The Second Computerization Wave; Validating Predictions for the Labor Market

ERASMUS UNIVERSITEIT ROTTERDAM

30/04/2020

Master thesis

Erasmus University Rotterdam

Erasmus School of Economics

Student: Sander van Heijningen

Student number: 401134

Supervisor: Delfgaauw, J.

Second assessor: Kapoor, S.V.

The content of this thesis is the sole responsibility of the author and does not reflect the view of either the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

In the long-running automation process, a new wave is about to break lose. The second computerization wave is again predicted to take over a lot of jobs. In this thesis that assumption is tested in US and UK labor markets. In an influential paper Frey and Osborne predicted 47% of US employment to be in danger of automation. In the same paper they came up with a probability of computerization for 702 individual occupations. This research paper tries to verify if there already are developments present in today's labor markets that relate to the probability of computerization. Evidence is found that relates to a negative correlation between the probability of computerization and recent changes in employment. However more striking is the strong evidence, which is found, that the probability positively correlates to wage. Moreover, contrary to general believe, this appears to be even more true for low-skilled occupations. At the same time high-skilled occupations seem to be more at risk.

Table of contents

1. Introduction

In the age of machine learning (ML) and artificial intelligence (AI) the newest automation wave is about to break loose. Because of developments in especially artificial intelligence, the workplace is predicted to be transformed in the next decade (Berg, Buffie, & Zanna, 2018). This will not be the first-time humans will be forced to spend their time differently than they have been used to. How exactly we will be filling that time, either by more and different types of leisure or by different work, is a mystery for now. However, what we can try to investigate is how this automation wave will affect present-day labor markets.

As stated, this is not the first-time things will change for us, the AI induced automation wave is just an extension of the long going cycles of automation waves. This also means that there is already vast literature around this subject, of which most of it is still relevant today. When an automation wave is coming up, generally experts predict that it will take out mostly low-skilled jobs (Autor D. H., 2015). This has been the case with every wave since the 19th century. In today's world it is again predicted that automation will take out a lot of middle-class jobs. The perspective that has been mostly neglected in predictions of the past is the macroeconomic perspective. Where with every wave expert feared for workers to lose their jobs, most of them just took upon other types of work. As people transitioned from farmland to factories and from factories to offices not only the nature of the jobs changed, also worker productivity rose to great heights. These efficiency gains created new job opportunities, as resources were saved due to these gains, they could be invested somewhere else.

In the literature no single answer is given regarding the way the current automation wave will spread over the labor markets. However, AI substituting mainly low-skilled workers does lie in the general expectation (Buera, Kaboski, & Rogerson, 2015) (Brynjolfsson & McAfee, 2014) (Acemoglu & Autor, 2011). With the current technology AI can take over human like tasks in a controlled environment. For instance, in the case of a help desk chat function, where a customer in need is no longer talking to a human, but to a computer program. If a customer has a more complex question that the AI is not able to answer, a human customer service employee can take over. The AI being the first line of defense, will free up time for the human employee to process other questions. This collaboration between the AI and the human will then make the customer service unit more productive per human worker. As automation is likely to increase productivity of high-skilled labor. There will be extra demand for high-skilled workers, because their labor becomes more valuable, this will result in a further increase of the college wage premium and therefore inequality (Autor D. H., 2014). The labor and capital shares in value added will also further skew to capital. However, as pointed out in the previous paragraph, it is not likely to just become a win for capital and high-skilled labor and a loss for low-skilled labor. Because the general equilibrium of the economy will shift up, automation will not be labor-displacing but will just reduce labor share in value added (Autor & Salomons, 2018).

The frontier of technological development is partly driven by the demand side. Organizations set up research and development (R&D) departments, purchase licensed technologies or acquire other enterprises in order to gain access to their technological advances (Goedhuys, Janz, & Mohnen, 2008). As it is in the interest of organizations to minimize labor cost, they will value labor cost saving technology. This valuation will express itself in the willingness to reward those that can bring them the desired technology. In turn this reward will incentivize other R&D firms to create labor cost saving technologies. Not all labor is equal, which means that different types of labor will have different cost. It would be most beneficial to substitute capital with high cost labor. However, as high cost labor generally will consist out of more complex tasks, which will also be more difficult to automate (Brandes & Wattenhofer, 2016). Closer to the present technological state of the world is the automation of routine tasks in a controlled environment. In the past century this effect has already been taking place in the manufacturing industry (Frohm, Lindström, Stahre, & Winroth, 2008). Momentarily, the technological developments in the field of machine learning and artificial intelligence make it viable to use and further develop these technologies in a commercial setting. As cost of this type of capital comes down, routine service occupations could be substituted by it.

In the article ‘The future of employment: How susceptible are jobs to computerisation?’ by Frey and Osborne (2017) the authors concluded that 47% of US employment is at risk due to computerization. The term computerization in this paper is used to describe todays automation wave, the second computerization wave. The first computerization wave covered mostly the digitalization of content, whilst the new wave is mostly processing and responding to this content either through software (AI) or hardware (robots). To compile a list of occupations, Occupational Information Network (O*NET) occupation categories and their task description were used. Those task descriptions served as the basis of a model to compute a corresponding computerization probability of occupation categories at detailed level. With the expertise of machine learning experts, they firsthand-labelled the 70 most evident occupations based on the task description. Using an algorithm, they then calculated the remaining occupations. Therefor they used a Gaussian process classifier, a method which has been developed in offshoring literature.

The result was a list of 702 unique detailed occupations and their corresponding future computerization probability. Information providing service task are mostly at risk while more complex managerial and social tasks are relatively safe. Each of the occupations was also labelled with a unique Standard Occupational Classification System (SOC) code, which made it relatable to the O*NET database and to Bureau of Labor Statistics (BLS) employment and wage data. The overall conclusion of the paper is that a substantial part of US employment is in danger. At the same time also recognizing the effect the efficiency gains might have on future job creation.

I want to build on their research and investigate to what extend we see developments in employment and wage data in conjunction with the probability of computerization of Frey and & Osborne.

Therefore, the main research question of this paper is: How does the predicted probability of computerization relate to developments in employment and wages across jobs?

The probability of computerization in this research question is a variable that represents the likelihood for an occupation of having an alternative for human labor. If this alternative is economically viable that should drive down the price of labor and have a negative effect on the total employment. After all the labor market is a demand and supply market and if demand is affected by automation, while supply stays the same, the equilibrium price and quantity should change. If the probability of computerization is high this signals that the occupation can easily and thus cheaply be taken over by capital. This is not necessarily true as the Frey and Osborne probability is partly a prediction for the future state of the world. This means that it is possible they expect an occupation to be in danger of automation with a lot of certainty, while with current technology it is not possible at all. For example, the probability of computerization different types of drivers is estimated at almost 1, while at this moment it is illegal for autonomous driving vehicles to drive on public roads in most places in the world. However, in general it is to be expected that the probability of computerization has a positive correlation with today’s possibility to automate. Thus, it is to be expected that the probability of computerization has a negative correlation with recent changes in employment and wages.

To answer the main research question, aside from the probability of computerization, data on employment and wage is required. The Frey and Osborne paper focused on the United States (US) labor market. The presence of the SOC code in the list of 702 occupations they provide makes is simple to engage in analyzing the BLS employment and wage data. Furthermore, particularly regarding external validity of this research, also employment and wage data from the United Kingdom (UK) is analyzed. This data is produced by the Office of National Statistics and uses its own classification system code. To facilitate the analysis both classification systems are linked using crosswalks, this does results in an inferior linkage to the probability of computerization variable. The data that is used is originating from recent yearly business surveys, to be able to link the data to the probability of computerization it is not possible to go too far back in time. The list of occupations is constantly updated as occupations appear, disappear and evolve. The choice is made to take 2012 as the first year, as the detailed SOC list from that year is similar to the Frey and Osborne list. The variables used to answer the main research questions are the change in total employment and change in both mean in median wage.

The aim of this paper is to ascertain the predicted probability of computerization is correlating with labor market developments. If this is true that will give some validation to the further predictions that are done. If 47% of the jobs in the labor market are indeed at risk, organizations and their employees should prepare for it. Because automation and education of workers are dependent on each other, workers need to keep up with the current technological progressions. Governments could also play a role here as they usually provide the education that is given to future job seekers. Therefore, it is important to identify the occupations that are demanded in the modern and future economy. This will give insight in the skill and therefore type of education that is required in this economy. To answer the research question while keeping the data sources in mind the following hypothesis were developed:

- The probability of computerization is negatively correlated with recent changes in employment.

- The probability of computerization is negatively correlated with recent changes in wage.

- Low-skilled occupations have a stronger negative correlation with the probability of computerization than high-skilled occupations.

The information in the thesis will be structured in the following way. The first part of the thesis will contain the literature review, which further elaborates on the historic automation waves. Moreover, the current state of the world will be discussed, which will include the current state of technology as well as the current state of the labor market. Thereafter, the research part of the paper is presented, which includes the methods used to estimate the results. Here the data sources will be clarified as well as the statically regression will be explicated and applied. The thesis will then end with a conclusion and some final remarks.

2. Literature review

The automation wave regarding the rise of artificial intelligence, machine learning and in more broad sense computerization is not the first automation wave. Technological progress has played an important role in the changing workplace the past two centuries. This phenomenon is called creative destruction and the principle has arguably been around since the agricultural revolution. Schumpeter defined the term as the innovation that revolutionizes the economic structure from within, while destroying the old way (Reinert & Reinert, 2005). The process takes place because actors have the incentive to embrace the new idea, however this does not always mean the outcome is optimal for total surplus. The current automation wave is feared for the effect it is supposed to have on the labor market (Autor D. , 2014). As with every automation wave it is highly probable that technological unemployment will arise. To predict which occupations will suffer from the completion of robots we have to first consider the current state of the world.

2.1. History of Automation

Technological process has been going on since the begin of humanity. As humans we are capable of using tools in things we do. As time has progressed those tools have been shown to be increasingly helpful. From making fire to cultivate farmland and all the way to using tools to mass produce more tools. If the tools operate in such independent manner that the process is performed with minimal or no humans involved, the process has been automated. Automation in that sense first took off during the industrial revolution.

During the 18th century the textile industry was the leading industry in Great Britain. Cotton was brought in from British colonies and processed in the Lancashire region. At the start of the 19th century with the rise of steam powered machines in the production process, productivity of weavers exploded (Hopkins, 2013). This was the first industrial revolution, which is characterized by the rise of mechanized factories.

A century later the second industrial revolution was powered by electricity, gas and oil. This manufacturing revolution lead to the production of the T-Ford on an assembly line (Mokyr, 1994). Compared to the craftsman in the middle ages, the work on the assembly line was done by low-skilled workers. Rather than one skilled worker being responsible for whole of the production process, multiple workers specialized in small parts of the production process. This specialization turned out to be incredibly efficient and got even more effective with time (Rasmussen, 1982). This was not only positive for the manufacturer but also for the unskilled factory worker, as unskilled blue-collar labor was required by the assembly line (James & Skinner, 1985).

However, as specialization on the assembly line continued to get better, things started to change. In the 20th century the low-skilled blue-collar workers came under pressure, because of electric machines which could replace their highly specialized job task. This was the starting point of the modern positive capital-skill relation, as these machines were operated by high-skilled blue-collar workers. At this point in time education got more important as the average job task was increasingly requiring higher skill level.

In the 19th century establishments already had been getting larger, they had the potential to grow because of the productivity gains. Moreover because of innovations in the transport sector, the market changed from local in the direction of global. These larger and more complicated organizations were in need of high-skilled white-collar worker, for managerial tasks and clerical jobs (Frey & Osborne, 2017).

From 1960 onwards, during the computer revolution, computers were introduced in the workspace, which again changed work and workers. A lot of the clerking jobs which had just been created in de last decades now were substituted with computers (Frey & Osborne, 2017). Education got even more important as more low-skilled white-collar jobs disappeared and high-skilled white-collar jobs were demanded.

All these changes dictated by technological change follow roughly the same path. In the adaption period people fear for their job and indeed jobs are lost. Not everyone profits from this change, but total welfare usually does. Then because of efficiency gains and adaption by the labor force, few desire the past state. Because of this all the literature regarding previous automation is relevant regarding the current one. Where it is very hard to predict the impact of the unknown, the past could give some hints.

2.2. Computerization

At this moment we are in the second computerization wave, jobs are already falling victim to this newest form of automation. Especially routine tasks are easily automated with current technology (Autor & Dorn, 2009). The problem with more complicated non-routine tasks is the very nature of automation. To automate a process, the designer must understand the process very precisely. And it is increasingly becoming clear that humans know more than they can tell. This is called the Polanyi paradox and is a key reason why computers lack the adaptability humans do have (Autor D. , 2014). In today's world the programmer is still the creator and it can only create what is understood.

Because of the rise of processing power and the widespread presence of data, this might change in the future. With artificial intelligence in combination with machine Learning the human creator might be replaced by the computer itself (Brynjolfsson & McAfee, 2014). However, despite early optimistic predictions the technology is not ready yet. The Polanyi paradox is still playing a role here, as computers are missing the baseline a human being has. The process of using machine learning to learn a computer what a cat is turns out to be very hard, while a task like this is easily done by a child. Also, while the first 99% of a process is quite easy to automate, the last 1 percent is the hardest part. This is something autonomous driving car designers are finding out right now (Maurer, Gerdes, Lenz, & Winner, 2016). However, automate all those truckers away you really need to get closer to 100%, as just assisting a driver won't bring close the efficiency gains. Of course, this is not the case in every field in which AI can play a role. In the introduction it was mentioned that AI is already being deployed in the form of chat bots. In contrast to driving a truck if an AI fails to deliver during an interaction with a customer, the customer can easily be connected to a human employee. If an AI fails to drive a truck the driver will not be there any more to save the day. Therefore, the developments in sectors in which imperfectly functioning AI can already be deployed will go faster. This contradiction provides an uncertainty in predicting how exactly this wave will flow over the labor market.

2.3. Predictions for the Labor Market

So still a lot is unclear regarding the development of the new technologies, however predictions are relevant. This is because technological unemployment is a race between technology and education, this has previously been noticed by Jan Tinbergen. As technology gets more advanced, to operate or complement this technology a higher skill level is required for the worker.

In 2017 Frey and Osborne concluded that 47% of current jobs will be lost in the future (Frey & Osborne, 2017). In the paper they come up with a list of 702 detailed occupations of which they have calculated the probability of computerization. They hand-labelled 70 occupations with the help of machine learning experts as either automatable (1) or non-automatable (0). The binary variable was used because only the occupations of which the team was the surest were assessed. They used the O*Net occupation descriptions as the assessment criteria and subsequently picked 9 O*NET variables as classifiers to calculate the probability of the remaining occupations. These 9 variables were subjectively chosen as computerization bottlenecks. If all the 9 variables where absent from the occupation's description, an occupation was fully automatable. The 9 variables where: 'Finger Dexterity', Manual Dexterity', Cramped Work Space', 'Originality', 'Fine Arts', 'Social Perceptiveness', 'Negotiation', 'Persuasion' and 'Assisting and Caring for others'. For the non-labelled occupations, they then used the 70 labelled occupations as a training set for the algorithm. Every labelled occupation corresponds to a vector with those 9 variables in it. As the classifying method they ran three gaussian models: exponentiated quadratic, rational quadratic and linear covariances. To find the best fitting model they randomly selected half of the training set of the 70 hand-labelled occupations and predicted the other half. This enabled them to find the exponentiated quadratic model as the algorithm that showed predictions most closely to their hand-labelled predictions. This model was then used to predict the probability of computerizations for the non-labelled occupations.

There are a few possible problems that can arise with the probability of computerizations using this method. Firstly, a lot is still unknown regarding the capabilities of computers and computer programs in the future. The hand labelling of the 70 occupations is done subjectively, the experts can only predict to the best of their knowledge. Job loss predictions historically tend to overestimate the impact of automation. Additionally, Frey and Osborne assumed in their prediction that if something is possible to automate, it will be automated. However, workers will not be taken over by automation if this will not make economic sense for the firm. Also, the algorithm takes just 9 tasks of the job description into account, while the time spend doing those tasks is neglected.

The predictions of Frey and Osborne have thereafter been tested in other labor markets. In Finland and Norway, a slightly smaller job loss number was found due to difference in composition of the economies (Pajarinen, Rouvinen, & Ekeland, 2015). Also, the methods of Frey and Osborne have been further elaborated with an additional layer. Where Frey and Osborne just use 9 bottleneck tasks, later papers use all the tasks from the O*NET description and also their relative importance in that occupation (Brandes & Wattenhofer, 2016). This later method was also used in an OECD research where lower job loss is predicted in OECD member states than in the Frey and Osborne paper (Nedelkoska & Quintini, 2018).

In this research not only the US labor market is relevant but also the UK labor market will be analyzed. Therefore, it needs to be evaluated if differences between the US and UK labor markets are present. The US and UK labor markets showed similar movements in job creation, declining unemployment and weak wage growth in 2016 (Forbes, 2016). Therefor the prediction to the US labor market are also likely to be true for the UK labor market.

As a similar but somewhat weakened effect as the in the Frey and Osborne result are found in other research papers. This validates creates the expectations that these results will also be true for this research. The external validity that was tested in Finland and Norway suggest that that external validation will be found in the UK labor market. This is because the UK labor market is more similar to the US labor market than those the labor markets of these Scandinavian countries.

2.4. Labor Market Equilibrium

To predict which occupations will be lost to automation, it is also important to take other factors than the state of technology into account. The Neoclassical Growth Model shows that economic growth is not just dependent on the current state of technology but also on other factors like labor and capital (Dimand & Spencer, 2008). Capitalist organization strive to maximize profit and thus grow their economic output while keeping cost to a minimum. To survive in a market-oriented economy, they constantly have to make an assessment what is the ideal mix of labor and capital, considering the current technological state of the world. Capital can both compliment and substitute for labor, in practice capital will fulfill those roles at the same time. It will substitute for low-skilled labor, while at the same time complement high-skilled labor.

The price of labor will be a factor in the investment decision of capital and thus the technological state of the world. When labor prices are high an organization is heavily incentivized to cut labor cost and will therefore invest in capital, those capital investments then drive technological developments. When labor is cheap, the incentive to invest in capital is less present. Therefore, the technological state of the world and the labor market are interconnected. While organizations would really like to substitute capital for their high cost labor this is generally not feasible. The high cost labor is high cost because the skill level required by the tasks of this labor is high. High-skilled task are often more complex tasks, which in turn are harder to automate than simple tasks. Therefore, organizations will substitute the most expensive labor under the condition that it is able to do so. If the investment in capital is made, the efficiency gains will mean that the labor that is complementing the capital will more efficient and valuable. The welfare gains that are made by the organizations will inevitably flow back into the economy. This will happen either through higher wages, for the now more valuable employee, or trough taxes and profit distributions. These efficiency gains that flow back into the economy will shift the economic equilibrium once again (Eden & Gaggl, 2016). So while automation will have a labor displacement effect on the labor market, the productivity effect leads to higher real wages and thus could lead to overall economic growth (Acemoglu & Restrepo, 2018).

Because of automation the labor market is polarizing, this effect has been going on for a while and is getting stronger (Autor & Dorn, 2009). The polarization effect means that one group of workers is separating itself from the other group of workers. In the case of the current labor market this translates into high-skilled workers getting paid more as they become more valuable, while mediocre-skilled workers are getting substituted by AI. This will also have a negative effect for the low-skilled worker, because those now need to rival with mediocre-skilled workers for low-skilled jobs. The demand for labor will become more U-shaped in required skill level. Low and mediocre-skilled workers are largely unable to change to high-skilled occupations, as this requires higher education. Two problems then arise in the absence of higher education. Firstly, society is built on the usual state of affairs that higher educations is obtained during adolescent. It is not easy to obtain higher education later in life, especially not if you need to provide for your family. Secondly, higher education will simply not be obtainable for everyone, as a certain intellectual capacity is a prerequisite. Workers who have been substituted by AI will then reallocate their labor to occupations where their low wage labor is more valuable (Autor D. , 2014). This will lead to extra supply in the labor markets for those occupations, which will lead to lower prices of labor. These low-skilled occupations were already at the bottom end of the income scale. This means that other factor like minimum wages and unemployment benefits start playing a role. If the minimum wage threshold is reached, the price of labor cannot drop any further. This then means that there will be a surplus in supply and thus more unemployment. If the price of labor drops to a level around or under the unemployment benefits, this will mean workers lose the financial incentive to go to work. Because of those last two remarks, in the automation literature universal basic income is frequently mentioned (Arntz, Gregory, & Zierahn, 2017). The advantage of this type of benefit versus a normal unemployment benefit twofold. On one hand the smaller incentive to leave the workforce and on the other hand the simplistic implementation of the distribution.

Zooming in on the situation of an office clerk, which has a probability of computerization of 0.96. The worker has to take on the threat of a computer program, which is ten times more productive then he is. The licensing fee for the computer program may be five times the yearly wage of the office clerk. This will mean the employer will now give the office clerk the choice between taking a 50 percent pay cut or leaving the firm. Finding a job at another firm is also hard because the total demand for office clerks has dropped. If the office clerk, whose education has now become worthless, is not in the position to obtain higher education he will need to find another occupation. However, he will not be the only mediocre-skilled worker changing occupations, which means that the supply of low-skilled occupations has been sky rocking. The upsweep in supply of low-skilled labor, while demand for low-skilled labor stays the same, will then results in a downturn in the price of low-skilled labor. Leaving the worker, who previously had a reasonably paid white-collar job, with a badly paid blue-collar job. Meanwhile the manager of the office clerk has seen his own productivity grow since the introduction of the computer program. This made the firm give him a raise, because managers have become more valuable to firms.

In the research section of this paper it is not possible to study all these different equilibrium effect separates from each other. The labor market equilibrium consists of an endless number of factors that influence the outcome. In that sense the available data is very limited, it just portrays the outcome of the labor market equilibrium. Labor market data consist of data on total employment, mean and median wage and limited basic demographic characteristic of workers. However, together with the likelihood of automation on a detailed occupational level the available data does give the opportunity to study if there is already a correlation between computerization and labor market outcome. It might be possible to recognize some of the effects mentioned in the paragraph above.

3. Research

This section will provide the research part of this paper. The data sources and regression variables are presented. Also, the research question of the paper is elucidated, as well as the logical model of the regressions is shown. Thereafter, the hypothesis is restated and further elaborated on.

3.1. Data Sources

The starting point is the list of occupations with the probability of computerization generated by Frey and Osborne in 2013 (2017). I retrieve this data from the from the data.world website (Kizer, 2020), however, this list contains just 659 of the 702 occupations originally listed by Frey and Osborne. Why this discrepancy exist is not clear, there does not seem to be a pattern in the missing occupations. It might be possible that the list I could retrieve was from an earlier version of the paper.

I then map this data to business survey data from the Bureau of Labor Statistics located in the United States of America, based in Washington. The object of the Occupational Employment Statics (OES) survey are 200,000 non-farming enterprises located in the US. The purpose of the survey is to estimate employment and wage data per detailed occupation, where the occupations are divided based on the North American Classification system and a four-digit Standard Occupational Classification (SOC) code. This SOC code is also used in the paper of Frey and Osborne, therefore it is simple to match the data entries from both data sources.

The second step is enlarging the dataset with occupational data from the United Kingdom. The purpose of the diversification of the data is to assess if the predictions of Frey and Osborne are valid in other labor markets. The UK data is from the Office of National Statistics (ONS). Like the OES data it is from a business survey the Annual Survey of Hours and Earnings (ASHE). However, the ASHE data is not of equal quality compared to the OES data. Firstly, the data entries of the total employment numbers are indicated as not being reliable enough for research purposes by the ONS. Secondly far more individual observations are missing in the UK data. Sometimes a single data entry is missing regarding one year and one occupation. Other times all the years of a variable of a specific occupation is missing. A similar four-digit SOC system is in place to distinguish individual occupations, however, the system differs from that from the US. To link the Frey and Osborne data, a crosswalk between both SOC system was used. Because no direct crosswalk was available, a translation trough the ISCO system had to be used. The ISCO system is the International Standard Classification of Occupations System, for which crosswalks to most national systems exist. As the occupational classification systems all have some issues in linking the systems together this does come with a cost. These issues consist of the absence of occupations from the list or occupations that are defined slightly different the different systems. The indirect crosswalk through the ISCO means that first the UK data had to be translated to ISCO codes, which then had to be linked to the US SOC system. As in every step data is lost, the UK list, which originally consisted of 498 occupations, after the translation loses around 150 observations. Also, the as the UK data was thus more incomplete than the US data, I chose to keep the observations which were not complete. Therefore, in the regressions later in this paper the number of observations differs far more per dependent variable than in the US data.

The website of the BLS includes a list of the educational attainment per detailed occupation. The educational attainment values are divided into six categories Most of the occupations from the Frey and Osborne probability of computerization are represented on this list. The year of publishing is 2018 and originating from the US data, however it will also be used for the other years and the UK data. I do not have a reason to believe that the educational attainment of employees will have drastically shifted in the five-year period or will be that different for the UK data set. Moreover, I will use a simplified fraction of high and low education in the analysis part of this paper, this will decrease the likelihood of imprecise measures biasing the results.

The data from all five different sources is merged together in one Excel file. Using the VLOOKUP function in Excel in combination with the SOC coding used in all the sources it was possible to create two all-encompassing Excel sheets. The original files existed of individual sheets per year, with redundant variables. Therefore, Total US and Total UK sheets were created containing the employment, mean wage and median wage variable of the years 2012 to 2018. This made possible to calculate and add the variables which represent the difference between the years so that the development could be tested.

A point of attention in this dataset is the economic cycle that is affecting the data sample. After the economic banking crisis of 2008, the world economy has been recovering. However, it is not to be expected that this recovery is affecting all the occupations in the dataset equally. This is because occupations that require higher education tend to be relatively safe. Because the probability of computerization is negatively correlated with the educational attainment this could bias the effect. As there was no way to directly control for the economic cycle effect per occupation a second sample was made. Rather than consisting of the ratio of the years 2012 until 2018 this sample consist of the years 2015 and 2018. As this is further from the initial years of economy recovery this should soften the bias.

3.2. Key Variables and Summary Statistics

The goal of this paper is to research if some of the effects of the current automation wave, that is flowing over the economy, translate into trends in labor markets. If the effects in the labor markets are presented, this should show itself in the outcome of the labor market equilibrium. The outcome of the labor market equilibrium is imprinted in employment and wage data. To capture this effect, it is thus needed to create variables that reflect employment and wage.

3.2.1. Dependent Variable

To answer the hypothesis multiple regressions are run, which is required because in this research as there is more than one dependent variable. The hypothesis includes two outcome effects the automation wave is supposed to have. On one hand the outcome on the employment and on the other hand the outcome on wage. Both outcomes are results of the labor market equilibrium.

Also, per regression two time periods are used, the main purpose of using two data entry is trying to combat the business cycle biasing the results. As the world economy was recovering from the banking crisis in 2008, this will have affected the data entries in 2012. Namely low-skilled occupations tend to be victims in an economic crisis. Thus the 2012 could have been biased downwards by the crisis. Furthermore, regressions are run on both labor markets, the US and UK labor markets.

As a measure of employment, the variable ‘total employment’ is chosen, as this is the only employment variable. However, the hypothesis talks about change in recent employment and two ‘total employment’ data entries are needed to be able to calculate a change. Because ratios have no negative outcome, I chose to work with them rather than with first differences. If outcomes could be negative this limits the mathematical possibilities, for instance taking the logarithms of the variables. This way the variable ‘Total Employment Ratio’ is created. The first measurement point taken is the 2012 data entry, as mentioned this is the earliest viable option. The most measurement point is the 2018 data entry, simply because this is the most recent data entry. Moreover, the data entry of 2015 is used is a separate data sample. Total Employment Ratio is calculated as the ratio of the two measurement periods. For the 2012 US data sample this means that the value is equal to the 2018 value divided by the 2012 value.

To test the hypothesis on the recent changes in wage, the same data entries are used as in the regressions on Total Employment Ratio. However, for the Wage changes there two variables are created. While there were multiple wage representing variables to choose from, I chose a median wage variable and a mean wage variable. In the US data sample Median Wage Ratio is the ratio of ‘Annual Median Wage’, while Mean Wage Ratio is the ratio of ‘Hourly Mean Wage’ of those years. For the UK data sample both mean and median wage are values of annual wage. The advantage of taken both values for mean and median wage is the fact that one could be skewed due to large income disparities. These variables will thus capture the change in wages over the designated time period.

In total this thesis will have 10 different dependent variables the two US data samples have 3 dependent variables, while the two UK data samples have 2 dependent variables each.

3.2.2. Main Independent Variable

The main independent variable is the Probability of Computerization that is calculated in the paper of Frey and Osborne. The probability is distributed between 0.03 and 0.99 with a mean of 0.56, median of 0.60 and a standard deviation of 0.36. As previously mentioned, the variable was constructed using a Gaussian process classifier that was trained using subjective evaluation of machine learning experts. Notice the U-shaped level in figure 1, this means that extreme probabilities are frequently present.

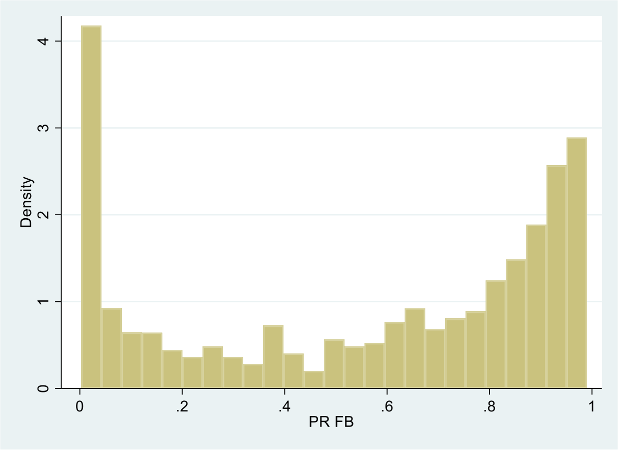


Figure 1: Histogram of the Probability of Computerization density

3.2.3. Control Variable

Education is the single control variable which is occupation specific. In the original data file that gives the educational attainment values there are six categories: ‘Less than high school diploma’, ‘High school diploma or equivalent’, ‘Some college, no degree’, ‘Associate’s degree’, ‘Bachelor’s degree’, ‘Master’s degree’ and ‘Doctoral or professional degree’. The variable is defined by dividing the percentage of workers with ‘bachelor’s degree’ or higher by the total number of workers. It expresses itself in a value between 0 and 1 with a mean of 0.34, a median of 0.20 and a standard deviation of 0.31.

3.3. Descriptive Statistics

Table 1 provides descriptive statistics of the variables in the US 2012 data sample. In this data sample there are 631 observations, except for the mean which has a few observations missing. The Probability of Computerization is distributed between 0.03 and 0.99 and has a mean of 0.559. Education is distributed between 0.015 and 1 and has a mean of 0.338. The ratios are calculated with the observation of 2018 as the divided and 2012 as divisor. This simple calculation will give immediately the percentage change, which is preferred while working with the data.

Table 2 provides the summary of the variables in the US 2015 data sample. Comparing the means with the US 2012 dataset the values of the total employment ratio and median wage ratio are lower. This means that the difference between the measuring points is smaller. Table 4 and 5 show the results for the UK dataset. The number of observations in the UK dataset differs more per variable than the US dataset.

This is the case because the UK dataset was less complete and therefore it was not viable to delete all the observations which had missing observations.

Table 1
Descriptive statistics US 2012

	Mean	Min	Max	StdDev	Observations
Probability computerization	0.559	0.003	0.990	0.362	631
Log probability computerization	-0.527	-2.553	-0.004	0.687	631
Education	0.338	0.150	1.000	0.313	631
Total Employment ratio	1.054	0.194	2.245	0.231	631
Median wage ratio	1.131	0.867	1.844	0.076	631
Mean wage ratio	1.132	0.912	1.657	0.064	631

Table 2
Descriptive statistics US 2015

	Mean	Min	Max	StdDev	Observations
Probability computerization	0.559	0.003	0.990	0.362	631
Log probability computerization	-0.527	-2.553	-0.004	0.687	631
Education	0.338	0.150	1.000	0.313	631
Total Employment ratio	1.010	0.163	2.097	0.162	631
Median wage ratio	1.076	0.844	1.587	0.066	631
Mean wage ratio	1.070	0.848	1.435	0.051	615

Table 3
Descriptive statistics UK 2012

	Mean	Min	Max	StdDev	Observations
Probability computerization	0.570	0.358	0.003	0.990	345
Log probability computerization	-0.511	0.684	-2.523	-0.004	345
Education	0.337	0.328	0.015	1.000	344
Median wage ratio	1.087	0.123	0.000	1.408	297
Mean wage ratio	1.089	0.095	0.810	1.460	335

Table 4
Descriptive statistics UK 2015

	Mean	Min	Max	StdDev	Observations
Probability computerization	0.576	0.003	0.990	0.356	351
Log probability computerization	-0.505	-2.523	-0.004	0.689	351
Education	0.328	0.015	1.000	0.326	351
Median wage ratio	1.027	0.000	1.325	0.168	317
Mean wage ratio	1.031	0.000	1.242	0.156	339

Figure 2 shows the distribution of the median ratio in the US 2012 data sample, it roughly follows a normal distribution. Moreover, it shows that most occupations have experienced growth in median wage. Figure 3 shows the relation between the Probability of Computerization and Median Wage Ratio in the US 2012 data sample. Eyeballing this graph does not show a clear linear or exponential relationship between both variables. Figure 4 shows the distribution of the Total Employment Ratio, which also follows a normal distribution. Notable that almost half of the occupations had a decline in total employment. Moreover, figure 5 shows the relation between the Total Employment Ratio and the Probability of Computerization. The observations are more spread out than in figure 3, however similarly no clear relation can be observed.

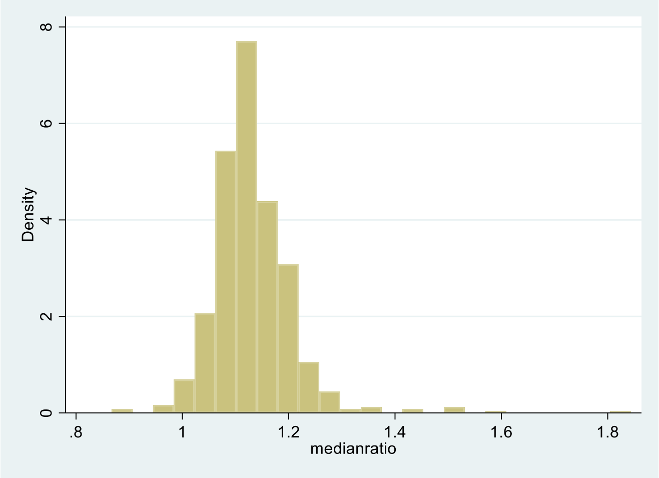


Figure 2: Histogram of the Median Wage Ratio density in the US 2012 dataset

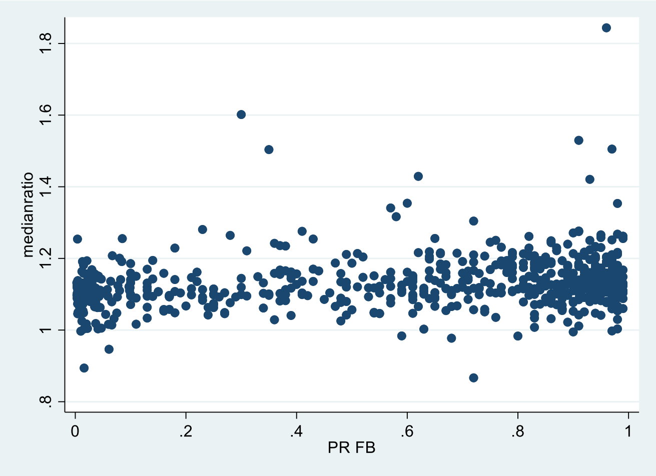


Figure 3: Scatterplot of the Median Wage Ratio and Probability of Computerization in the US 2012 dataset

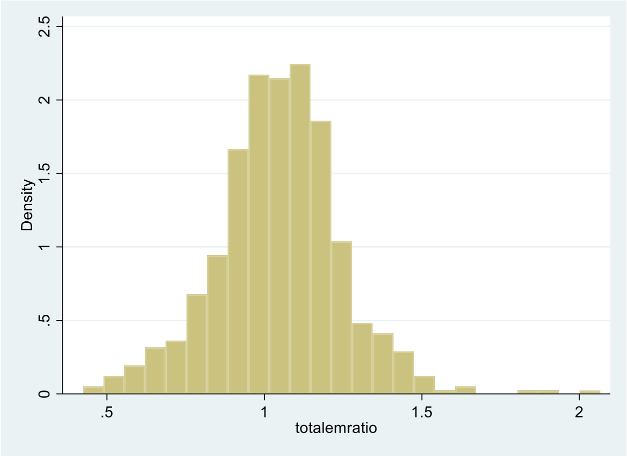


Figure 4: Histogram of the Total Employment Ratio density in the US 2012 dataset

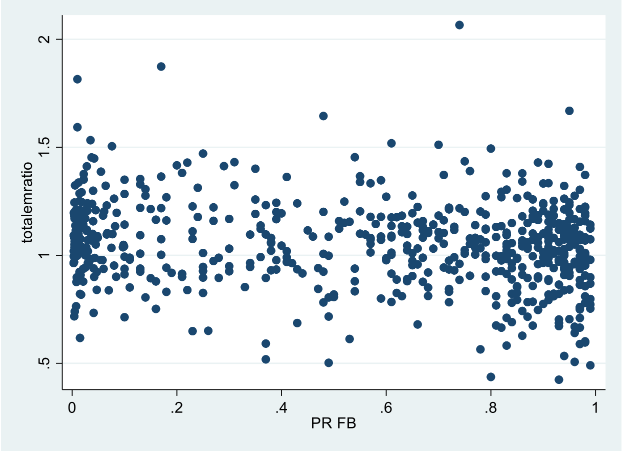


Figure 5: Scatterplot of the Total Employment Ratio and Probability of Computerization in the US 2012 dataset

3.4. Research Question

In the literature review two effects of automation turned out to be present. The labor displacing effect and the productivity effect. In an ideal experiment there would be data available to expose those effects individually. However, due to the nature of the data this is not possible, the only data available is the outcome of the labor market equilibrium. It is possible to use the probability of computerization and check for correlation with trends in that data, however no causal claims can be made from this research.

The research question in this thesis is as follows:

How does the predicted probability of computerization relate to developments in employment and wages across jobs?

A regression analysis is a statistical method which is used to uncover a possible correlation between a dependent and one or multiple independent variables. As described above there are 4 different data samples, US 2012, UK 2012, US 2015 and UK 2015. Moreover, there are multiple variables that will be subject of the regression. Because of the nature of the data all these different data samples need their own regression function.

The variables present in this question have already been outlined in the previous chapter. However, to research a topic also a regression model is needed. The regression method used in this thesis will be an ordinary least squares (OLS) regression. This regression method estimates unknown parameters using the outcome that minimizes the sum of squares between observed and predicted dependent variables.

The first OLS regression used to estimate the β_1, where pr_i is the Probability of Computerization of occupation i and X_i is a vector of control variables. The dependent variable $wageratio_i$ is the wage ratio between year '2012' or '2015' and '2018'. The logical regression equation 1 is used to estimate the wage ratio for both the US and UK labor market.

$$wageratio_i = \beta_0 + \beta_1 pr_i + \ \beta_2 X_i + \varepsilon_i \qquad (1)$$

In a similar manner the expected ratio in total employment in the US data can be estimated using the equation 2:

$$Totalemploymentratio_i = \beta_0 + \beta_1 pr_i + \ \beta_2 X_i + \varepsilon_i \qquad (2)$$

3.5. Hypotheses

In the literature review it is argued that if an occupation has a high probability of computerization this will have a negative influence on wage. This is due to the outcome of the labor displacing effect and the productivity effect. This alternative option of computerization this leads to a deteriorated competitive position of the worker. In this research it will not be possible to filter out all the equilibrium effects independently, only the labor market outcome is observed.

The labor market is a market of supply and demand, thus changes in supply and demand should be visible in equilibrium outcomes. If demand for a certain type of labor declines, while demand stays as it is the price of labor should also decline. This should then mean the change in employment becomes is skewed downwards in the data set. Also, the change in wage should be influenced in a negative manner by the market developments. A remark must be placed here that the labor markets in the US and UK are not perfect markets. As employees and firms are subjected to contracts the changes in value will not immediately translate into job loss and lower wage. However, on a longer time period when new agreements are made, this should be getting visible.

The probability of computerization is not only estimated on the current state of the world but mostly on what is likely to happen in the future. Despite that fact, in this thesis I already try to see if labor market data shows trend that indicates the prediction to be legitimate. The trend that should be present in the data is the negative correlation between probability of computerization and recent employment and wage changes.

In accordance with this expectation a hypothesis is formed.

Hypothesis 1A: The probability of computerization is negatively correlated with wage.

Additionally, the alternative option for firms to invest in capital rather than in labor is predicted to result fewer job openings. This will eventually lead to lower total employment numbers as jobs of human being are taken over by computers. Therefore, a second hypothesis is composed:

Hypothesis 1B: The probability of computerization is negatively correlated with total employment.

As stated previously in the thesis education and technological developments are connected to each other, the relation is especially important in in the labor market. The second computerization wave is expected to take out mostly low-skilled service jobs. In the labor market equilibrium chapter these types of jobs were referred to as mediocre-skilled jobs.

Because of this in analyzes the relation of education and the probability of computerization is also considered. It is to be expected that low-skilled occupations will be the main victim of this automation wave. Therefore, the following hypothesis is introduced:

Hypothesis 2: Low-skilled occupations have a stronger negative correlation with the Probability of Computerization than High-skilled occupations.

4. Results

In this chapter the Regression results for all the data samples will be discussed. Thereafter, some robustness checks will be done ensuring the validity of the chosen regression model. Thenceforth an extension, that uses major occupations groups rather than detailed occupations, will be discussed.

4.1. Regressions

For the US data samples six regressions are run per sample. The dependent variables are the employment, median wage and mean wage ratios. The main independent variable is the Probability of Computerization with Education as a control variable. For every different dependent variable two regressions are run, one with and one without an interaction term between the Probability of Computerization and Education. The first regression of the dependent variable is the regression without the regression term, while the regression on the right is the one with the interaction term.

4.1.1. US

Table 5 shows the results of the regressions with robust standard errors of the US 2012 dataset. The R-squared for all the regressions is below 10%, this low value means that the regression can only explain a small part of the variance in the observations. Although a regression with a higher R-squared is preferred, the low R-squared is not necessarily bad. As the data provided limited variables to control with the low R-squared should be expected.

The Probability of Computerization in this dataset gives a highly significant coefficient for every regression that is run. The coefficient for the Total Employment regressions shows a negative relation between the Probability of Computerization and the Total Employment Ratio. This relates to a higher probability to predicts a lower value of Total Employment relative to the measuring point in 2012. To be exact a 1 percentage point increase in the Probability of Computerization, all else equal, decreases the expected Total Employment Ratio with $\frac{1}{100}$ of the regression coefficient.

In the first regression this relates to a 1 percentage point increase in the Probability of Computerization, ceteris paribus (CET PAR), decreases the expected Total Employment Ratio with $\frac{1}{100} * 0.142$, which in turn relates to the expected in 2018 total employment decreases with 0.142% relative to 2012 . The coefficient of 0.142 is not only statistically significant, it is also economically significant. The coefficient expresses that an occupation that has a Probability of Computerization of 1, has a predicted value of total employment 2018 that is 14.2% lower relative to an equal occupation with a Probability of Computerization of 0. 14.2% is quite a large number, we are predicting averages in a macroeconomic setting. The sign of the coefficient is negative which relates to a negative correlation between the Probability of Computerization and recent changes in total employment, as was expected in hypothesis 1b.

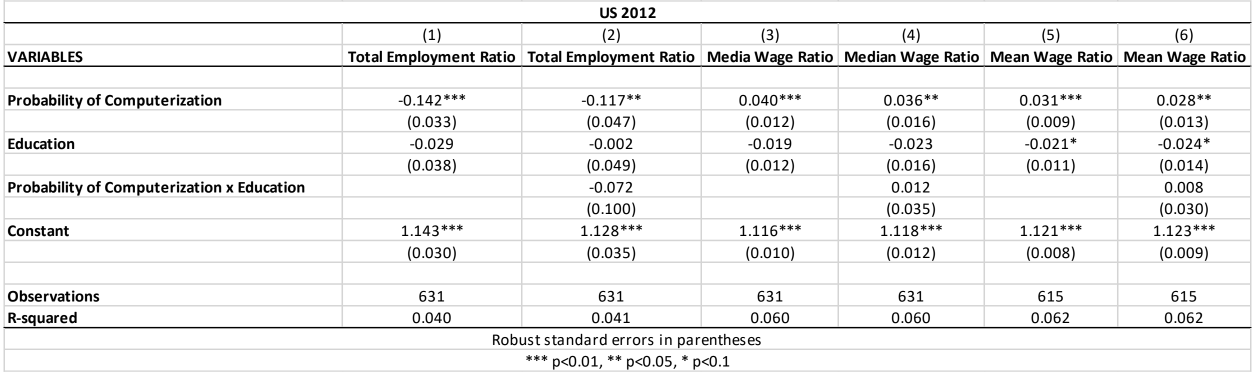
The second regression with the interaction term has a slightly lower coefficient for the Probability of Computerization. The coefficient has a value of -0.117, the difference can be explained by the difference in interpretations of the coefficients. Namely the coefficient for the Probability of Computerization now estimates the expected value under the condition that education is 0. The interaction term has no significant coefficient.

The wage regressions, also highly significant in the Probability of Computerization coefficient, show counter to expectations a positive relation between Probability of Computerization and the change in wages over that time period. In hypothesis 2, the expectations were expressed that low-skilled labor would be stronger correlated to the Probability of Computerization than high-skilled labor. From this estimate we read the opposite effect is present. The first regression coefficient is a value for average educational attainment, while the second regression the educational attainment is 0. Based on the predictions the second coefficient should thus be higher, which it is not.

In the third regression a 1 percentage point increase in the Probability of Computerization means Median Wage Ratio rises 0.00040, CET PAR, which means that expected median wage rises with 0.04%.

The fourth regression with the interaction term has a slightly lower coefficient for the Probability of Computerization, this is again contrary to the prediction of hypothesis 2.

The fifth and sixth regressions on Mean Wage Ratio regressions show similar but somewhat lower coefficients of -0.031 and -0.028 compared to the regressions on median wage. The Education variable in these regressions have a negative coefficient meaning that a higher average educational attainment level predicts lower values for employment and wage ratios. A 1 percentage point increase in the share of higher education in an occupation has a negative influence of 0.0021 on the expected Mean Wage Ratio if all else is equal. This relates occupations that have a large high education fraction, according to the data if the share of high education becomes larger this negatively correlates to the wage growth in the 2012 until 2018 period. For an occupation that has a 10% larger share of high-skilled workers than an otherwise equal occupations the expected difference in wage is negative 2,1%. The sign is in line with the previous stated expectations that low-skilled labor was at a low level in 2012 due to the 2008 banking crisis.



US 2012						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	**Total Employment Ratio**	**Total Employment Ratio**	**Media Wage Ratio**	**Median Wage Ratio**	**Mean Wage Ratio**	**Mean Wage Ratio**
Probability of Computerization	-0.142***	-0.117**	0.040***	0.036**	0.031***	0.028**
	(0.033)	(0.047)	(0.012)	(0.016)	(0.009)	(0.013)
Education	-0.029	-0.002	-0.019	-0.023	-0.021*	-0.024*
	(0.038)	(0.049)	(0.012)	(0.016)	(0.011)	(0.014)
Probability of Computerization x Education		-0.072		0.012		0.008
		(0.100)		(0.035)		(0.030)
Constant	1.143***	1.128***	1.116***	1.118***	1.121***	1.123***
	(0.030)	(0.035)	(0.010)	(0.012)	(0.008)	(0.009)
Observations	631	631	631	631	615	615
R-squared	0.040	0.041	0.060	0.060	0.062	0.062
Robust standard errors in parentheses						
*** p<0.01, ** p<0.05, * p<0.1						

Table 5

In the US 2015 data sample the same pattern is found as in the 2012 dataset as seen in table 6. Thus, a negative relation between Probability of Education and employment ratios and a positive relationship between the Probability of Education and wage ratios.

The coefficients for the Probability of Computerization are significant in all the regressions. For the first two regressions on Total Employment Ratio the value of the regression coefficient is -0.106 for the regression without interaction term and -0.094 for the regression with the interaction term. Although the difference is smaller the same unexpected coherence is found as in the US 2012 data sample.

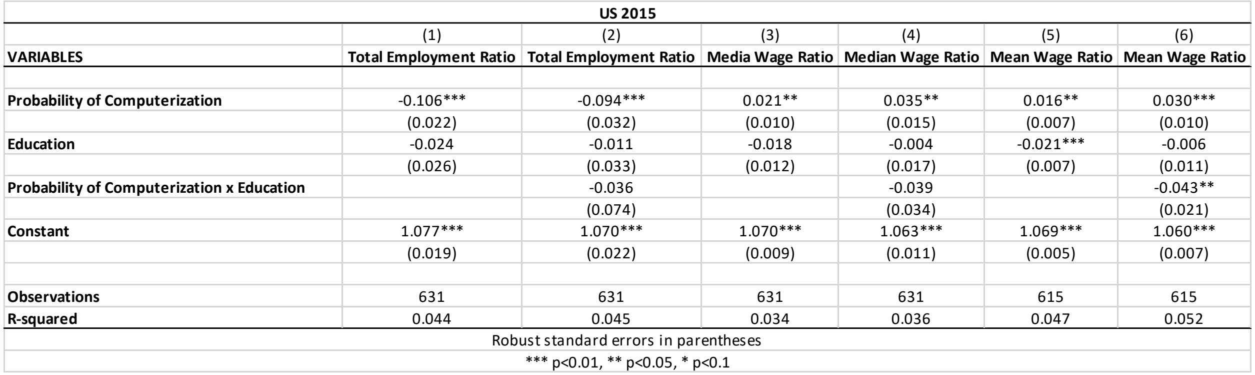
In comparison to the US 2012 dataset the coefficients are slightly lower in this sample. This is to be expected as the difference in time between the two measuring is now three rather than six years. Therefore, the absolute difference between the years will be greater, as a somewhat similar year to year difference would make the most sense. The higher absolute value of the difference between the years will then result in a higher ratio.

In the third regression the coefficient for the regressions without interaction term is also a bit lower than in the US 2012 data sample at 0.021. What is also notable in contrast to the US 2012 data sample is that the coefficient of the fourth regression with interaction term is 0.035, it is higher than the one mentioned previously. Occupations where the Education variable is thus 0, have a higher expected value in this data sample. This is completely different results than was expected, especially in the 2015 data sample. Low-skilled occupations were expected to be negatively correlated with the Probability of Education.

In the fifth and sixth regression the same trend is present regarding the coherence of the Probability of Computerization coefficients. Against expectation the coefficient of the regression with interaction term is higher than the coefficient of the regression without interaction term.

Again, in this data sample, the Education variable is only significant in regression on the Mean Wage Ratio. The coefficient of the regression without the interaction effect is -0.0021, the same as it was in the US 2012 data sample. The fact that the coefficient is equal while the period between measurements was shorter does not comply with the theory that low-skilled occupation experienced more wage growth in the aftermath of the 2008 banking crisis.

In the sixth regression the interaction term in this data sample is significant for the Mean Wage Ratio regression. The coefficient is -0.043, which implies that a 1% increase in the Probability of Computerization, while the Education variable has value 1, has a negative effect of $0.03 - 0.043 = -0.013$ the expected Mean Wage Ratio. As this only is relevant in full unit increases of the variables in Figure 6 this interaction term is dissected for more levels of Education (levels go from 0 to 1 with increments of 0.1). What stands out in this figure is the changing of the slope from positive for low values of Education to negative for high values of Education. This implies that for low-skilled work an increase in the Probability of Computerization has a positive effect on mean wage, while for high-skilled work the opposite is true. This in line with the inference we had from comparing the regression with and without the interaction term and is goes completely against that which we had expected.



	US 2015					
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	**Total Employment Ratio**	**Total Employment Ratio**	**Media Wage Ratio**	**Median Wage Ratio**	**Mean Wage Ratio**	**Mean Wage Ratio**
Probability of Computerization	-0.106***	-0.094***	0.021**	0.035**	0.016**	0.030***
	(0.022)	(0.032)	(0.010)	(0.015)	(0.007)	(0.010)
Education	-0.024	-0.011	-0.018	-0.004	-0.021***	-0.006
	(0.026)	(0.033)	(0.012)	(0.017)	(0.007)	(0.011)
Probability of Computerization x Education		-0.036		-0.039		-0.043**
		(0.074)		(0.034)		(0.021)
Constant	1.077***	1.070***	1.070***	1.063***	1.069***	1.060***
	(0.019)	(0.022)	(0.009)	(0.011)	(0.005)	(0.007)
Observations	631	631	631	631	615	615
R-squared	0.044	0.045	0.034	0.036	0.047	0.052

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 6

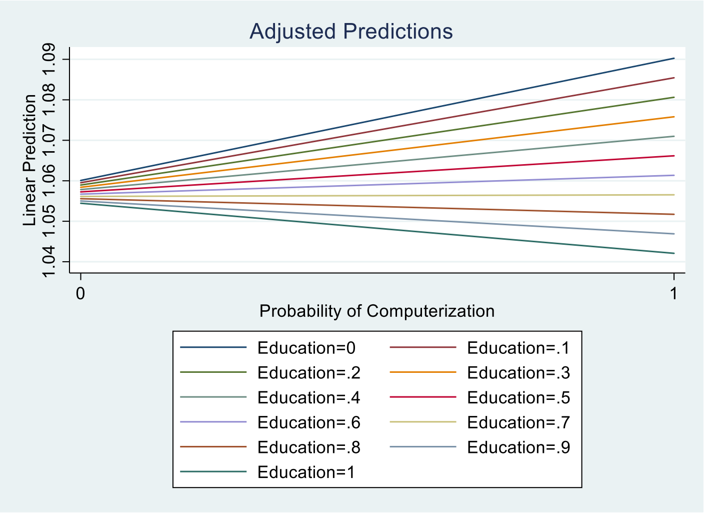


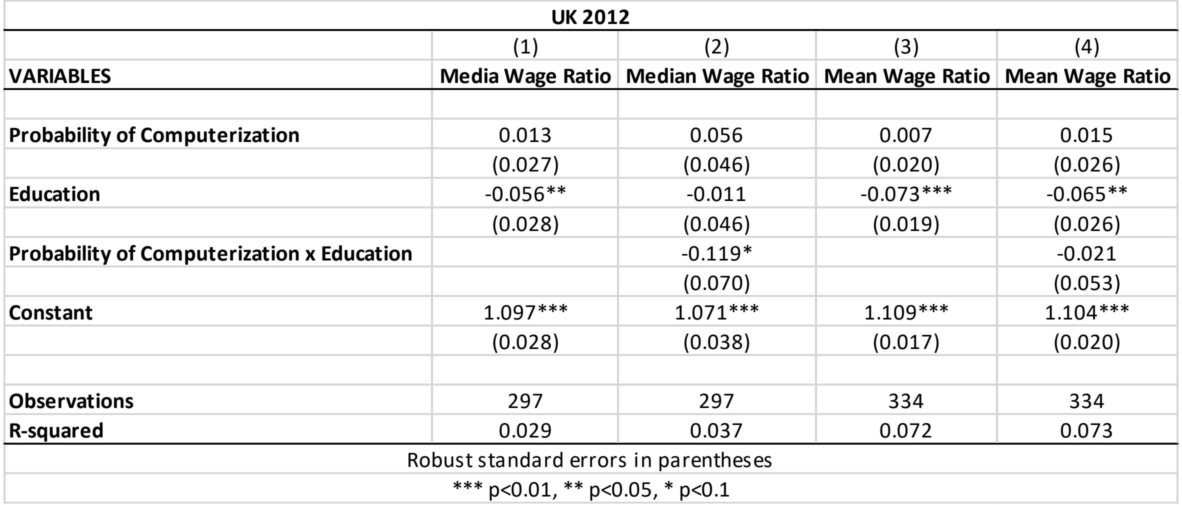
Figure 6: Slope of the interaction term for different levels of Education

4.1.2. UK

In tables 7 and 8 the results of the regression with robust standard errors of the UK data samples are shown. Because the employment data was not reliable enough for research purposes, these regressions are not present. In contrast to the US regression the coefficients for the Probability of Computerization in the UK sample are not significant. This is striking as in the US data sample all the coefficients were significant.

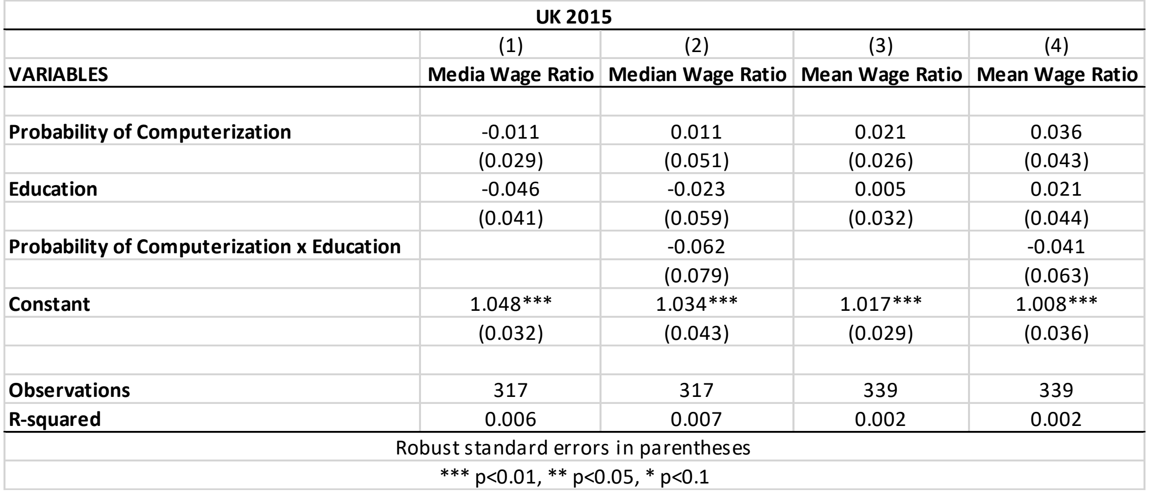
The coefficients for Education are significant and show a negative relation with the wage ratios. Notable in these regressions is that the Education variable is more significant than in the US data samples. This could hint that the crosswalk translation worked out better than we would expect based on the significance level of the other variables.

Furthermore, in the UK 2012 data sample the interaction term is significant in the Median Wage Ratio regression, the coefficient is -0.062. This is striking as it is the same effect that was found in the US data samples that were contrary to our expectations. If this effect is even present as the only effect in the UK data sample this is really under scribing the validity of the results.



UK 2012				
	(1)	(2)	(3)	(4)
VARIABLES	**Media Wage Ratio**	**Median Wage Ratio**	**Mean Wage Ratio**	**Mean Wage Ratio**
Probability of Computerization	0.013	0.056	0.007	0.015
	(0.027)	(0.046)	(0.020)	(0.026)
Education	-0.056**	-0.011	-0.073***	-0.065**
	(0.028)	(0.046)	(0.019)	(0.026)
Probability of Computerization x Education		-0.119*		-0.021
		(0.070)		(0.053)
Constant	1.097***	1.071***	1.109***	1.104***
	(0.028)	(0.038)	(0.017)	(0.020)
Observations	297	297	334	334
R-squared	0.029	0.037	0.072	0.073
Robust standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Table 7



UK 2015				
	(1)	(2)	(3)	(4)
VARIABLES	**Media Wage Ratio**	**Median Wage Ratio**	**Mean Wage Ratio**	**Mean Wage Ratio**
Probability of Computerization	-0.011	0.011	0.021	0.036
	(0.029)	(0.051)	(0.026)	(0.043)
Education	-0.046	-0.023	0.005	0.021
	(0.041)	(0.059)	(0.032)	(0.044)
Probability of Computerization x Education		-0.062		-0.041
		(0.079)		(0.063)
Constant	1.048***	1.034***	1.017***	1.008***
	(0.032)	(0.043)	(0.029)	(0.036)
Observations	317	317	339	339
R-squared	0.006	0.007	0.002	0.002
Robust standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Table 8

4.2. Robustness Checks

The first assumptions of OLS regression is the linearity of coefficients and the error term. To control if linear regression fits the data the residuals (error) term of the regression displayed below. The Residuals should follow a normal distribution, the normal distribution line is portrayed in figures 7 and 9. The residuals for both the regression with and without seem to follow this normal distribution. The scatterplots of the residuals need to show a box display and should lack any pattern. The residuals do show this box display, however there are some outliers present.

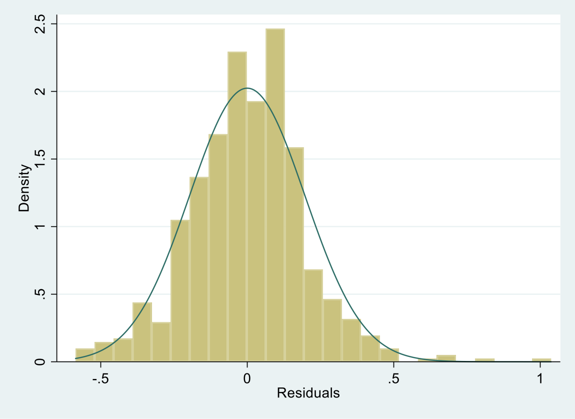


Figure 7: Distribution of residuals for the regression without interaction term

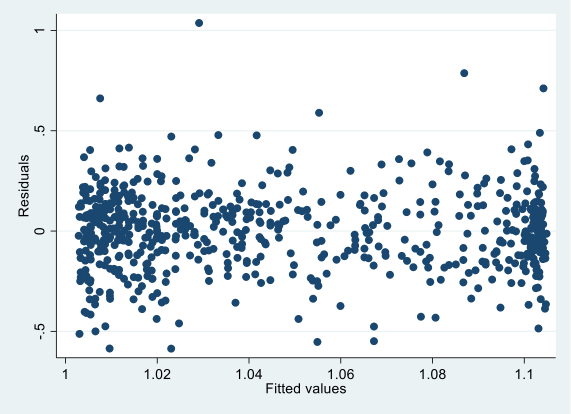


Figure 8: Scatterplot of residuals for the regression without interaction term

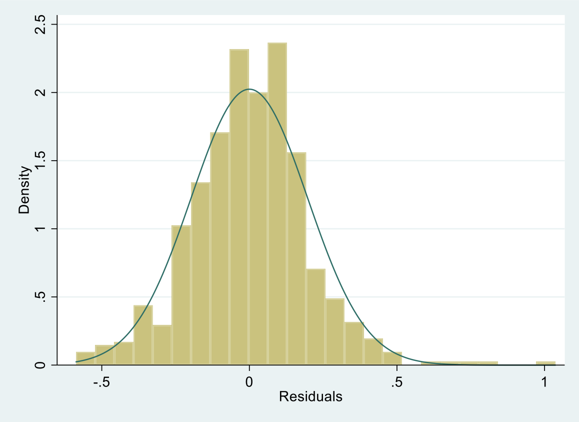


Figure 9: Distribution of residuals for the regression with interaction term

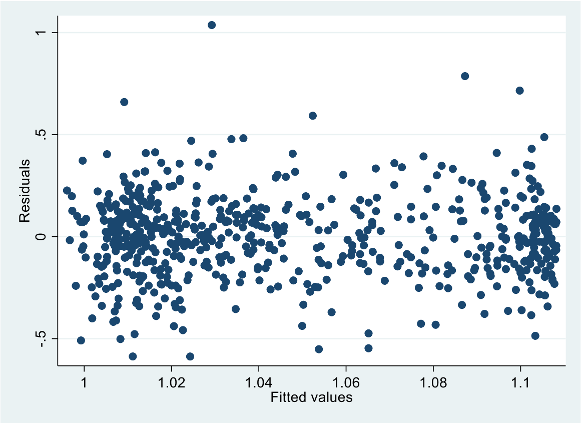


Figure 10: Scatterplot Distribution of residuals for the regression with interaction term

What figures 2 and 3 failed to show were clear cut linear or non-linear relations between the Probability of Computerization and the dependent variables. As described above the residuals of the linear regression are not very alarming. However, to make sure that there is not a very strong non-linear relation between the Probability of Computerization and the dependent variables, all the regressions are run again with the Log of that probability. This enables comparing the R-squared values of the regression directly. From table 9 no very strong non-linear relation exist, linear regression seems to outperform the non-linear regressions almost every time.

What is also clear from this table is the limited effect of the added 2015 data sets. The idea behind those was to limit the effects of the rebound from the 2008 financial crisis in the regressions. However, the R-squared and thus the fit of the regression model on the data does not seem superior. From the results in the last chapter is was already clear that the data did not show the trends that were expected.

R-squared						
	Total Employment Ratio	**Total Employment Ratio**	**Median Wage Ratio**	**Median Wage Ratio**	**Mean Wage Ratio**	**Mean Wage Ratio**
US 2012	0.040	0.041	0.060	0.060	0.062	0.062
US 2012 Log	0.030	0.031	0.058	0.058	0.058	0.058
US 2015	0.044	0.045	0.034	0.036	0.047	0.052
US 2015 Log	0.038	0.040	0.034	0.037	0.043	0.051
UK 2012			0.029	0.037	0.072	0.073
UK 2012 Log			0.029	0.034	0.072	0.072
UK 2015			0.006	0.007	0.002	0.002
UK 2015 log			0.009	0.009	0.001	0.001

Table 9: Values of the R-squared for all the regressions

4.3. Extension

The quality of the previous data samples is the detailed occupational level data. A downside to this is the lack of control variables and thus a not so great overall fit of the model. An advantage of the data being linked through the SOC is the possibility to go back a level from detailed to major occupational groups. This gives the opportunity to add demographic characteristics to the regression, as this data is available for the larger major groups.

4.3.1. Descriptive Statistics

The major groups in the SOC system consist of 23 different groups, from which “Military Specific Occupations” is disregarded due to lack of data. This leaves 22 observations in the US 2012 data sample, the only sample that will be considered. In the appendix the whole list is provided, this includes an average probability of computerization per major occupational group. The 22 observations are immediately the biggest drawback of using the major groups, the limited observations translate into a lack of statistical power. The control variables that are added are: “% women”, “%White” and multiple age categories which are displayed in table 10, those control variables were not available in for the detailed list of occupations.

Variable	Obs	Mean	Std.Dev.	Min	Max
Probability of Computerization	22	.482	.266	.048	.838
Log Probability of Computerization	22	-.419	.352	-1.322	-.077
Education	22	.408	.279	.059	.83
Median Age	22	41.773	3.372	29.9	47
% Women	22	44.936	24.422	3.5	86.9
% White	22	77.345	6.377	64.2	89.3
% 16-19 years	22	3.004	3.634	.205	16.46
% 20-24 years	22	8.942	4.102	2.762	21.616
% 25-34 years	22	23.736	3.207	18.183	29.765
% 35-44 years	22	21.294	2.895	14.192	26.42
% 45-54 years	22	19.92	2.583	12.258	24.92
% 55-64 years	22	16.571	2.827	9.441	21.196
% 65> years	22	6.53	2.21	3.232	12.839
Total Employment Ratio	22	1.072	.074	.906	1.248
Median Wage Ratio	22	1.131	.03	1.092	1.214
Mean Wage Ratio	22	1.132	.027	1.099	1.206

Table 10: Descriptive Statistics of the Major Groups data sample

4.3.2. Regression

In table 11 the results of the regression for the major group data set are portrayed. From the age variables the 16-19 age groups are left out as a separate explanatory variable. The first and second regression in the table do not show any interesting significant results. However, the third and fourth regression do have significant coefficients for the Probability of Computerization. The coefficients have a similar economic relevance as in the previous regressions. Moreover, the sign is also positive which is in line with the previous found results but against the predictions of hypothesis 1a.

The sixth regression on wage has a significant coefficient for the Probability of computerization with positive sign, while the fifth regression does not. In the sixth regression also the interaction term has a statistically significant coefficient. As with the previous regression the sign is negative, suggesting an unexpected stronger negative correlation if education level is high.

US 2012						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	**Total Employment Ratio**	**Total Employment Ratio**	**Median Wage Ratio**	**Median Wage Ratio**	**Mean Wage Ratio**	**Mean Wage Ratio**
Probability of Computerization	-0.040	-0.227	0.038**	0.117**	0.020	0.093**
	(0.168)	(0.262)	(0.015)	(0.043)	(0.013)	(0.036)
Education	-0.140	-0.311	-0.051	0.021	-0.055	0.010
	(0.113)	(0.232)	(0.033)	(0.036)	(0.031)	(0.035)
Probability of Computerization x Education		0.452		-0.191		-0.174*
		(0.522)		(0.106)		(0.083)
Women	0.000	0.000	-0.000	-0.000	-0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
White	-0.007	-0.008*	-0.002**	-0.001*	-0.001**	-0.001**
	(0.004)	(0.004)	(0.001)	(0.001)	(0.001)	(0.000)
% 20-24 years	-0.020	-0.019	-0.011	-0.011*	-0.012*	-0.012*
	(0.026)	(0.025)	(0.006)	(0.005)	(0.006)	(0.006)
% 25-34 years	-0.010	-0.017	-0.005*	-0.003	-0.006**	-0.003
	(0.010)	(0.011)	(0.003)	(0.002)	(0.002)	(0.002)
% 35-44 years	0.014	0.022	-0.006	-0.009**	-0.007	-0.010**
	(0.024)	(0.024)	(0.004)	(0.004)	(0.004)	(0.004)
% 45-54 years	-0.034	-0.043*	-0.007	-0.003	-0.007	-0.004
	(0.020)	(0.019)	(0.005)	(0.003)	(0.005)	(0.004)
% 55-64 years	-0.012	-0.008	-0.009***	-0.011***	-0.009***	-0.010***
	(0.015)	(0.015)	(0.003)	(0.002)	(0.002)	(0.002)
% 65> years	0.005	0.000	-0.004	-0.002	-0.007	-0.005
	(0.023)	(0.021)	(0.004)	(0.004)	(0.004)	(0.004)
Constant	2.607**	2.933**	1.944***	1.806***	1.970***	1.845***
	(1.033)	(1.059)	(0.331)	(0.248)	(0.319)	(0.259)
Observations	22	22	22	22	22	22
R-squared	0.439	0.483	0.839	0.887	0.823	0.873
Robust standard errors in parentheses						
*** p<0.01, ** p<0.05, * p<0.1						

Table 11

5. Conclusion

This research tries to answer the question: How does the predicted probability of computerization relates to developments in employment and wages across jobs? Therefore, the wage and employment data of hundreds of individual occupations in the United States and the United Kingdom have been analyzed.

To answer the main research question three hypothesis were formed. The first hypothesis involves the wage data. In contrast to expectation in the data a positive correlation between the Probability of Computerization and the wage changes were found in the US data. A side note must be placed here that some evidence was found that in the case of highly educated workers, a negative correlation was found. The evidence was very consistent for all the regressions, suggesting that results are reasonably valid.

The second hypothesis involves the employment data. Here the expected correlation was found in US data, a negative correlation between the Probability of Computerization and employment changes. While this result could be defended the true results, as employment changes might be likely to show results sooner than fixed wages. This specific result is divergent from the other results in the thesis, it is also supported by the least number of regressions.

The third hypothesis involves the correlation between the Probability of Computerization at certain levels of education. Contrary to the expectations all the evidence points in the directions that low-skilled occupations have a weaker negative correlation between the Probability of Computerizations and recent employment and wage changes. While the unanimity in the expectation of automation to hit low-skilled workers is very present in the literatures. Every significant source of evidence, even the UK data, in this thesis points directly into the contrary direction.

To then answer the main research question, the probability of computerization does show correlation with developments in employment and wage. It shows negative correlation with the recent change in total employment and unexpected positive correlation with recent change in wages. Furthermore, it shows more negative correlation for high skill occupation than for low skill occupations.

The references to the US data in the paragraphs above imply the lack of evidence present in the UK data. This means that the research failed in the contribution to test if the predicted probability of computerization has validity outside of the United States. The reason this is true could very well be the detour that was made to link this data set to the Frey and Osborne. However, it is a shame that no external validity was found as automation is a global phenomenon, which will not stop at the US border.

What the research did accomplish was a modest contribution in the ever-growing discussion regarding automation and economical and social changes this will bring forth. To my knowledge this is the first research that has used a statistical test to verify if the predictions done in the contemporary automation literature correspond to real world data. It will therefore also be the first research paper to come to the remarkable conclusion that automation may not be terrible news for low-skilled workers. Furthermore, I would like to stress that is would be wise for organizations and governments to closely follow the automation trends in the labor market. As in the end it might not even be needed to put everybody trough college.

5.1. Limitations

The biggest limitation of this research is the relatively short time that has elapsed since Frey and Osborne have done their predictions. Those predictions, first made in 2013, are long term predictions and a lot of the occupations they marked as easily automatable in the future are anno 2020 far from automated. A good example from the dataset is the occupation group “Taxi Drivers and Chauffeurs” which is denoted with a Probability of Computerization of 89, while not a single taxi driver or chauffeur has lost his job due to automation at the time of writing.

Another limitation this research has is the lack of control variables used in the main regression. Because the detailed level of occupational wage and employment data the associated control variables are hard to come by. Especially the way the business cycle has affected the low-skilled jobs seems to be a problem for this regression model. A future research with more detailed data and/or less during steep economic cycle would resolve this problem for research in the same field.

Finally, the matching of data between the United States and the United Kingdom was not ideal. The translating route through the ISCO has resulted in an incomplete dataset weakening the likelihood to find external validity for this research topic in other labor markets.

Bibliography

Acemoglu, D., & Autor, D. H. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of labor economics Vol 4*, 1043-1171.

Acemoglu, D., & Restrepo, P. (2018). Artificial Intelligence, Automation and Work.

Arntz, M., Gregory, T., & Zierahn, U. (2017). Revisiting the risk of automation. *Economics Letters, 159*, 157-160.

Autor, D. (2014). *Polanyi's paradox and the shape of employment growth.* Cambridge, MA: National Bureau of Economic Research.

Autor, D. H. (2014). Skills, Education, and the Rise of Earnings Inequality. *Science 344, no. 6186*, 843-851.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic perspectives 29(3)*, 3-30.

Autor, D. H., & Salomons, A. (2018). Is automation labor-displacing? Productivity growth, employment, and the labor share. *National Bureau of Economic Research*.

Autor, D., & Dorn, D. (2009). The Growth of Low Skill Service Jobs And The Polarization of The Labor Market. *NATIONAL BUREAU OF ECONOMIC RESEARCH*.

Berg, A., Buffie, E. F., & Zanna, L. F. (2018). Should we fear the robot revolution?(The correct answer is yes). *Journal of Monetary Economics, 97*, 117-148.

BLS. (2020, January 12 dec). *Occupational Employment Statistics*. Retrieved from bls: https://www.bls.gov/oes/tables.htm

Brandes, P., & Wattenhofer, R. (2016). Opening the Frey/Osborne black box: Which tasks of a job are susceptible to computerization?

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* New York: WW Norton & Company.

Buera, F. J., Kaboski, J. P., & Rogerson, R. (2015). Skill biased structural change (No. w21165). *National Bureau of Economic Research.*

Dimand, R. W., & Spencer, B. J. (2008). Trevor Swan And The Neoclassical Growth Model. *NATIONAL BUREAU OF ECONOMIC RESEARCH*.

Eden, M., & Gaggl, P. (2016). On the welfare implications of automation.

Forbes, K. (2016). *a tale of two labour markets .* Retrieved from Bank of England: https://www.bankofengland.co.uk/-/media/boe/files/news/2016/january/a-tale-of-two-labour-markets-the-uk-and-us-speech-by-kristin-forbes.pdf

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *echnological forecasting and social change 114*, 254-280.

Frohm, J., Lindström, V., Stahre, J., & Winroth, M. (2008). Levels of automation in manufacturing. *Ergonomia - an International journal of ergonomics and human factors, 30(3)*.

Goedhuys, M., Janz, N., & Mohnen, P. (2008). What drives productivity in Tanzanian manufacturing firms: technology or business environment? *European Journal of Development Research, 20:2*, 199-218.

Hopkins, E. (2013). *Industrialisation and society: a social history, 1830-1951.* Routledge.

James, J. A., & Skinner, J. (1985). resolution of the labor-scarcity paradox. *The Journal of Economics History*, 513-540.

Kizer, J. (2020, January 14). *Occupation Computerization*. Retrieved from data.world: https://data.world/jonathankkizer/occupation-computerization

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2016). *Autonomous driving.* Berlin: Springer Berlin Heidelberg.

Mokyr, J. (1994). *Institutions, Technological Creativity and Economic History. In Innovation, Resources and Economic Growth.* Berlin: Springer.

Nedelkoska, L., & Quintini, G. (2018). *Automation, skills use.* OECD.

Pajarinen, M., Rouvinen, P., & Ekeland, A. (2015). Computerization and the Future of Jobs in Norway. *Oslo: Statistisk sentralbyrå*.

Rasmussen, W. D. (1982). The Mechanization of Agriculture . *Scientific American 247(3)*, 76-89.

Reinert, H., & Reinert, E. S. (2005). *Creative destruction in economics: Nietzsche, Sombart, schumpeter.* Boston, MA.: Springer.

Appendix

Major	Major Group	Average Probability of Computerization
11	Management Occupations	0.09
13	Business and Financial Operations Occupations	0.55
15	Computer and Mathematical Occupations	0.11
17	Architecture and Engineering Occupations	0.28
19	Life, Physical, and Social Science Occupations	0.25
21	Community and Social Service Occupations	0.05
23	Legal Occupations	0.50
25	Educational Instruction and Library Occupations	0.25
27	Arts, Design, Entertainment, Sports, and Media Occupations	0.24
29	Healthcare Practitioners and Technical Occupations	0.17
31	Healthcare Support Occupations	0.49
33	Protective Service Occupations	0.37
35	Food Preparation and Serving Related Occupations	0.78
37	Building and Grounds Cleaning and Maintenance Occupations	0.78
39	Personal Care and Service Occupations	0.50
41	Sales and Related Occupations	0.65
43	Office and Administrative Support Occupations	0.84
45	Farming, Fishing, and Forestry Occupations	0.78
47	Construction and Extraction Occupations	0.73
49	Installation, Maintenance, and Repair Occupations	0.68
51	Production Occupations	0.82
53	Transportation and Material Moving Occupations	0.70

Table 12: The Major group list