automation, computerization and future employment in Singapore". In: *Journal of Southeast Asian Economies*, pp. 388–399.

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

Graetz, Georg and Guy Michaels (2018). "Robots at work". In: *Review of Economics and Statistics* 100.5, pp. 753–768.

Hatzius, J et al. (2023). "The potentially large effects of artificial intelligence on economic growth". In: *Goldman Sachs Economic Research.*

Hawksworth, John, Richard Berriman, and Saloni Goel (n.d.). *Will robots really steal our jobs? - PWC.* URL: https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf.

Hayflick, Leonard (2007). "Biological aging is no longer an unsolved problem". In: *Annals of the New York academy of Sciences* 1100.1, pp. 1–13.

Himmelreicher, Ralf K., Christine Hagen, and Wolfgang Clemens (2009). "Skills and age of retirement: Have high skilled the highest age of retirement?" ger. In: *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 61.3, pp. 437–452. ISSN: 0023-2653.

Hollander, Samuel (May 2019). "Retrospectives: Ricardo on Machinery". In: *Journal of Economic Perspectives* 33.2, pp. 229–42. DOI: 10.1257/jep.33.2.229. URL: https://www.aeaweb.org/articles?id=10.1257/jep.33.2.229.

Jylhävä, Juulia, Nancy L Pedersen, and Sara Hägg (2017). "Biological age predictors". In: *EBioMedicine* 21, pp. 29–36.

Keynes, John Maynard (2010). "Economic Possibilities for Our Grandchildren". In: *Essays in Persuasion*. London: Palgrave Macmillan UK, pp. 321–332. ISBN: 978-1-349-59072-8. DOI: 10.1007/978-1-349-59072-8_25. URL: https://doi.org/10.1007/978-1-349-59072-8_25.

Klemera, Petr and Stanislav Doubal (2006). "A new approach to the concept and computation of biological age". In: *Mechanisms of ageing and development* 127.3, pp. 240–248.

Kotter-Grühn, Dana, Anna E. Kornadt, and Yannick Stephan (Aug. 2015). "Looking Beyond Chronological Age: Current Knowledge and Future Directions in the Study of Subjective Age". In: *Gerontology* 62.1, pp. 86–93. DOI: 10.1159/000438671. URL: https://doi.org/10.1159/000438671.

Kozicki, Bartosz and Marcin Gornikiewicz (2020). "Unemployment rate in Poland and USA during COVID-19 pandemic: a case study". In.

Krugman, Paul (2019). "Globalization: What did we miss?" In: *Meeting Globalization's Challenges*, pp. 113–120.

Kyodo (Sept. 2019). *Elderly citizens accounted for record 28.4% of Japan's population in 2018, Data Show*. URL: https://www.japantimes.co.jp/news/2019/09/15/national/elderly-citizens-accounted-record-28-4-japans-population-2018-data-show/.

Lancet, The (2022). *Population ageing in China: crisis or opportunity?*

LaPrade, Annette et al. (2019). "The enterprise guide to closing the skills gap: Strategies for building and maintaining a skilled workforce". In: *IBM Institute for Business Value*.

Levine, Morgan E (2013). "Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age?" In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 68.6, pp. 667–674.

Luo, Yanan, Binbin Su, and Xiaoying Zheng (2021). "Trends and Challenges for Population and Health During Population Aging - China, 2015-2050". eng. In: *CHINA CDC WEEKLY* 3.28, pp. 593–598. ISSN: 2096-7071.

Manning, Jennifer E. (Dec. 2022). *Membership of the 117th congress: A profile*. URL: https://crsreports.congress.gov/product/pdf/R/R46705/2.

Marešová, Petra, Hana Mohelská, and Kamil Kuča (2015). "Economics aspects of ageing population". In: *Procedia economics and finance* 23, pp. 534–538.

Munger, Kevin (2022). *Why the Baby Boomers Still Dominate American Politics and Culture*. New York Chichester, West Sussex: Columbia University Press. ISBN: 9780231553810. DOI: doi:10.7312/mung20086. URL: https://doi.org/10.7312/mung20086.

Munnell, Alicia H (2013). “Social Security’s Real Retirement Age Is 70”. eng. In: *IDEAS Working Paper Series from RePEc*.

Noah, Timothy (Sept. 2019). *America, the gerontocracy.* URL: `https://www.politico.com/magazine/story/2019/09/03/america-gerontocracy-problem-politics-old-politicians-trump-biden-sanders-227986/`.

Organization, World Health et al. (2010). *Definition of an older or elderly person.*

Orimo, Hajime et al. (2006). “Reviewing the definition of “elderly””. In: *Geriatrics & gerontology international* 6.3, pp. 149–158.

Pajarinen, Mika, Petri Rouvinen, Anders Ekeland, et al. (2015). “Computerization threatens one-third of Finnish and Norwegian employment”. In: *Etla Brief* 34.22, pp. 1–8.

Parsons, Alexander J.Q. and Stuart Gilmour (2018). “An evaluation of fertility- and migration-based policy responses to Japan’s ageing population”. eng. In: *PloS one* 13.12, e0209285–e0209285. ISSN: 1932-6203.

Phiromswad, Piyachart, Sabin Srivannaboon, and Pattarake Sarajoti (Feb. 2022). “The interaction effects of automation and population aging on labor market”. In: *PLOS ONE* 17.2, pp. 1–16. DOI: `10.1371/journal.pone.0263704`. URL: `https://doi.org/10.1371/journal.pone.0263704`.

Rao, Anand S. and Gerard Verweij (n.d.). *PWC’s Global Artificial Intelligence Study: Sizing the prize.* URL: `https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html`.

Romao, Licio, Kostas Margellos, and Antonis Papachristodoulou (2021). “Tight sampling and discarding bounds for scenario programs with an arbitrary number of removed samples”. In: *Learning for Dynamics and Control.* PMLR, pp. 312–323.

Romao, Licio, Antonis Papachristodoulou, and Kostas Margellos (2022). “On the exact feasibility of convex scenario programs with discarded constraints”. In: *IEEE Transactions on Automatic Control*, pp. 1–1. DOI: `10.1109/TAC.2022.3165320`.

Rose, Elizabeth L. (2021a). “The Decline of US Manufacturing: Issues of Measurement”. eng. In: *Management and organization review* 17.1, pp. 24–28. ISSN: 1740-8776.

Rose, Stephen J. (Oct. 2021b). *Do not blame trade for the decline in manufacturing jobs.* URL: `https://www.csis.org/analysis/do-not-blame-trade-decline-manufacturing-jobs`.

Soto-Perez-de-Celis, Enrique et al. (2018). “Functional versus chronological age: geriatric assessments to guide decision making in older patients with cancer”. In: *The Lancet Oncology* 19.6, e305–e316. ISSN: 1470-2045. DOI: `https://doi.org/10.1016/S1470-2045(18)30348-6`. URL: `https://www.sciencedirect.com/science/article/pii/S1470204518303486`.

Steigerwald, David (2010). “Walter Reuther, the UAW, and the dilemmas of automation”. eng. In: *Labor history* 51.3, pp. 429–453. ISSN: 0023-656X.

The Oxford Dictionary of Sports Science & Medicine (Jan. 2006). Oxford University Press. DOI: `10.1093/acref/9780198568506.001.0001`. URL: `https://doi.org/10.1093/acref/9780198568506.001.0001`.

Thompson, Derek (Mar. 2020). *Why do such Elderly People Run America?* URL: `https://www.theatlantic.com/ideas/archive/2020/03/why-are-these-people-so-freaking-old/607492/`.

Tinker, Anthea (2002). “The social implications of an ageing population”. In: *Mechanisms of Ageing and Development* 123.7, pp. 729–735.

Tipping, Michael E (2001). “Sparse Bayesian learning and the relevance vector machine”. In: *Journal of machine learning research* 1.Jun, pp. 211–244.

Vespa, Jonathan, David M Armstrong, Lauren Medina, et al. (2018). *Demographic turning points for the United States: Population projections for 2020 to 2060.* US Department of Commerce, Economics and Statistics Administration, US ...

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.

Woirol, Gregory R and Roger E Backhouse (1997). “The technological unemployment and structural unemployment debates”. eng. In: *Journal of economic literature* XXXV.4, pp. 2075–2076. ISSN: 0022-0515.

World Economic Forum, V (2020). “The future of jobs report 2020”. In: *Retrieved from Geneva.*

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

World population ageing, 1950-2050. (2002). eng. New York: United Nations. ISBN: 9789210510929.

Appendix

Standardisation process for BLS dataset

While the BLS did provide documents[15] outlining and explaining the changes to the SOC, they are 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 merge 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[16].

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.

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

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

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 ‘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 each inconsistency, 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.