University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Kevin Kanarbik

An empirical investigation on wage inequality in Estonian firms

Master's Thesis (30 ECTS)
Software Engineering Curriculum

Supervisors:

Rajesh Sharma PhD
Jaan Masso PhD

Tartu 2019

Abstract:

An empirical investigation on wage inequality in Estonian firms

Wage inequality has been thoroughly researched throughout the world, but less research has been conducted on the general wage inequality of Estonia. The country is intriguing because its wage inequality is high with its gender wage gap being the highest in Europe. The purpose of this thesis is to empirically explore matched employee-employer data of Estonia from 2006 to 2014 in order to find correlations between with-in firm wage inequality and other firm traits. In addition, the data is studied with linear regression and other predictive models. A secondary goal is to find a correlation between wage inequality and company growth. We discovered that wage inequality depends on the sector and region of the company with wage inequality generally decreasing from 2006 to 2014. Furthermore, we identified that income inequality is moderately correlated with firm growth, size and average wage.

CERCS: P160 Statistics, operation research, programming, actuarial

Keywords: wage inequality, firm growth, exploratory analysis, predictive analysis

Eesti firmade palgalise ebavõrdsuse uurimine

Palgaline ebavõrdsust on põhjalikult uuritud üle maailma, kuid Eesti üldist palgalist ebavõrdsust on vähem uuritud. Antud riik on huvitav, kuna riigi sisemine palgaline ebavõrdsus on kõrge ja sooline palgalõhe on kõige suurem Euroopas. Käesoleva töö eesmärk on empiiriliselt uurida Eesti tööandjate ja töötajate ühendatud andmeid aastatest 2006 kuni 2014, et leida korrelatsioon firmasisene palgaline ebavõrdsuse ja teiste firma omaduste vahel. Andmed samuti uuriti lineaarse regressiooniga ja teiste prognoosivate mudelitega. Teisejärguline eesmärk on leida korrelatsioon palgalise ebavõrdsuse ja firma kasvu vahel. Me avastasime, et palgaline ebavõrdsus sõltub tihti firma majandusharust ja regioonist ning palgaline ebavõrdsus on pidevalt langenud aastast 2006 kuni 2014. Lisaks leiti, et palgaline ebavõrdsus on mõõdukalt seotud firma kasvuga, suurusega ja keskmise palgaga.

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Keywords: palgaline ebavõrdsus, firma kasv, empiiriline analüüs, ennustav analüüs

Contents

1 Introduction

In modern times people live their lives by working in a job. Their job provides them with a monetary wage that they can use to buy food supplies, home and other essentials for everyday life. Different people earn different amounts from their job because occupations are diverse and thus they are rewarded differently. The distinction of wages between different groups of people is by definition called wage inequality. High wage inequality occurs when the pay of highly-skilled workers is much larger than the pay of low-skilled workers.

Wage inequality shows income distribution across individuals, families and population hubs in specific regions or countries. It is a very important research subject for academia because of the many articles and research papers that have investigated the reasons, trends and effects of income inequality on the economy as a whole [14]. Researchers have found that a nation's income inequality can affect crime, corruption, economic stability and investment levels [1]. Additionally, wage inequality can affect people more directly - income inequality effects life expectancy, infant mortality rate and the murder rate [3]. Inequality has negative social and economic consequences for the population, but new studies have found something positive in inequality. By looking at employee level data in different countries and companies academic papers have found that wage inequality is positively related to growth [21]. Academics have seen through research and survey's that wage inequality occurs throughout the world by different levels depending on the region and this inequality has also changed differently by region. For example, during the last 30 years wage inequality in the United States of America has increased while the wage inequality in the nations of Europe has decreased [8].

In this thesis, we study wage inequality in the case of Estonia. The country of Estonia is an interesting research subject in regards to wage inequality. The country of Estonia has gone through very extensive economic reforms when it transitioned from a centrally planned economy to a free-market one and the wage inequality of Estonia is quite high compared to its Western European partners [25]. In addition, it has been reported that the wage inequality between different genders in Estonia is the highest in the European Union [26]. Although a broad range of research has been conducted on income inequality in different parts of the world, the relationship between wage inequality and other features of companies has been largely overlooked in Estonia. Academics in Estonia have looked into more specific topics regarding wage differential like the effect of minimum wage and unions on wages and the gender wage gap by Ferraro, Meriküll and Staehr [13], Eamets and Kallaste [10] and Vassil, Eamets and Mõtsmees [26] accordingly.

To the best of our knowledge, this is the first study with the wage data of Estonia in this that has uses predictive analysis. Other studies have looked at wage inequalities in other countries like the United Kingdom than the wage inequalities in Estonia [21]. Our exploratory analysis shows findings in different sectors. Academics have previously researched wage inequality in firms and countries in the following ways. They have conducted qualitative research - Akerman,

Helpman, Itskhoki, Muendler and Redding [1] and Bernstein [7] look at what effects does wage inequality have socially on society by discussing the effects independently and by referring to previous social studies. Quantitative research has also been conducted by Barro [4] and Mueller, Ouimet and Simintzi [21] who look at how wage inequality effects a specific feature of a company or a country by applying statistical methods on wage data. Researchers also look at the effects of wage inequality in specific countries - Alvarez, Benguria, Engbom and Moser [2] investigate wage data of Brazil. Akerman, Helpman, Itskhoki, Muendler and Redding [1] investigate Swedish wage data. Academics have also focused on wage inequality between different types of groups. Vassil, Eamets and Mõtsmees analyzed the gender wage gap and wage inequality between different ethnicities was investigated by Leping and Toomet [19]. Finally, some research papers look at wage inequality with-in a specific sector and use survey data with small sample sizes. The sector focused papers with survey data are the following:

1. Bell and Reneen [6] focus on the financial sector in the UK with survey data of 5000 households.

2. Campos-Soria, Ortega-Aguaza and Ropero-Garcia [9] focus on the accommodations and food service sector in Andalusia with survey data of 3211 workers.

3. Sarkar and Singh Mehta [23]. focus on the ICT sector in India with survey data of 1087472 workers

4. Hyder and Reilly [16] focus the wage inequalities between public and private sector workers in Pakistan with survey data of 7352 workers.

In this work, we use a large dataset of payroll taxes from the Statistics Estonia facility. The data set describes an average annual number of 18,000 companies and 378,000 employees during the period between 2006 and 2014. During the research part, we take a two-fold approach - for the first step we conduct exploratory analysis and on the next step we do predictive analysis on the data using machine learning methods (linear regression, gradient boosting methods, random forest, etc.).

Our contribution to the prior literature on the topic of wage inequality is to provide data analytic research by an empirical analysis of a very large data set that describes 8 years of national wages. This data has previously not been looked at by comparing wages between different economic sectors and firm sizes. We find that the mining sector has significantly lowered its wage inequality, but it still has one of the highest inequalities along with the health and electricity sector. In addition, we provide proof that the capital city has the highest wages and highest wage inequality in the country. More importantly regarding the relationship of wage inequality and other firm attributes, we provide evidence that there is a slight correlation between wage inequality and the company's mean wage, firm size and growth.

The rest of the paper is organized as follows. In the related work section, we discuss literary works of a similar topic like wage inequality, income inequality with growth, wage inequality

in different sectors, wage differentials in Estonia and economic analysis methods. Different levels of wage inequality in different economic sectors and regions are brought out. On the data description chapter we characterize the data, show the metrics of the data set. We have manipulated the data as well so the data could be explored more and manipulation methods are also described. The main chapter of this thesis is the research part, where we report the background of used methods and presents the results of our exploratory and predictive analysis. We conclude the thesis with a discussion of future directions in the conclusion chapter.

2 Related work

On this chapter of the thesis, we examine academic works that focus on wage inequality in different aspects. Firstly, we bring out papers that have looked into income inequality socially - what is the reason behind it, what does it effect, how to limit it and is it generally positive or negative. Secondly, we talk a bit about wage inequality levels in different sectors of the economy. In addition, we introduce Estonia, its economic history and its current situation regarding wage inequality. Finally, we briefly talk about the relationship between income inequality and other economic features.

2.1 Wage inequality

Wage inequality is defined as the income difference between different groups of people. Wage inequality is high if the wages of a population are unevenly distributed and the opposite occurs if wages are evenly distributed. In modern times, income inequality is a popular and important topic by economists, because it is related to the prosperity and stability of people in different countries. Academics have theorized that there would be a lot of social and economic problems if the income gap between the wealthy and the underprivileged is too wide. Thus, wage distribution is a popular topic for academia and governments, who usually investigate the reasons behind wage inequality [27].

2.1.1 Reason

There has been a lot of research into the question of why has wage inequality increased throughout the years. An analysis by Lemieux [18] found different reasons for the increase during different decades. The paper argues that the main reason for the wage inequality increase is the demand for skilled workers. This skill-demand was not a new phenomenon, because the increase in demand began in the 70s. This demand increased even further in the 90s because in the 70s and 80s an educated "baby boom generation" met this skill-demand. During the 90s not only were there less educated young people, but the computer revolution also increased the demand for skilled workers. A steep competition to find skilled workers increased the wages of these skilled workers ergo the inequality increased. Secondly, the paper mentions that while the technological and demographic change was global, inequality increase was not. Some countries kept their inequality low, unlike the US. This research found that another explanation for inequality was wage-setting institutions and unions. Some countries have strong unions and institutions that contribute to compressed wage distribution while in other countries there is weaker regulation of wages.

One of the main theories about inequality in economics is called the Kuznets curve which also explains the reasons behind wage inequality. Figure 1 shows the Kuznets curve. The curve describes the rise and fall of wage inequality when an economy moves from an agricultural fo-

cus to industrial. When an economy is focused on agriculture then most people earn a similarly low amount, thus the inequality is low. During the transition period from agriculture to industry, the minor amount of people who went into the industry at the beginning is earning a significant amount compared to the majority who are still in the agricultural sector. This rises inequality, but as more and more people go to work in the industry sector than more are also getting a significant pay rise. In addition, after the main transition phase when the majority of the population is in the industry sector then the wages of the agricultural sector start to rise because it is now a more competitive and unpopular sector and thus wage inequality lowers again [4].

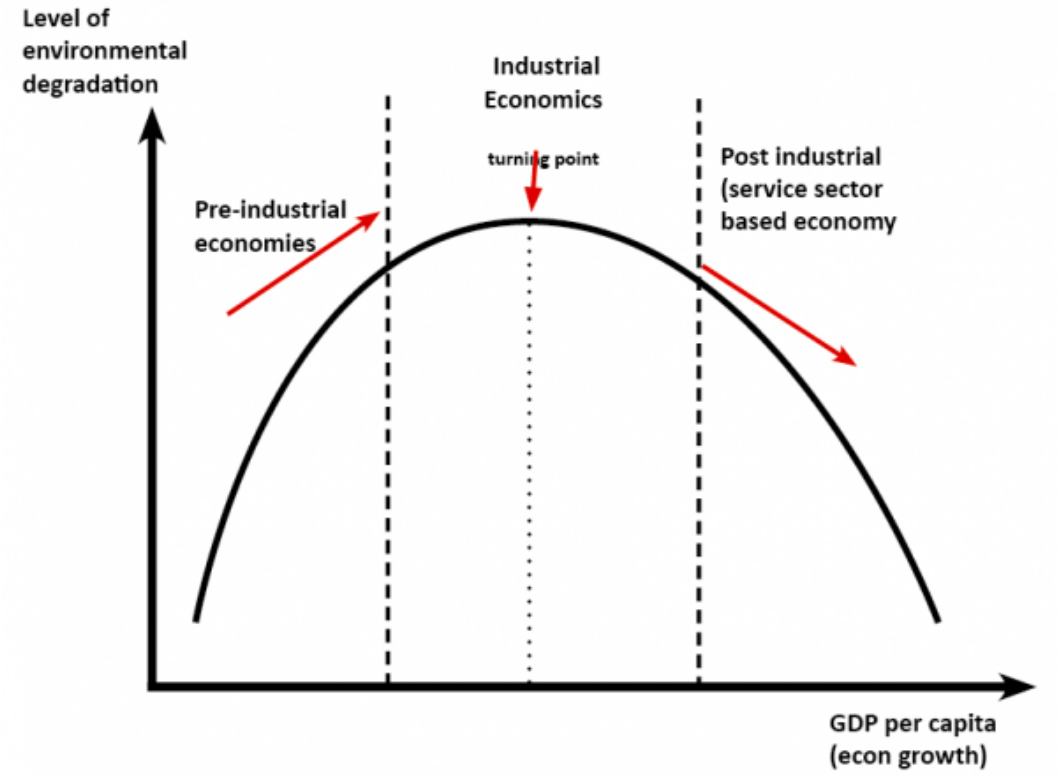

Figure 1: Kuznets curve

Source: https://www.economicshelp.org/blog/14337/environment/environmental-kuznets-curve

Another research paper by Song, Price, Guvenen, Bloom and von Wachter [24] was specifically targeting wages within companies. They found that inequality within firms has increased because the earnings of top employees have grown relative to the firm itself. This means that throughout the years these companies have experienced growth in revenue and size. Because of this growth, the owners and the top workers in there firms also get a boost in their pay through promotion, wage increase and benefits.

2.1.2 Effects

Many researchers have looked into how wage inequality effects nations. Babones [3] looks at how inequality is related to population health while investigating if it is causal. He finds that life

expectancy and infant mortality rates are directly related to wage inequality. What needs to be taken into regard is that these 2 are individual-level variables, but inequality is an aggregate level variable. In addition, the murder rate is also weakly related to wage inequality [3]. Additionally, a contradicting theory related to the savings rate suggested that inequality is positive for growth. Savings rates rise with the level of income because richer people have more money to put aside for investment. Policies that lower inequality by redistributing resources from the wealthy to the poor also lower the saving rate of the rich, which lowers the average saving rate of the country. This means that with high inequality the saving rate is higher meaning there is more investment and thus growth is faster [4].

Berenstein [7] theorized that inequality promotes credit bubbles. This is explained by the following: high inequality means that middle and low-class people have a low wage and this makes them borrow more money from banks in order to heighten their living standards while the high-class people have a high income and culminate greater wealth. Because of these reasons, the financial sector quickly grows as average debt-to-income numbers also increase. Because of low and stagnant incomes and the large middle-class debt the financial sector grows unstable, eventually crashing and causing an overall financial crisis, similar to the one that happened in 2008.

2.1.3 Reduction

There are many ways on how to reduce inequality and most of these methods have been used in the real world. Alvarez, Benguria, Engbom and Moser [2] investigated policies and work institutions in order to find out why Brazil has significantly decreased its wage inequality. The researchers found that firms themselves play a very significant role and that worker heterogeneity is very important in determining inequality. The first reason for the decrease in inequality is related to educational attainment. The researchers found that more experienced and educated workers are paid more. This explains the inequality decrease because data shows that throughout the years the average education level and age of workers has steadily increased. A larger, more educated worker population in Brazil has decreased inequality. The second reason is related to the bargaining power of the unions of Brazil. Unions have regulated the wages of the firms by forcing them to implement higher and equal worker wages. The bargaining power throughout the years has become stronger and more centralized which means that the pays got less dependent on firm or worker specific performance. More equal pay by the regulators in unions has decreased inequality. The final reason is related to the minimum wage in Brazil. It has risen 119% so low-skilled workers have gotten a good pay raise which means the inequality has decreased.

2.2 Inequality by sector

The characteristics of a company are often related to the business sector that it belongs to. Some companies in one sector are very low-wage worker intensive while companies in another sector have a lot of high-wage workers. The business sector can be related to the percentage of men versus women in a company or it can be related to the region of the company as well. Researchers have compared different broad sectors with each other and looked into their within sector wage inequality, but some have also investigated the wage distribution in a specific sector.

The wage distribution in private and public sectors in Pakistan has been thoroughly investigated by Hyder and Reilly [16] using survey data of 7352 workers. They found that overall the wages in the private sector are larger than those in the public sector. This is because private sector companies are motivated by profit and thus the wages of workers are set by their productivity. These pay rewards generally do not exist in the public sector. In regards to the wages between different groups of people in the private and public sectors than these wages also differ from each other. In the private sector, more educated workers get greater wages, but the gender wage gap is stronger here compared to the public sector. In the public sector, the correlation between wage and age is stronger. In respect to wages, inequality public sector wages are more compressed meaning the wage inequality is smaller than in the private sector because high-skill workers get paid less than their private sector counterparts. The opposite applies to low-skill workers - they get paid more in the public sector.

Bell and Reenen [6] looked into the rise of inequality from the 90s, noticing that the main reason for this is the increase of wages for workers at the top of the wage distribution aka the workers who earn the most out of all the workers. In addition, they focus on wage inequality in a specific business sector - finance. The financial crisis brought the wages in the financial sector into attention, thus researchers looked into it using survey data of 5000 households in the United Kingdom. The general rise of wage inequality across all business sectors is heavily related to the rise on inequality in the financial sector because 60 percent of the top high-wage workers, whose wage has risen, are financial workers. This is significant because in the 90s only 12 percent of workers in the top high-wage category was in the financial business. One of the reasons behind the increase in financial worker wages are the bonuses that they get. While in other sectors, bonuses make up about 5 percent of the overall wage then in the financial sector the bonuses are generally 25 percent of the overall wage and this percentage is even higher for high-income workers there. The top 10 percent of workers in the financial sector share 2.2 percent of the total gross income of all financial workers. One of the possible reasons for this wage inequality and massive wages is related to talented workers in finance getting a lot of bonuses while less talented do not. Finance workers like traders excessively take more risks and gather massive profits upon success. Unsuccessful workers may lose their job, but they still keep their bonuses. So overall it seems that in the banking sector the inequality is quite high. This wage inequality is

also high between different genders. Studies show that high-wage professionals in the financial sector are mostly male with female workers being either discriminated or are less inclined to ask for higher wages.

Campos-Soria, Ortega-Aguaza and Ropero-García [9] researched wage differences in the hotel and restaurant industry by focusing on the gender wage gap. Their data set consisted of 3211 workers in Andalusia. This tourism focused industry is known for having a large proportion of women compared to other sectors which lower the average median wage and cause stronger gender wage gaps of wages. Studies find that in this sector women are paid 10-20% less than males and this is the main cause of the overall high wage difference between workers in the hotel and restaurant sector.

Information and Communication Technology (ICT) sector in India consists of the manufacturing and services industries. The wage inequality of this sector in the country of India was analyzed by Sarkar and Mehta [23]. by using a household survey data of 1087472 workers. They analyzed inequality with different measures like mean log deviation, Theil index and the Gini coefficient. They found that wage inequality is lower in the ICT sector than in other sectors. One index showed that between higher wage workers wage inequality is stronger than low wage workers in this sector. Lastly, this paper saw that the workers in the ICT sector are paid more than workers in other sectors.

Fleming and Measham [14] looked into the effects of the mining industry in smaller state areas in Australia. For their analysis, they used national census data. In that country, there was a mining boom in the years 2001 to 2011 and researchers looked into areas where mines were established to see if wage inequality in those areas were affected. Wage inequality was calculated with the Gini coefficient and the data was gathered from national censuses. The mining sector usually employs a limited amount of people who get high wages so it's logical that mining increases the inequality. What they found was that overall income inequality has risen substantially in Australia, but this rise was weaker in mining areas. This means that compared to other areas in the nation, especially urban-city areas, the mining industry has kept inequality in lower levels. This means that mining benefits the whole community thus the income inequality in the mining sector is low.

2.3 Estonia

Estonia is a small Baltic nation that regained its independence in 1991 from the Soviet Union. The capital of Estonia is Tallinn where almost half the population lives. The country has a large national minority of Russians living there. During the transition from a state-owned and centrally planned economy to a free-market one, Estonia went through very extensive economic reforms. The result of these reforms gave Estonia high growth rates which rose income levels, but these were still minuscule compared to the incomes of Western European countries. A major objective for politicians is to continue the economic growth in the order of achieving similar income levels as their western partners. This objective is important for the benefit of the population because the distribution of income in the country is quite unequal - 20 percent of the population with the highest income earn 6-7 times more than the 20 percent of the population with the lowest income [25].

This inequality of wages is even more significant between different ethnicity's. In the capital, Estonian workers on average get paid 30 percent more than their Russian counterparts [19]. In addition, gender wage inequality plays a major role in the Estonian economy. The country has the highest gender-based wage gap in the Europe Union. Women are paid 30 percent less than men and of course, this difference is even higher between Estonian men and Russian women [26].

Fortunately, the high wage inequality of Estonia is steadily being lowered by different policies. For example, compulsory minimum wage has played a significant role in lowering inequality in the country [13]. Estonia has a unique economic situation because the country's economy is growing and the wage inequality is high between Estonians, between different ethnicity's and between genders as well. This means that studying the country should yield interesting results [22].

2.4 Wage inequality and firm growth

Academic research has been conducted on the relation of inequality and the growth of firms. There are many papers on these topics separately by solely looking at either inequality or growth. Most existing papers look at this relation within the scope of a country and not of a company [2]. Barro [4] studied the effects of inequality on growth and produced theories on how inequality affects both growth and investment. One theory is called "political economy", which refers to the concept of redistribution of wealth through the political process because when inequality is high then the low-wage earning majority will vote for policies that support redistribution of wealth. These kinds of redistributions usually distort and reduce investment. Therefore, in the sense of politics, inequality hinders growth. Another theory called "sociopolitical unrest" says that high inequality tends to make the poor population riot, engage in criminal activity and disturb the peace. A higher crime rate in a country means a lot of energy wasted by the population on things that do not benefit the economy and it deters investment. In short

– high inequality means higher crime rates, which means less growth. Overall his findings on the relation between growth and inequality within a nation were interesting - wage inequality is negative to growth in poor countries aka low developed countries but it is positive to growth in rich countries aka first-world countries [4].

Similar results regarding poor and rich countries were found by Akerman, Helpman, Itskhoki, Muendler and Redding [1]. Their meta-analysis describes the results of previously done work in the field of inequality, which has sometimes concluded that wage inequality research results say that sometimes inequality is positive to growth and sometimes inequality is negative to growth. The author of the meta-analysis admits that the results of these analyses quite often depend on estimation methods, data type and data quality.

Forbes [15] tries to prove that the previously known fact in academia that inequality is negative to growth is wrong and what he found is interesting. Firstly, papers that concluded that inequality is negative to growth is flawed because there are measurement errors and biases. Secondly, previous research was wrong because corruption is taken into account which is positive to inequality and negative to growth. Lastly, government spending on public health and education is also taken into account even though it's negative to inequality and positive to growth.

Mueller, Ouimet and Simintzi [21] focused on how wage inequality is related to firm growth. Their paper uses statistics by looking at companies of different sizes and investigating their inequality. Inequality investigation is conducted by categorizing the workers by hierarchies: workers that have basic skills and low wages have a low hierarchy level while workers with more complicated skills and high wages have high hierarchy levels. These levels are compared with each other throughout the different companies in the dataset. What this paper found is the wages of high-skill workers increase relative to the wages of the low-skill workers when the size of the company increases. Wages of middle-level workers in firms do not increase relative to the wages of low-level workers. This means that wage inequality increases with firm size because high-wage managers in companies get higher wages and bonuses when the company size increases. In short, the larger the company the larger the inequality. Another conclusion that this paper made is that in more developed countries inequality is more positively associated with the growth of the largest firms. In short, this paper provides evidence that wage inequality is positively correlated with growth and company size.

3 Data description

On this chapter of the thesis, we describe the data that we will analyze for this thesis. First of all, we explain the origins of the data, what features are important for our research and what they mean. Secondly, we introduce the data by showing some metrics of it. Lastly, we describe how we altered the data for our analysis.

3.1 Description

We analyze the data from Statistics Estonia, a governmental agency working under the Estonian Tax and Customs Board. For our analysis, we collected two main data sets from this facility. The first source of the data is the Estonian Tax and Customs Office dataset of monthly wage tax dataset of Estonian workers for the period 2006-2014. The features of this data set that is significant for our analysis are worker ID, worker's company ID, wage, year, birth date and gender. These features are important because we want to distinct workers from each other and we want to know what company the worker belongs to in order to calculate the company's wage inequality. The year of the wage is important so we can compare the levels of wage inequality throughout different years. Birthdate is important because we want to calculate age and see if age is related to other features and gender is needed in order to calculate the gender wage gap in our data. The second source of the data is the Estonian Commercial Registry data on companies annual reports (balance sheets, profit and loss statements). The features of this data set that are important for our analysis are company ID, business sector code and county code. Company ID is important because we will use it to put this dataset together with the wage dataset which also features company IDs. The business sector code is needed because we will compare the features of companies between firms in different sectors. The county code is needed because we will compare the features of companies between firms in different counties. Overall, the data set has an average annual number of 18,000 companies and 378,000 workers. Estonia has 15 counties, but in our data, there are 16 county categories, because the capital city of Tallinn is separate from its home county Harjumaa.

In addition, there are 20 different categories of businesses that the companies are defined by the Estonian Classification of Economic Activities aka EMTAK which is the Estonian national version of the international classification called NACE. These economic activity groups are classified by an alphabetical class, which can be seen on most figures involving sectors. The economic sectors and their class are as follows: agricultural sector is A, mining is B, the manufacturing sector is C, the electricity sector is D, the water supply sector is E, construction sector is F, the wholesale and retail sector is G, the transportation and storage sector is H, accommodation and food service sector is I, IT sector is J, financial sector is K, real estate sector is L, professional and scientific sector is M, private sector administrative sector is N, public sector administrative sector is O, education sector is P, health sector is Q, the arts and entertainment

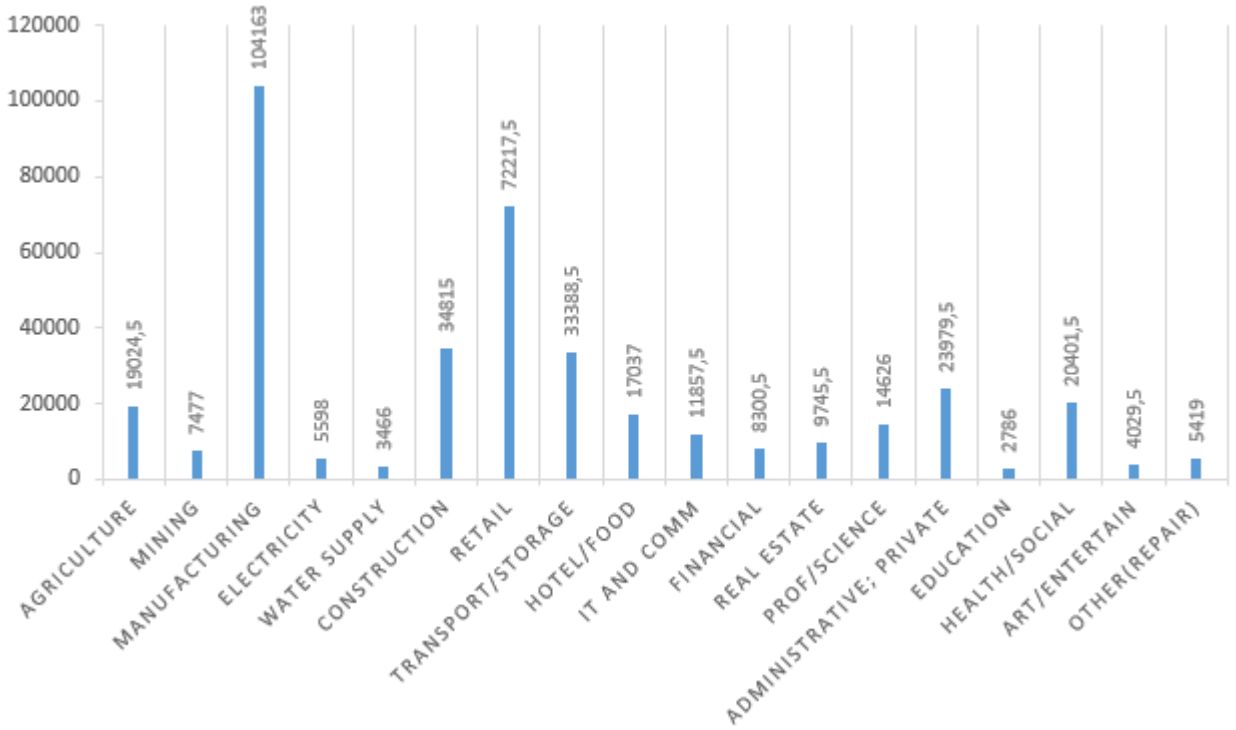

Figure 2: Annual number of workers in sectors

sector is R and other services like repairing sector is S. The average annual number of workers in each sector can be seen in figure 2.

Companies can be characterized based on the amount of workforce they have. A very small company with five workers is significantly different than a very large company with a thousand workers. For this reason, companies are categorized by the size of their work-force. For the purposes of this thesis, the European Commission's definition of companies by their size is used - micro-companies are firms with 1 to 9 workers, small companies are firms with 10 to 49 workers, medium-sized companies are firms with 50 to 249 workers and other companies with more than 249 workers are called large companies[1]. Small and medium companies separately belong to a class called SME which means small or medium enterprises. SMEs play a large role in smaller economies like Estonia, where they make up 80 percent of the economy. Generally, SMEs dominate the service business sectors like retail, IT, transportation and large companies dominate the manufacturing sector [12]. The number of workers and companies in different types of firms can be seen on table 1.

One important event has to be taken into account when exploring the data. In 2010 there were a limited amount of companies compared to 2006 and 2014. This is because during the middle of the period when the firm data was gathered the global financial crisis took place. This crisis had far-reaching and different effects on the country of Estonia in the years 2009 and 2010, but the significant one is the increase of company bankruptcies during this crisis. The exit of firms increased from 10% to 15% which can be seen by the number of companies in

[1]https://ec.europa.eu/regional_policy/sources/conferences/state-aid/sme/smedefinitionguide_en.pdf

Company type	# companies	# workers
firms with 1-9 workers	11075	48587
firms with 10-49 workers	5350	109218
firms with 50-249 workers	1172	114407
firms with 250+ workers	165	106328

Table 1: Num. of workers and companies

2009-2010 on figure 3. Fewer companies mean less employed workers which mean fewer data [20].

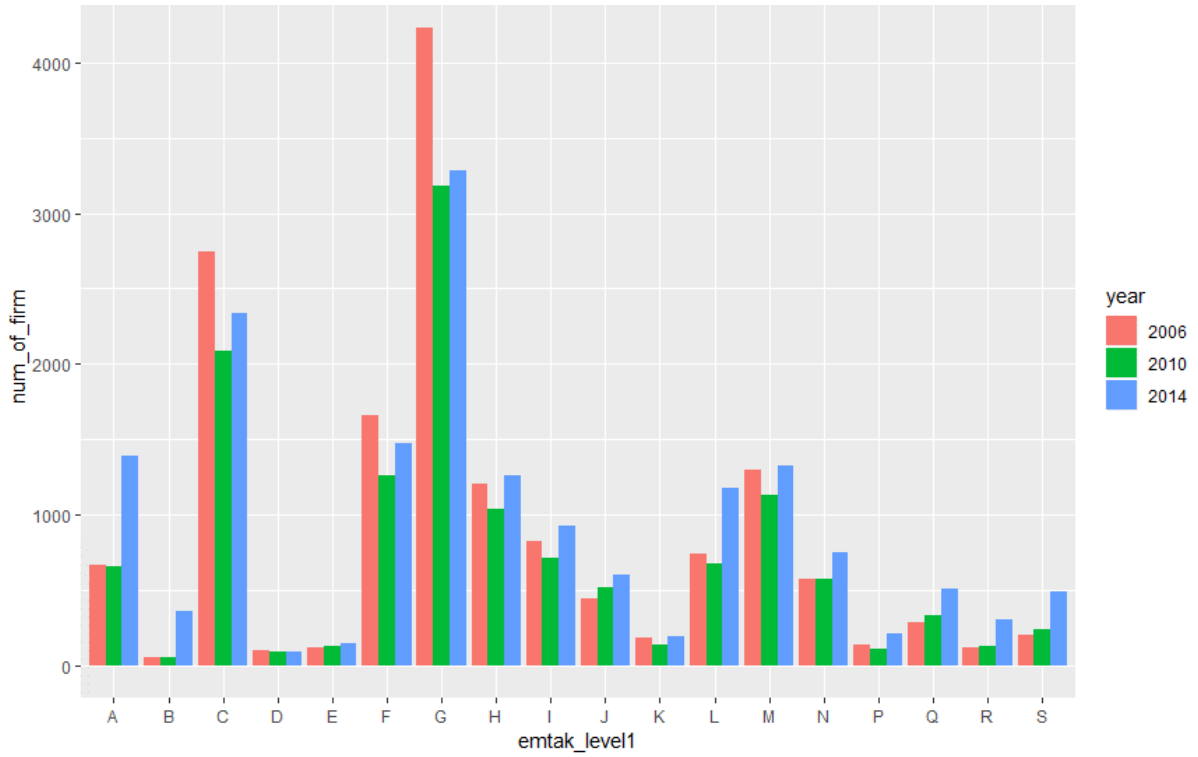

Figure 3: Number of firms (Y-axis) in company sectors (X-axis)

3.2 Data preparation

In order to make the raw data more readable and usable in our analysis, the data was modified in different ways. Firstly, the employee data set was aggregated to the company level and year to see the data of workers in companies by different years. Using the employees' wages within companies, the Gini coefficient was calculated for every company. Secondly, more calculations were conducted to find the mean of the wages within companies, the mean age of workers within companies, the number of workers within companies and the percentage of male workers within companies. The number of workers and male workers were calculated by counting their

amount for every company and the percentage of male workers was found by dividing their number with the total number of employees. Lastly, we calculated the growth of companies by using the change in the number of employees in the company. We calculated this growth differently than just subtracting the number of workers in year t with the number of workers in the previous year $t-1$ and then dividing it with the number of workers in year $t-1$. We use the equation (1), which is called the midpoint method[2]. This method is different from a standard method of calculating percentage change by dividing the subtracted value of current and previous with the median of these two values instead of dividing the subtracted value with the previous year. This midpoint method is better because the value is immediately normalized - the calculated growth percentage cannot be lower than -2 or bigger than 2.

$$\frac{(numofworkers_t - numofworkers_{t-1})}{\frac{(numofworkers_t + numofworkers_{t-1})}{2}} = growth_t \tag{1}$$

For the predictive analysis part, the categorical features had to be altered, so that they could be used in the linear regression model. For this dummy values were created for the 4 size classes, 20 business sector classes and the 16 county classes. Altogether there are 7 numerical features and the 40 additional dummy values that describe what are the classes of the companies.

[2]http://www.econport.org/content/handbook/Elasticity/Calculating-Percentage-Change.html

4 Methods

Gini coefficient is a statistical measure of dispersion which is used to serve as an indicator of the income distribution of a population. The value of the coefficient is between 0 and 1. A zero value of the Gini coefficient means pure equality where everyone has the same wealth. A Gini coefficient of 1 means ultimate inequality where one person has all of the wealth of the population and everyone else has nothing. Thus in general, a higher coefficient value means income is distributed unevenly [27].

Dummy values are used in this analysis for categorical features of the data. For every category a dummy column is created which indicates if an item belongs to the specific category or if it does not. For example, a person can be male or female, so we create columns $dummy.male$ and $dummy.female$. If the person is male than column $dummy.male$ will have a value of 1 and $dummy.female$ will have a value of 0.

Dummy values are needed because throughout this analysis we will use machine learning algorithms and some algorithms do not work with categorical features. Predictive algorithms try to predict the given value called the independent variable by using the other values linked with the independent variable which are called dependent variables. These algorithms are used to build different machine learning models that take dependent variables and return the predicted value of the independent variable. We used the following algorithms: linear regression, lasso regression, ridge regression, gradient boosting model, extreme gradient boosting and random forest.

Linear regression is a model that predicts using a linear function on one or many independent variables [17]. Lasso (Least Absolute Shrinkage and Selection Operator) or L1 regularization improves the accuracy of the prediction by decreasing the size of the coefficients of some features, sometimes even eliminating coefficients. Ridge or L2 regularization like L1 regularization improves the accuracy of the prediction, but all of the coefficients are decreased by the same factor, coefficients are never eliminated [3]. Gradient boosting model (GBM) is a machine learning model that can be used for regression and classification. It generates a model that consists of a collection of weak prediction models, mostly decision trees. Extreme gradient boosting utilizes the concept of gradient boosting, but it uses a more regularized model interpretation to lower over-fitting so it's performance is better than other gradient boosting models while being faster as well [4]. Random forest is a machine learning model that predicts by using multiple decision trees [17].

Models predict values and different metrics are used to judge the accuracy of the predictions. During this analysis, we will use MAPE aka Mean Absolute Percentage Error which is a statistical method of measuring the accuracy of a prediction. The formula for calculating

[3] https://www.statisticshowto.datasciencecentral.com/regularization/

[4] https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7

MAPE can be see on equation (2) where F_t is the forecast value and A_t is the real value [5].

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{2}$$

For every dependent value, a null hypothesis is tested. This hypothesis suggests that the coefficient is zero meaning the value has no effect. If the p-value is less than < 0.05 than we can reject the null hypothesis which means that this dependant value is valuable to our model [6].

[5]https://www.statisticshowto.datasciencecentral.com/mean-absolute-percentage-error-mape/

[6]https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

5 Results

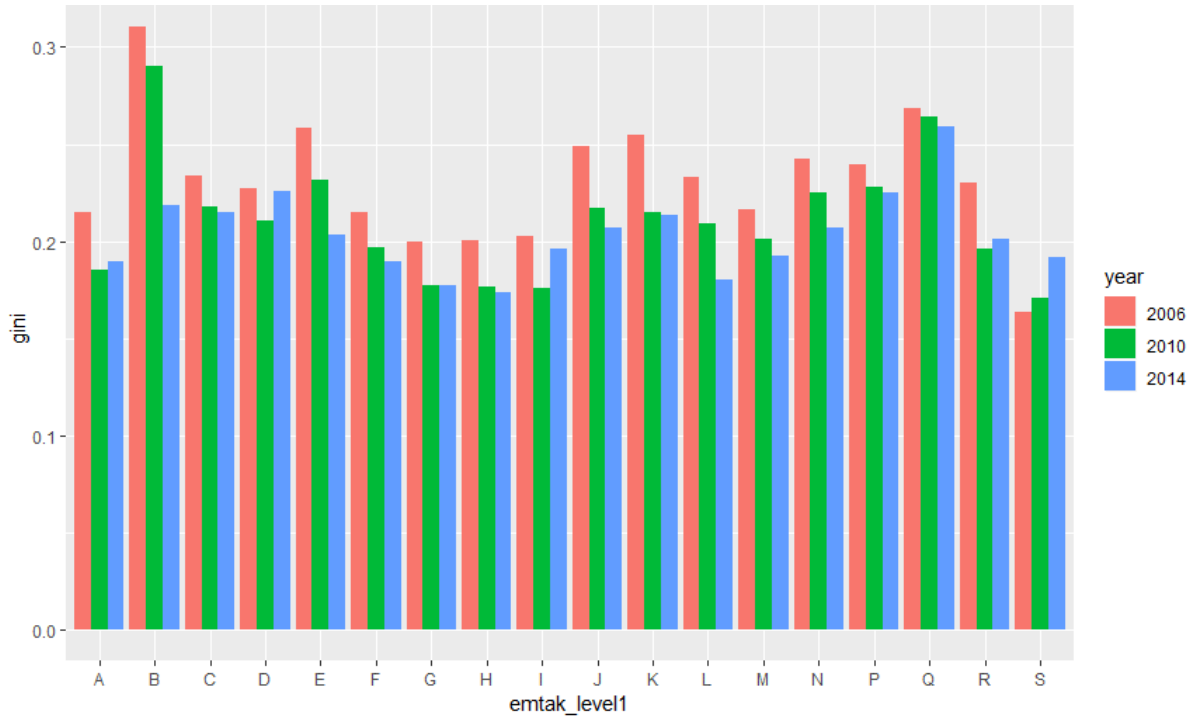

Figure 4: Wage inequality (Y-axis) and company sectors (X-axis)

The merged data set is ready to be explored after shaping it to be more readable. On the first part of the research, we conduct a descriptive analysis. The purpose of this analysis is to give a better idea of what is in the data set and what are the characteristics of the data and its many features.

One part of descriptive analysis is to group the data together and use tabulation summarizing to see the characteristics of the different categories. With these methods, we see how the number of companies and workers are divided between different company size classes, sectors and counties. Additionally, we analyze the annual inequality, average wage, number of firms and number of workers differences between the categories. Lastly, the relationships between the different features in the data set are analyzed in hopes of finding interesting correlations.

5.1 Descriptive analysis results

The average annual inequality in different sectors is analyzed as can be seen in figure 4. On the figure, we show inequality values on the beginning of our sample period, on the end of our sample period and on the middle of it so we can see how it is the wage inequality right after the financial crisis. What is interesting is that the mining sector (B) has lowered its inequality from $0,32$ in 2006 to 0,22 in 2014. The inequality of the health sector (Q) is high and has overall stayed the same level throughout the years. The sector that offers other services like repairing is the only one where wage inequality has actually increased its wage inequality level because it

has doubled its number of workers. All sectors have generally lowered their inequality. This is an interesting finding, showing Estonia is progressing into a less unequal society. Possibly one of the main reasons for this trend of wage inequality lowering is because of the minimum-wage in Estonia has been constantly raised which has increased the wage of low-wage workers thus decreasing inequality [13].

One exception to this continuous decrease of inequality is the year 2010 when wage inequality in Estonia was generally the lowest in some sectors but has since increased a little bit. This is most likely related to the global financial crisis of 2007-2009. This crisis severely affected nations throughout the world and Estonia was no exception. During this crisis employment of workers plummeted in 2009-2010 and increased after the crisis in 2011. Unemployment lowered the number of workers in the market and additionally cuts to wages decreased the nominal wages of workers. The decrease in inequality can be explained partly by the increase in unemployment and the decrease of nominal wage [10].

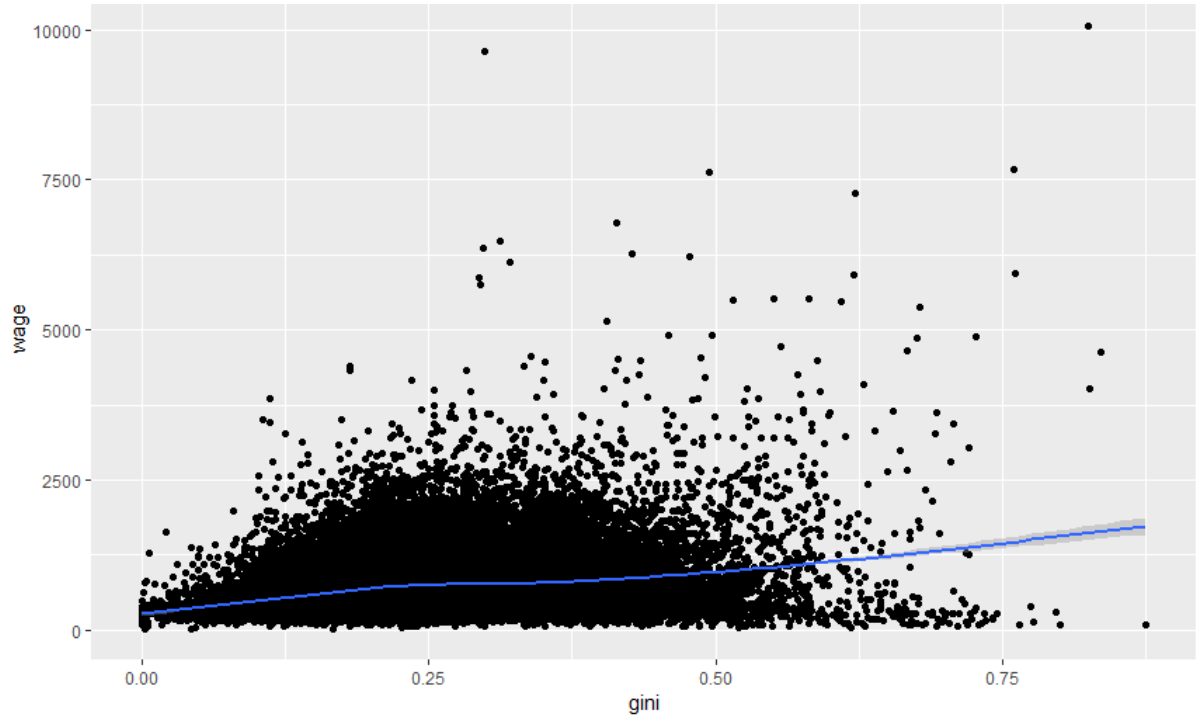

Figure 5: Average wage (Y-axis) and wage inequality (X-axis)

From this graph it is noticed that in the year 2014 the health (Q), electricity (D), education (P) and mining (B) sectors have a generally high inequality and the transportation (H), wholesale/retail (G), real estate (L) and agriculture (A) sectors have generally low inequality compared to others. One possible explanation for some sectors having higher inequality levels than others is that a few sectors have unions while others do not. In the European Union trade unions are strong institutions that significantly impact employment and wages, stronger unions mean more rigid wages. In Estonia, trade union membership is very unpopular with only 14 percent of workers belonging in unions in 2002. The construction sector and finance sector

are union-free. Collective bargaining coverage which plays a key role in setting wages in other countries is also very low with the exception of the transportation, health, education and culture sectors [11].

Thirdly, the average wages of companies is compared with the wage inequalities of companies which can be seen in figure 5. There seems to be a correlation between these two features meaning that the higher the average wage in a company, the higher the wage inequality in the company. The correlation between these two features was 0.274. This relationship should be looked into more.

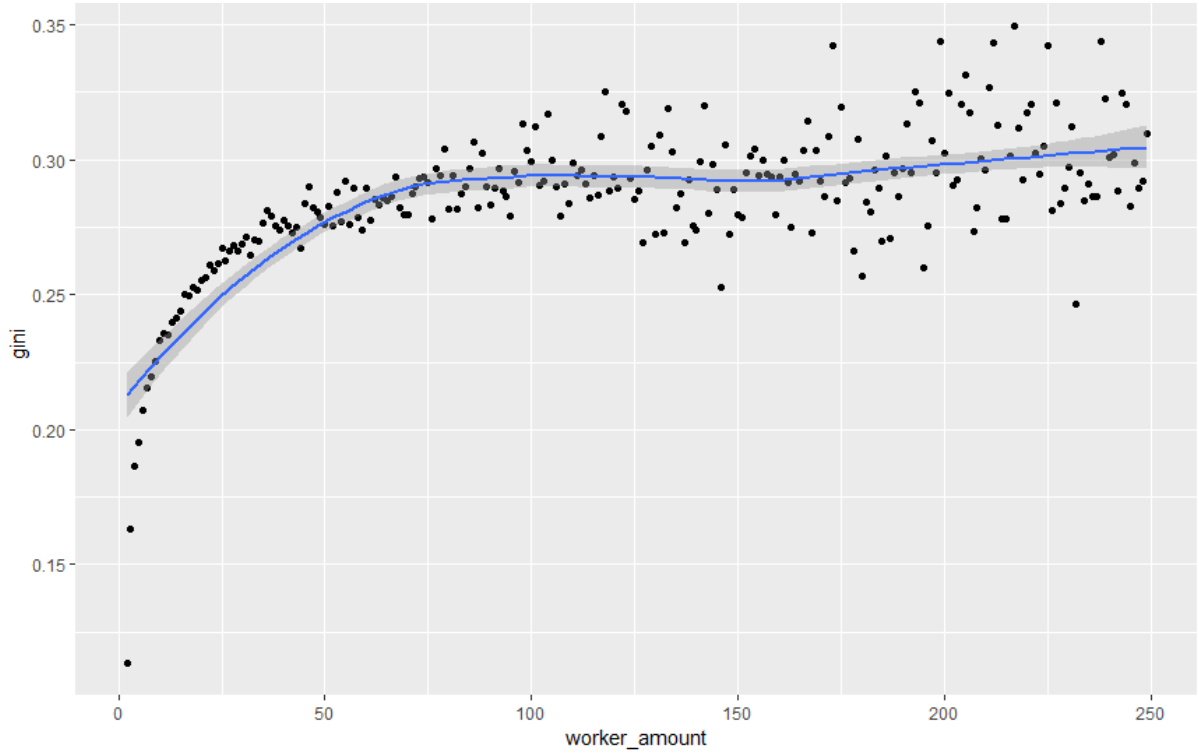

Figure 6: Number of workers and wage inequality in micro, small and medium companies. The worker amount value is on the X-axis and the Gini coefficient value is on the Y-axis.

In addition, the correlation between the number of workers in a firm and wage inequality within the firm is investigated. Large companies are not taken into account, because they consist of less than 1% from the whole amount of companies and the relationship between the size of a large company and its wage inequality varies wildly. But the relationship between these two features for micro, small and medium-sized companies follows a pattern as can be seen in figure 6. The more workers there are in a company the higher the wage inequality in it - this is proved by the graph and also by their correlation value which is 0.585.

Furthermore, the relations of wage inequality with features like the percentage of men and the average age of workers are also investigated, but there does not prove to be a correlation. The figure between the percentage of men in the company and firm wage inequality can be seen in figure 7. The correlation is -0.08 and the graph also shows that basically there is no correlation. The same can be said for the correlation between inequality and the average age of

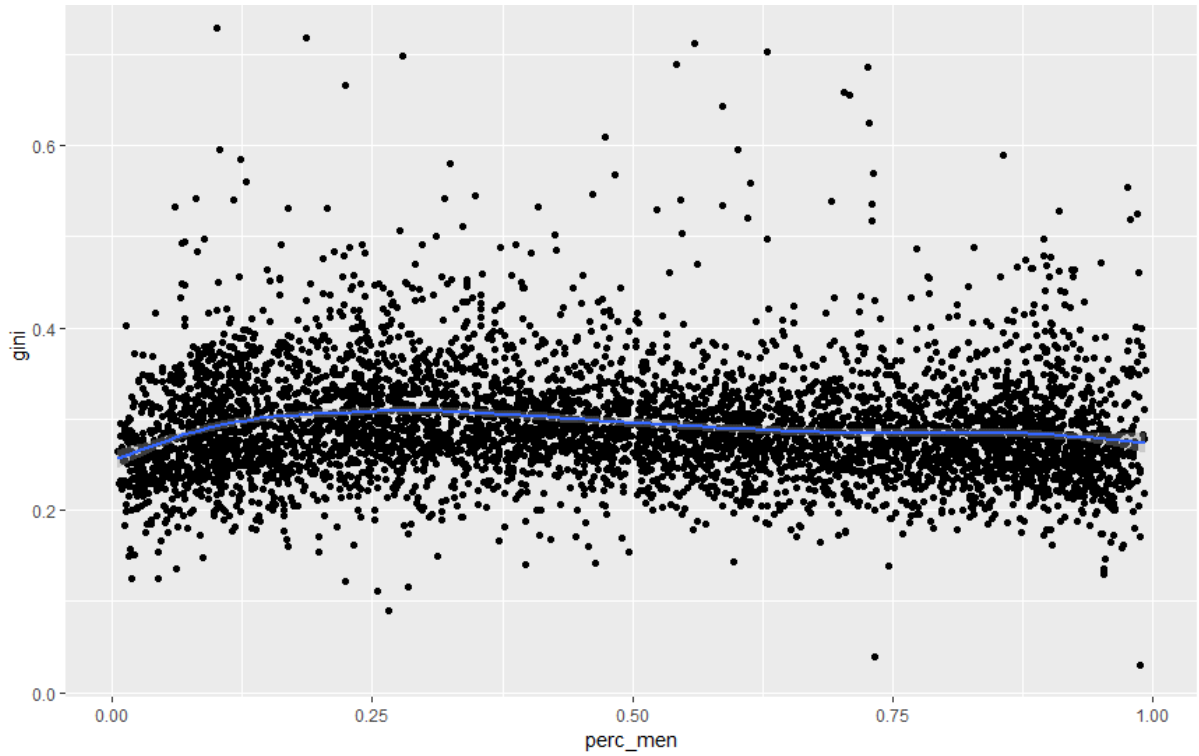

Figure 7: Wage inequality (Y-axis) and percentage of men (X-axis)

workers - its figure 8 shows no correlation and its value is -0.09.

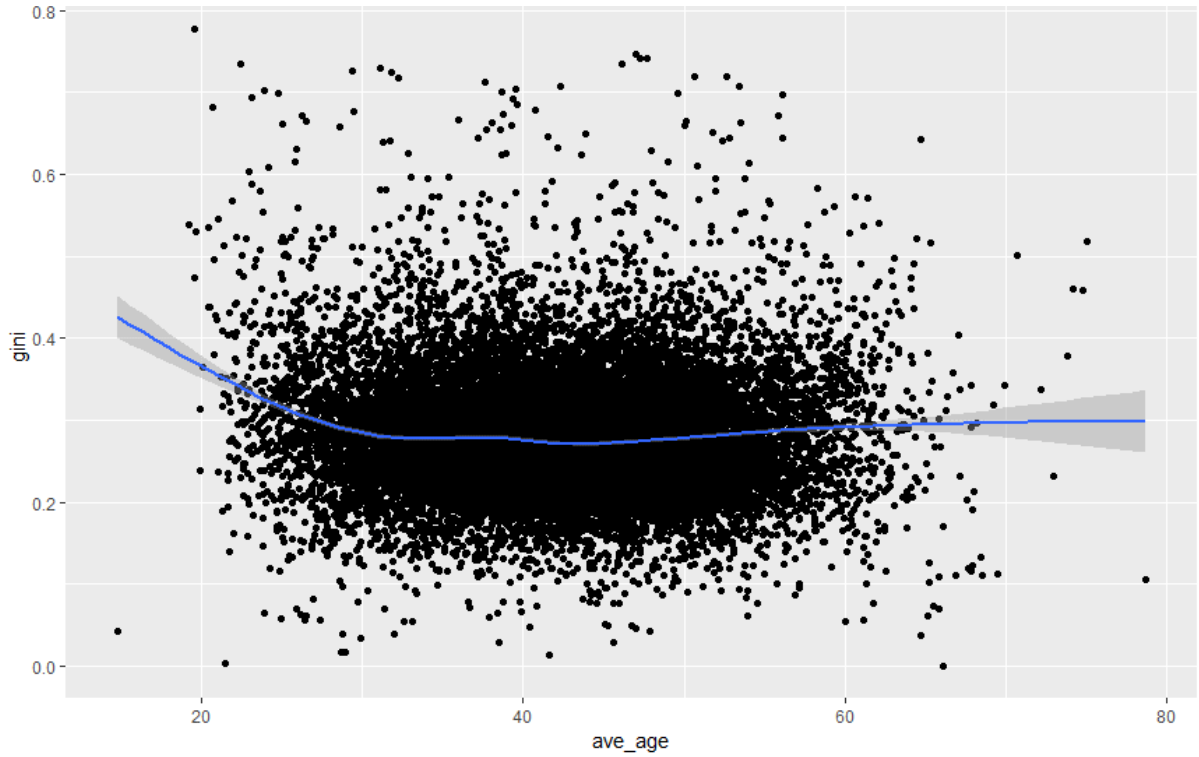

Figure 8: Average age (X-axis) and inequality (Y-axis)

The relationship between wage inequality and the growth of workers are also looked into.

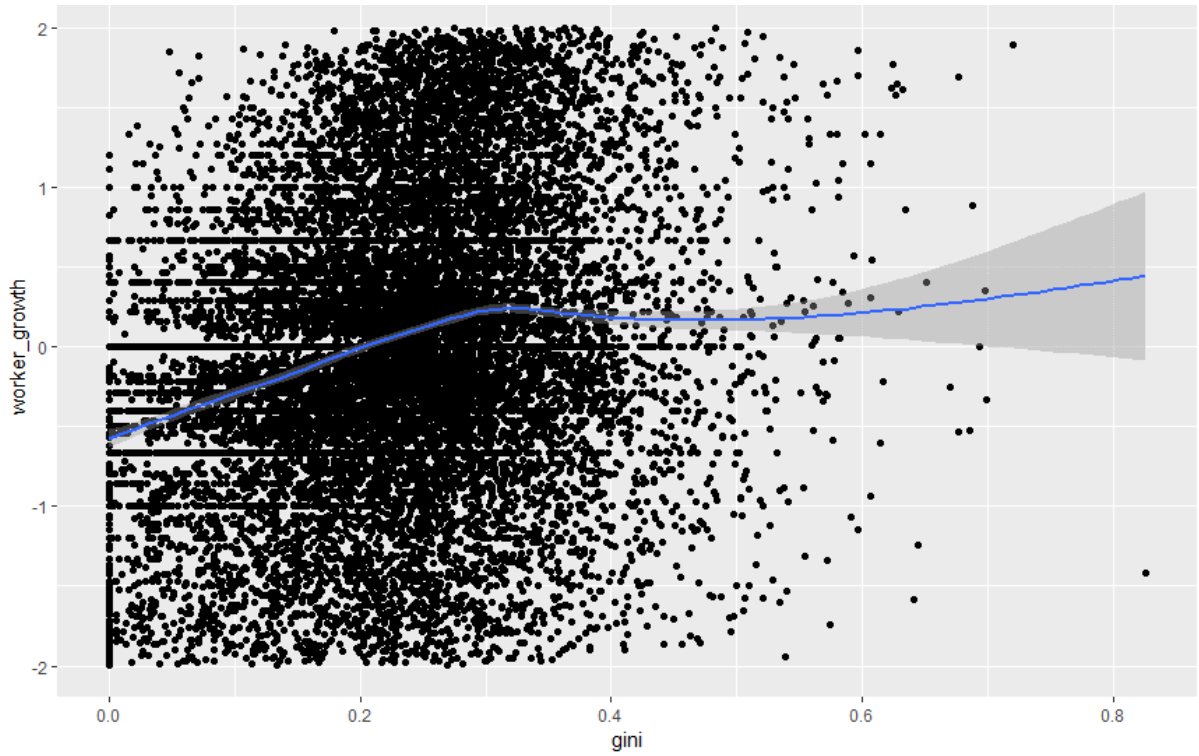

Figure 9: Wage inequality (X-axis) and firm growth percentage (Y-axis)

The yearly growth of companies as in how much has the amount of workers in the company increased or decreased compared to the previous year was calculated during the analysis and compared with wage inequality of the company. Also, we calculate the correlation between inequality and growth for all the firms which was 0.1324, quite a small number. In figure 9 we also see that there is a small positive correlation between these two values.

During our data exploration part, we looked at other relevant numbers that are not related to wage inequality. For example, the number of workers in different sectors is analyzed, as can be seen in figure 10. The graph shows that the manufacturing and retail sectors have 2-3 times more workers than in other sectors. The agriculture, mining, education, health, art/entertainment and other (repair) sectors have more than doubled their worker amount. Manufacturing and construction sectors have lowered their worker amount by 25 percent. This decline of workers in the manufacturing sector makes sense as mentioned earlier in related work.

One key location in the economy of Estonia is its capital, the city of Tallinn. The city has gone through rapid economic and population growth firstly by the Soviet Union and its industrialization of Tallinn which developed the economic output of the capital and brought in a massive amount of workers. Secondly, during the establishment of the independent nation in the 90s, Tallinn again went through rapid economic growth with new companies being established and with young people migrating to the capital from rural areas. This all raised the wages and living standards of people in the city and separated it from the rest of Estonia.

While development and investments are being concentrated in the capital city of Tallinn which is constantly raising the quality of life there, then the opposite is happening in the south-

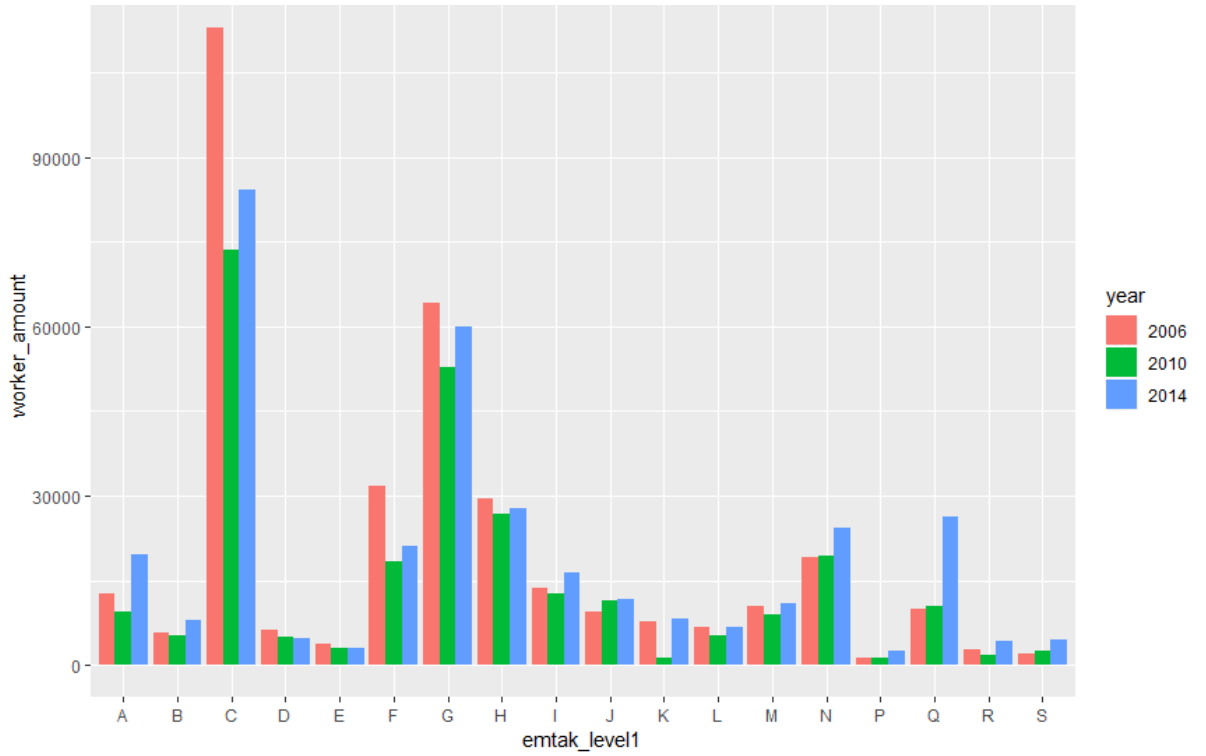

Figure 10: Number of workers (Y-axis) and company sectors (X-axis)

ern and north-eastern regions of Estonia. The problem with the outermost southern regions is that they are far away from any centers of economic development and ports significantly decreasing their development. The north-eastern region mostly consists of the Ida-Virumaa county which is suffering under declining large industry area symptoms. The county has low entrepreneurial activity, high black market activity and a high crime rate. Both south and north-eastern areas of Estonia are described by high unemployment, low wages and by having the poorest quality of life [22]. Overall the wages in Estonia have risen throughout the years, which is logical because the quality of life and incomes in the country has risen a lot because of economic growth after the financial crisis and the constant increase of the minimum wage [13].

Lastly, the relationship between the wages of men and the wages of women are compared with each other. Figure 12 compares the average wages of men and the average wages of women between different sectors. Overall, in our data set men have a higher average wage than women by 18.3 percent in 2014. The largest difference was in the electricity sector(D) with a difference of 26.5 percent and the lowest difference was in the education sector (P) with 4 percent. Fortunately, these differences are an improvement compared to the year 2006 when the difference according to our data was 21.8 percent with the highest difference of 36.4 percent in the financial sector (K) and the lowest difference of 10 percent in the agriculture sector (A). In Estonia, studies have shown that female workers earn a quarter less than their male counterparts. This gender wage gap is higher than in other countries in the European Union [26]. In our data, the gender wage gap is lower than previous research papers because we examine the data of all workers, but previous work looked at Eurostat 2014 data which examined employees in firms

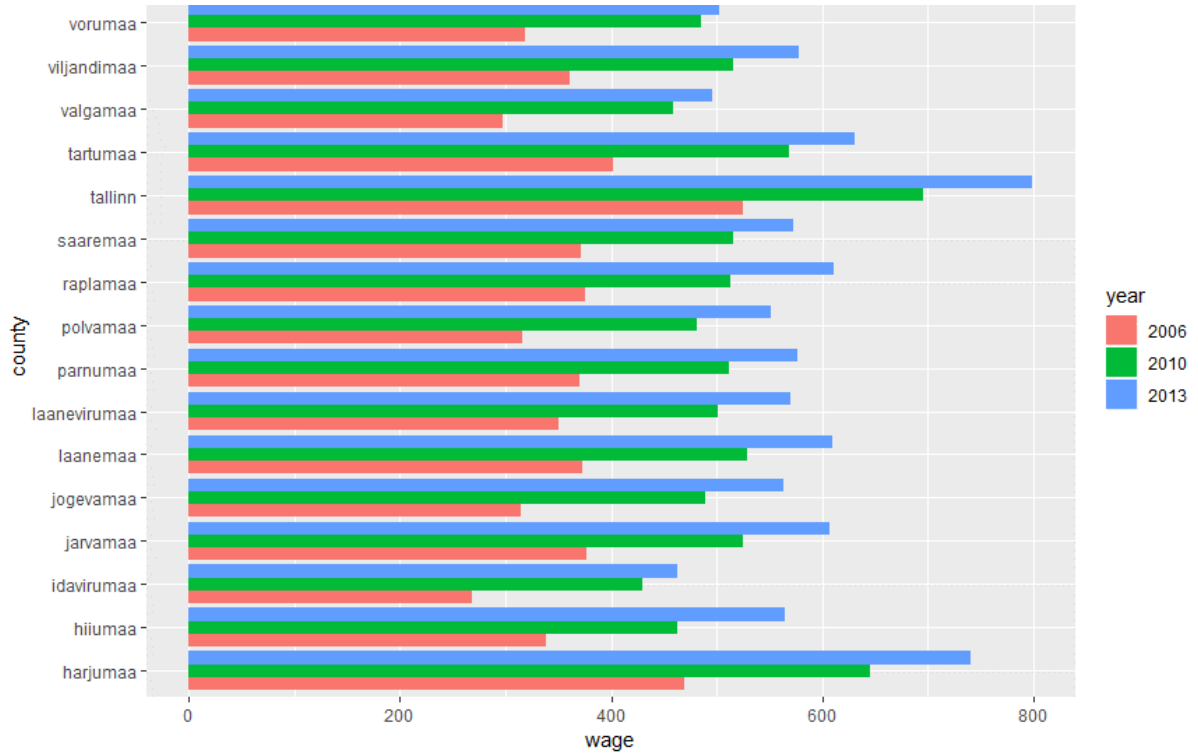

Figure 11: Company county (Y-axis) and average wage (X-axis)

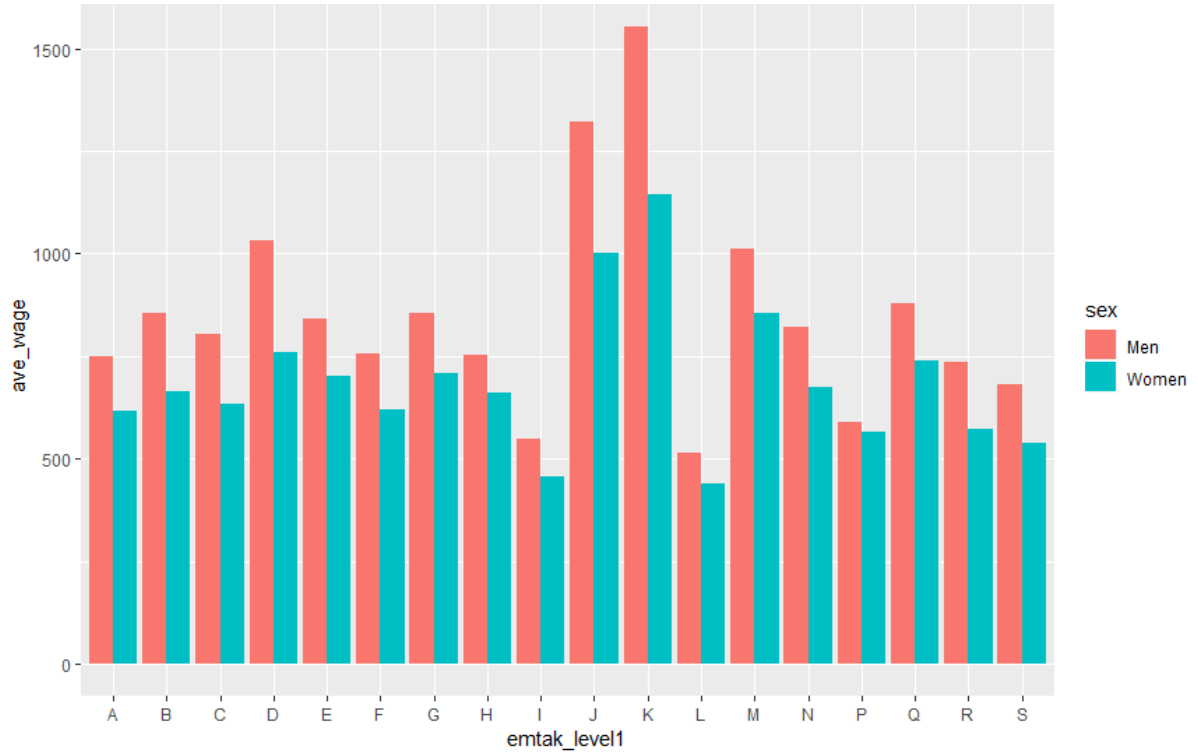

Figure 12: Average wage (Y-axis) and gender in company sectors (X-axis) in 2014

with more than 10 workers. Hence it seems that micro workers have more balanced wages which makes sense. The highest differences between male and female wages are in the IT and finance sectors. This was also mentioned in related work, these sectors are male-dominated and

female workers there earn less [5, 6].

5.2 Predictive analysis results

During the second part of our research, we conduct predictive analysis. The purpose of this analysis is to build many different models with all of the features in the data set to predict the wage inequality in a firm. The coefficients of the features in the model show which features are either significant or insignificant for the inequality of a company. Because linear regression only works with numerical values, we convert categorical values to a set of dummies and include these dummies as control variables in the regression. We also ignore the year, which should not play a role in our prediction model.

Other algorithms are also used on the full data set and the smaller data set, in order to build wage inequality, prediction models. The accuracy of the linear regression model has enhanced through L1 and L2 regularization, these new enhanced models are named separately "Lasso regularized model" and "Ridge regularized model". Additionally, we use boosting algorithms to get better results. Gradient boosting model and extreme gradient boosting models are put into effect. Lastly, a random forest algorithm is also used.

While linear regression and random forest models are "white-box" models that show the importance of its variables, then the other models are "black-box" models meaning we do not see the importance of the variables. The purpose of these models is to get better predictions of wage inequality and limit the number of mismatches. When training linear regression and other models 70 percent of the data set is used as the training set while 30 percent of the data set is used as the test set. The test set is used in the models to predict the inequalities and the results are compared with the real inequalities in the test data set. The differences between predicted inequality and actual inequalities can be used to calculate different metrics that help understand the accuracy of the model. The metric that we look at is MAPE.

The models are built with the training data. The dependent value is the wage inequality and the independent values are the other features. Categorical features like the company sector, company size and company county have been replaced with dummy values. In order to increase accuracy, additional values were added: the natural logarithm of the number of workers, the squared value of average age value and the squared value of the percentage of men. So all-together the following features were used for the models: number of workers, natural logarithm of number of workers, average age of workers, average age of workers, squared value, percentage of men squared value, natural logarithm of the average wage, 4 dummies for the company size classes, 20 dummies for company sectors and 16 dummies for company counties.

Once the model is built with the training data, it is used to predict the test set inequalities. The predicted values are compared with the actual values and the previously mentioned metric is calculated. The attributes of the models are also analyzed. The p-value of the coefficients is evaluated and the coefficients which have a higher p-value than 0.05 are removed because they are not significant in the model. Additionally, the betas of the coefficients and the p-values are analyzed. The beta and the p-value of a coefficient differs depending on the specific coefficient.

The higher the value of the beta and lower the p-value - the higher the significance of the feature in regards to wage inequality. In addition, models are built with a smaller data set. The average wage, wage inequality, percentage of men and the average age in micro-companies vary wildly and do not follow a logical pattern. Micro companies make about 80 percent out of all the companies in our data set, but the number of workers in these companies represents only 13 percent out of all the workers. The models trained and tested with only small, medium and large companies are more accurate and produce interesting results.

5.2.1 Full data analysis results

A normal linear regression model is built with all of the data. The linear regression model calculates the betas and p-values of independent values. In order to avoid multicollinearity, we investigate the correlations between the different independent features of the data set. The correlation values are visualized as a heatmap and this visualization can be seen in figure 13. From this figure we notice many correlations and thus we remove the following features which are correlated with other dependent features, but have a smaller correlation between wage inequality: average wage of workers, average age of workers, average age of male workers, average age of female workers, average wage of male workers, average wage of female workers, percentage of male workers, percentage of female workers.

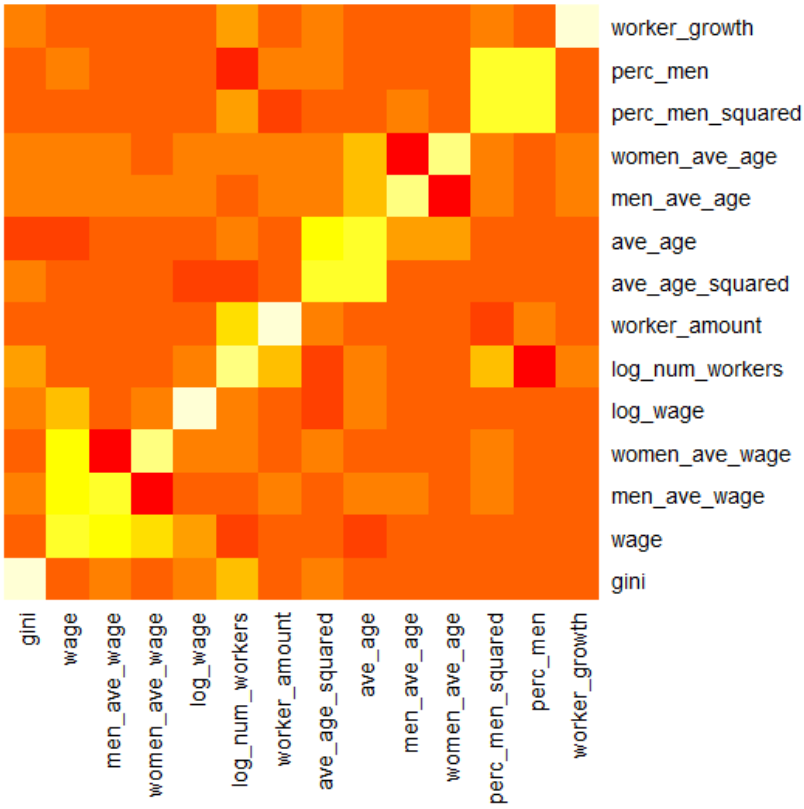

Figure 13: Heatmap of correlations between features of the data

Thus numerical features that are left in the model are the number of workers, the natural logarithm of the number of workers, the average age of workers squared value, percentage of men squared value and the natural logarithm of the average wage. Both the number of workers and its natural logarithm values are kept because the correlation between them was only 0.5. All the dummy values of counties have a p-value over 0.05 except for Tartumaa, Ida-Virumaa, Järvamaa and Põlvamaa. This means that except for these four the company country location is not significant and thus we rebuild the model without these features. In addition, to avoid collinearity between all dummy values, one random dummy is deleted. We remove sector T aka Other economic sectors, a large type of company and Hiiumaa county values from the data. The new values of betas and p-values of coefficients can be seen on table 2.

Feature	Beta	p-value
log(average-wage)	0.036	<2e-16
num-of-workers	-0.001 <	<2e-16
log(num-of-workers)	0.058	<2e-16
squared(average-age)	<0.001	<2e-16
squared(percentage-of-men)	-0.03	<2e-16
firm-growth	**0.004**	**0.0002**
medium-company	0.047	<2e-16
small-company	0.102	<2e-16
micro-company	0.12	<2e-16
county-tartumaa	0.006	3.67e-08
county-polvamaa	0.01	0.000414
county-idavirumaa	-0.012	7.46e-16
county-jarvamaa	0.007	0.014

Table 2: Summary of linear regression model with the full data set

Economic sectors have separate coefficients and p-values which are more or less the same. Coefficient value is between -0.3278 to -0.383 and the p-value is between 0.0002 to 0.001. This means that all economic sectors affect inequality negatively meaning they do no affect wage inequality separately. Features that have high coefficient values are the following: log(average wage), log(number of workers), all firm size dummy features and Ida-Virumaa county dummy feature. What this means is that the average wage of a company and the number of workers in the company play a role in the wage inequality within the company. Also, inequality is affected if the company is located in the county of Ida-Virumaa or Põlvamaa.

The other features are not significant for wage inequality as they do not matter in regards to the value of the wage inequality within a company which means that further analysis of these features is not necessary. Regarding the relation between wage inequality and firm growth, we

see from the linear regression analysis that even though its coefficient is only 0.004 its p-value is strongly below the value 0.05 thus the growth of a company weakly affects the inequality.

5.2.2 Small data set analysis results

In addition, a normal linear regression model is built with the small data set which does not contain data about micro companies. This smaller data-based model has different results than its full-data counterpart. In order to avoid collinearity between dummy values, we remove economic sector O, a large type of company and Läänemaa county from the data. All business sector dummy values have a p-value over 0.05 meaning they do not affect wage inequality and thus we remove them from the model. The values of betas and p-values of coefficients can be seen on table 3.

Feature	Beta	p-value
log(average-wage)	0.034	<2e-16
num-of-workers	-0.001 <	<2e-16
log(num-of-workers)	0.026	<2e-16
squared(average-age)	<0.001	<2e-16
squared(percentage-of-men)	-0.044	<2e-16
firm-growth	**0.003**	**0.008**
medium-company	0.024	<2.46e-08
small-company	0.033	<8.24e-11
county-tallinn	0.013	<2e-16
county-idavirumaa	-0.009	3.91e-06
county-tartumaa	0.01	3.96e-13
county-polvamaa	0.022	2.37e-07
county-jarvamaa	0.013	0.0003

Table 3: Summary of linear regression model with the data set without micro companies

The results are similar to the linear regression analysis with the full data set. The average wage of a company and the number of workers in the company once again play a role in the wage inequality within the company. Additionally, the results show that the wage inequality of companies is a bit higher if the firms are located in the counties of Põlvamaa, Järvamaa, Tartumaa or the city of Tallinn compared to the companies in other counties. Regarding the relation between wage inequality and firm growth the results are basically that same as with the analysis with the full data set - the coefficient is only 0.003 its p-value is strongly below the value 0.05 so in a weak way the growth of a company affects the inequality.

5.3 Analysis results of other models

In our final part of our analysis, we use other algorithms as well to try and predict wage inequality values with the traits of companies. The other algorithms that we use are the following: lasso regularized regression, ridge regularized regression, gradient boosting, extreme boosting and random forest. We build these models with both the full data set and the data set without micro companies. When they are built we calculate different measures of accuracy's for these models in order to see what model has better results and what models have worse results compared to the main linear regression model. The metrics that we compute are the following: root mean square error (RMSE), mean squared error (MSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and R squared (R2).

Models	RMSE	MSE	MAPE	MAE	R2
Linear regression	0.103	0.011	0.445	0.085	0.2507
Lasso regularized	0.103	0.011	0.445	0.081	0.2488
Ridge regularized	0.103	0.011	0.449	0.081	0.2466
Gradient boosting	0.118	0.014	0.462	0.091	0.2243
Extreme boosting	0.135	0.018	0.434	0.108	0.3596
Random forest	0.124	0.115	0.442	0.1	0.2401

Table 4: MAPE values of prediction algorithms with full data set

Models	RMSE	MSE	MAPE	MAE	R2
Linear regression	0.086	0.008	0.259	0.066	0.1073
Lasso regularized	0.086	0.008	0.26	0.066	0.1052
Ridge regularized	0.086	0.007	0.26	0.066	0.1073
Gradient boosting	0.084	0.007	0.267	0.064	0.1
Extreme boosting	0.082	0.007	0.251	0.063	0.179
Random forest	0.082	0.007	0.251	0.063	0.181

Table 5: MAPE values of prediction algorithms with small data set

As can be seen from table 5 the results of the models are more or less the same. Linear regression still stands very well, regularization makes the model worse. Extreme boosting and Random Forest give better results than linear regression but the differences are small.

5.4 Features correlated with inequality

One of the goals of this thesis was to find interesting correlations between with-in company wage inequality and other features of companies. To fulfill this goal the features and inequalities are visualized on graphical representations, their correlations are calculated and a linear

regression model is built in order to see significant features through the betas of the coefficients in the model. The first correlation that we find between the average wage of the firms and wage inequality. In the descriptive analysis part of this thesis, figure 5 shows that the higher the average wage of a company, the higher the inequality within it. The correlation between these two values is 0.214. In the predictive analysis part, the linear regression model says that the average wage is a significant value by having a very low p-value (less than $2e-16$). The beta of its coefficient is <0.001. We can conclude that the average wage in a company is positively but moderately correlated with the wage inequality of a company because the results of the descriptive, correlative and predictive analysis show this as well.

The second weak correlation that we found is the more important focus of this thesis - the relation between firm growth and wage inequality. In the descriptive analysis part of this thesis, figure 9 shows that, at least until a Gini coefficient of 0.3, the higher the growth of a company the higher the inequality within it. The correlation between these two values is 0.214. In the predictive analysis part, the linear regression model says that the average wage is a significant value by having a very low p-value (less than $2e-16$). The beta of its coefficient is <0.004. We can conclude that the average wage in a company is positively but lowly correlated with the wage inequality of a company because of the results of the descriptive, correlative and predictive analysis show this as well

The last moderate correlation that we found is similar to the growth - relation between the number of workers in a firm and wage inequality. In the descriptive analysis part in this thesis, figure 6 shows that for micro, small and medium companies the size of a company is clearly correlated positively with the wage inequality of the company. The correlation between these two values is 0.585. The correlation between these two features is smaller if we take into account other companies aka firms with more than 250 workers. Furthermore, in the predictive analysis part, the linear regression model built with the full data set shows that the number of workers in a company is by itself significant by having a very low p-value and a high coefficient value compared to other coefficients. In the model, the size classes of micro, small and medium firms were also important by having very low p-values. We can conclude that the number of workers in a company is positively but weakly correlated with the wage inequality of a company because the results of the descriptive, correlative and predictive analysis show this as well. This correlation between these two values is very strong for firms with less than 250 workers.

6 Conclusion

The current thesis analyzes the wage inequality with-in firms in Estonia. We look at inequality in companies through various perspectives like company size, county, sector, firm growth, etc. The analysis is performed using a large dataset which consists of 18,000 companies and 378,000 workers from the years 2006 to 2014.

The descriptive analysis part reveals some interesting findings. For instance, wage inequality in Estonia has lowered from 2006 to 2014. In addition, we find that economic sectors vary by average wages and inequalities: the IT sector has high wages, high inequality and a high gender wage gap. Wage inequality has decreased the most in the mining sector from 2006 to 2014. The economic sectors of medicine, electricity, education and mining have high wage inequality while transportation, retail, real estate and agriculture sectors have low inequality. When we analyzed other features of the companies we noticed that there exists a gender wage gap that coincides with previous research. We see that this wage gap exists in all of the economic sectors with an average value of 18.3%.

Furthermore, we observe that when the wages and inequalities of counties are compared with each other then the capital city of Tallinn stands out as the place with the highest wage levels, highest wage inequalities and youngest workers. During the second phase of the analysis - the predictive part - we saw that the average wage of a company, the number of workers employed and some counties play a role in determining wage inequality.

The average wage of a company and it's wage inequality is moderately correlated which can be seen on their figure and by their correlation having a value of 0.274. A similar low correlation exists between the size of the company and its wage inequality. The correlation is much stronger for companies with less than 250 workers.

Regarding the correlation between inequality and growth, the results of our analysis show a correlation with a value of 0.1324 and this weak correlation is also witnessed from the descriptive and predictive analysis part. When conducting the descriptive analysis the coefficient of growth in the linear regression was quite low. During the predictive analysis, we implemented other algorithms like Lasso, Ridge, Gradient Boosting, Extreme Boosting and Random Forest to predict wage inequality with Estonian firm data. While extreme boosting and random forest algorithms were useful in getting more accurate results, the other models produced the same or worse results. Our results show that wage inequality is possible to predict with machine learning algorithms. The average mean absolute percentage error (MAPE) of models built with the whole data set was 0.44 while the MAPE of models built with the data set without micro companies was on average 0.25, which is a good metric.

In conclusion, our findings support the results of Mueller, Ouimet and Simintzi (2017) who also found that wage inequality is positively correlated with firm size and growth. The correlation in our research is not very strong, so even though it exists in one way or another, this correlation should not affect the policies of companies and governments. More research has

to be conducted on the subject of different wage inequality levels in different sectors. Investigating the reasons behind low or high wage inequalities would be interesting in an academic standpoint and would also have good implications for policymakers. We show here evidence that it is possible to predict wage inequality with-in firms that may paw the way for better and more accurate prediction methods.

We have several suggestions for further research in this area. Taking into account the results of this thesis we recommend a few different perspectives on what specific aspects should be looked at. First of all, a very limited amount of academics have researched wage inequality with-in specific economic sectors and looked into the economic and social reasons for why inequality is high or low in this specific sector. As could be seen on this thesis, income inequality in different sectors of Estonia varies and we plan on researching the reasons behind this. Secondly, we plan to look more into prediction regarding wage inequality. Throughout this thesis by using the traits of companies we built predictive models using different algorithms. These algorithms produced different results and we believe that trying to predict wage inequality through machine learning is a rare subject that needs to be researched more. Other algorithms and firm traits could be used in order to build predictive models that produce better results. Lastly, the secondary goal of this thesis was to find a correlation between income inequality and firm growth and we found a weak correlation between these two traits. Previous research has shown that this relation is stronger so in the future we plan to investigate this correlation once again by using other ways of measuring inequality like the Theil index or mean log deviation.

References

[1] A. Akerman, E. Helpman, O. Itskhoki, M.-A. Muendler, and S. Redding. Sources of Wage Inequality. 103(3):214–219, 2013. doi: 10.1257/aer.103.3.214.

[2] J. Alvarez, F. Benguria, N. Engbom, and C. Moser. Firms and the Decline in Earnings Inequality in Brazil. *American Economic Journal: Macroeconomics*, 10(1):149–89, January 2018. doi: 10.1257/mac.20150355. URL `http://www.aeaweb.org/articles?id=10.1257/mac.20150355`.

[3] S. J. Babones. Income inequality and population health: correlation and causality. *Social science & medicine*, 66(7):1614–1626, 2008.

[4] R. J. Barro. Inequality and Growth in a Panel of Countries. 5:5–32, 2000. doi: 10.1023/A:1009850119329.

[5] A. Belgorodskiy, B. Crump, M. Griffiths, K. Logan, R. Peter, and H. Richardson. The gender pay gap in the ICT labour market: comparative experiences from the UK and New Zealand. *New Technology, Work and Employment*, 27(2):106–119, 2012.

[6] B. Bell and J. Van Reenen. Bankers' pay and extreme wage inequality in the UK. 2010.

[7] J. Bernstein. The Impact of Inequality on Growth. `https://www.americanprogress.org/issues/economy/reports/2013/12/04/72062/the-impact-of-inequality-on-growth/`. Accessed: 2019-05-14.

[8] G. Bertola and A. Ichino. Wage inequality and unemployment: United States vs. Europe. *NBER macroeconomics annual*, 10:13–54, 1995.

[9] J. A. Campos-Soria, B. Ortega-Aguaza, and M. A. Ropero-Garcia. Gender segregation and wage difference in the hospitality industry. *Tourism Economics*, 15(4):847–866, 2009.

[10] R. Eamets. Labour Market in Estonia: Responding to the Global Finance Crisis. *CESifo DICE Report*, 10(2):34–39, 2012.

[11] R. Eamets and E. Kalaste. The lack of wage setting power of Estonian trade unions? *Baltic Journal of Economics*, 5(1):44–60, 2004.

[12] M. G. Entrepreneurship. Entrepreneurship at a Glance, 2011.

[13] S. Ferraro, J. Meriküll, and K. Staehr. Minimum wages and the wage distribution in Estonia. *Applied Economics*, 50(49):5253–5268, 2018.

[14] D. A. Fleming and T. G. Measham. Income inequality across Australian regions during the mining boom: 2001–11. *Australian Geographer*, 46(2):203–216, 2015.

[15] K. J. Forbes. A reassessment of the relationship between inequality and growth. *American economic review*, 90(4):869–887, 2000.

[16] A. Hyder and B. Reilly. The public and private sector pay gap in Pakistan: A quantile regression analysis. *The Pakistan Development Review*, pages 271–306, 2005.

[17] T. Kasekamp. A Web Application to Support Researchers in Predictive Process Monitoring Tasks. 2018.

[18] T. Lemieux. The changing nature of wage inequality. 21(1):21–48, 2008. doi: 10.1007/s00148-007-0169-0. URL `https://app.dimensions.ai/details/publication/pub.1048006222`.

[19] K.-O. Leping and O. Toomet. Emerging ethnic wage gap: Estonia during political and economic transition. *Journal of comparative Economics*, 36(4):599–619, 2008.

[20] J. Masso, J. Meriküll, and P. Vahter. Gross profit taxation versus distributed profit taxation and firm performance: effects of Estonia's corporate income tax reform. *The University of Tartu Faculty of Economics and Business Administration Working Paper*, (81-2011), 2011.

[21] H. M. Mueller, P. P. Ouimet, and E. Simintzi. Wage Inequality and Firm Growth. 107(5): 379–383, 2017. doi: 10.1257/aer.p20171014.

[22] G. Raagmaa. Shifts in regional development of Estonia during the transition. *European Planning Studies*, 4(6):683–703, 1996.

[23] S. Sarkar and B. Singh Mehta. Employment profile of ICT sector in India. 2008. URL `https://www.researchgate.net/publication/262425937_Employment_Profile_of_the_ICT_Sector_in_India`.

[24] J. Song, D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter. Firming up inequality. Working Paper 21199, National Bureau of Economic Research, May 2015. URL `http://www.nber.org/papers/w21199`.

[25] K. Stæhr and E. Pank. Economic Developments in the Baltic States: Success and New Challenges. *Danmarks Nationalbank Monetary Review 4th Quarter*, 2007.

[26] K. Vassil, R. Eamets, and P. Mõtsmees. Socio-demographic model of gender gap in expected and actual wages in Estonia. 2014.

[27] C. Zhu, G. Yang, K. An, and J. Huang. The leverage effect on wealth distribution in a controllable laboratory stock market. *PloS one*, 9(6):e100681, 2014.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kevin Kanarbik**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   An empirical investigation on wage inequality in Estonian firms,

   supervised by Rajesh Sharma and Jaan Masso.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Kevin Kanarbik
16/05/2019