VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Magistro baigiamasis darbas

VEIKSNIŲ, LEMIANČIŲ PAJAMŲ NELYGYBĘ EBPO ŠALYSE, EKONOMETRINĖ ANALIZĖ

Econometric analysis of the factors determining income inequality in OECD countries

Edgaras Juodsnukis

Vilnius, 2020

MATEMATIKOS IR INFORMATIKOS FAKULTETAS
TAIKOMOSIOS MATEMATIKOS INSTITUTAS
STATISTINĖS ANALIZĖS KATEDRA

Darbo vadovas doc. Jurgita Markevičiūtė _____

Darbas apgintas _____

Registravimo NR. _____

Contents

Santrauka

Veiksnių, lemiančių pajamų nelygybę EBPO šalyse, ekonometrinė analizė

Šiame darbe nagrinėjamas ryšys tarp pajamų nelygybės ir socialinių bei ekonominių rodiklių EBPO šalyse 2007 - 2016 m. laikotarpiu. Teorinėje darbo dalyje apžvelgiama pajamų nelygybės mokslinė literatūra leido veiksnius, lemiančius pajamų nelygybę, suskirstyti į penkias kategorijas. Remiantis šiuo suskirstymu, tyrime naudojami kintamieji aprėpia ekonominio augimo arba ekonominio vystymosi, demografijos, politinius, kultūros bei aplinkos ir makroekonominius veiksnius. Praktinėje tyrimo dalyje iš pradžių sukonstruojami jungusis ir vienpusiai savųjų veiksnių paneliniai modeliai visoms šalims. Modelių liekamosios paklaidos netenkino panelinių duomenų regresinės analizės prielaidų, o modelių struktūra, kurioje posvyrio koeficientai visoms šalims yra vienodi, nepadeda identifikuoti reikšmingų veiksnių, pajamų nelygybę veikiančių skirtingose šalyse. Todėl norint pasiekti darbo tikslą, kaip alternatyva kiekvienai šaliai skirtingų posvyrio koeficientų vertinimui buvo pasiūlytas klasterinės analizės ir panelinių modelių bei kintamųjų atrankos metodų derinys. Atlikus klasterinę analizę, kuri leido šalis sugrupuoti pagal jų geografinę padėtį bei jose veikiančius socialinius modelius, pajamų nelygybė buvo modeliuojama kiekviename klasteryje. Klasterizuojant naudojamas K-vidurkių metodas leido išspręsti paklaidų heteroskedastiškumo ir autokoreliuotumo problemas. Nors dvigubas ir dalinis kintamųjų atrankos metodai ne visais atvejais sumažino parametrų įverčių standartines paklaidos, tačiau leido identifikuoti statistiškai reikšmingą tiesinį ryšį tarp BVP augimo ir pajamų nelygybės Šiaurės Europos ir anglosaksų šalyse bei patvirtino, jog keturiuose iš šešių klasterių didesnės mokestinės pajamos yra susijusios su mažesne pajamų nelygybe.

Raktiniai žodžiai: pajamų nelygybė, paneliniai modeliai, daliniai kintamųjų atrankos metodai, LASSO regresija, klasterinė analizė

Abstract

Econometric analysis of the factors determining income inequality in OECD countries

This paper examines the relationship between income inequality and socioeconomic indicators in OECD countries in the period of 2007 - 2016. The theoretical part of the thesis reviews the scientific literature on income inequality, which allowed to classify the factors determining income inequality into five categories, that might be described as: economic growth or economic development, demographic, political, cultural - environmental and macroeconomic factors. In the practical part of this work, firstly, pooled and one-way fixed effects panel models are constructed for all countries. The residual errors of the models has not met the assumptions of the regression analysis of the panel data, and the structure of the models, where the slope coefficients are the same for all countries, does not help to identify significant factors affecting income inequality across different countries. Therefore, in order to achieve the aim of this work, a combination of cluster analysis, panel modelling and variable selection methods have been proposed as an alternative of the estimation of the different slope coefficients for each individual. After clustering that allowed countries to be grouped according to their geographical location and the social models operating in them, income inequality were modelled separately in each cluster. The K-means method used in clustering allowed to solve the problems of residuals heteroscedasticity and autocorrelation. Although double and partial selection methods has reduced standard errors of parameter estimates not in all cases, however these techniques helped to identify a statistically significant linear relationship between GDP growth and income inequality in North European and Anglo-Saxon countries and confirmed that higher tax revenue is associated with lower income inequality in four of the six clusters.

Key words: income inequality, panel models, partial variable selection methods, LASSO regression, cluster analysis

INTRODUCTION

Over the last three decades rising in income inequality in most developed countries in the world has become the subject of increasing debate. Supply-side policy based on neoliberal ideas have failed to create an equal opportunities society, even in countries with high levels of development. Some of the regained and newly established democracies, including Lithuania, have turned in the direction of a neo-liberal model of economic development, which in the long run has led to declining financing for public services and the failure of the state to fulfill its commitments to its citizens. The crisis in the public finance identity has led to tolerance of the shadow in health system, a growing gap between regions, which is characterized by inequality of access to public goods. Therefore, the growing sense of injustice and the social exclusion of a significant part of the population in developed countries pose new threats to the economic, social and political stability of these countries, that are the key factors to ensure peace and stability worldwide.

In recent years, researches of the OECD and EU countries on income inequality examines the impact of one or more variables on income inequality, without strictly distinguishing between groups of factors determining income inequality in different directions. Furthermore, most of the researches on income inequality are performed by modelling without clustering countries according to their socio-economic indicators, therefore not taking into account of the heterogeneity of countries, may led to the loses of the ability to identify factors that are significant in the groups of countries. Thus, in order to identify the main factors of income inequality across different groups of countries, new and more detailed studies are needed to model income inequality in groups obtained by clustering countries according to their socio economic indicators.

Aim. The main purpose of this work is to identify the influence of different groups of factors on income inequality in different clusters of OECD countries.

Tasks. The aim of master thesis was achieved by accomplishing the main tasks of the work:

1. To complete a comprehensive analysis of income and social inequality and distinguish between separate factor groups that affect income inequality differently;
2. To fit pooled and fixed effects panel models on income inequality for non-clustered data of the OECD countries;
3. To cluster countries according to their social and economic indicators into a selected number of groups, which were determined by literature analysis;
4. To fit pooled and fixed effects models and variable selection methods based on LASSO regression in order to estimate income inequality in different clusters.

The first section reviews the concept of income inequality, its negative impact on societies and economies, the most popular indicators of income inequality, the factors that influence economic inequality, current economic inequality situation in Lithuania and some inequality studies, which includes studies that use machine learning algorithms. The second chapter presents panel models and their assumptions, statistical tests and hypothesis used in modelling panel data, LASSO regression including double selection and partialling-out approaches and cluster analysis in terms of K-means algorithm. The third section includes methodology, used in this thesis and modelling results. The methodology presents variables that were used in models and cluster analysis, economic and econometric models of income inequality and proceed of modelling. Meanwhile, the second part of the third chapter includes results of modelling both non-clustered and clustered data and cluster analysis. Residuals analysis of panel models for non-clustered data and robust standard errors of panel models for each cluster are provided in the Appendix.

1. State of the art

1.1. The concept of income inequality

The Organisation for Economic Co-operation and Development (OECD) describes income inequality as uneven distribution of financial goods in societies (OECD, 2005). In scientific literature, the most frequently used concepts describing inequality are economic inequality (Allub & Erosa, 2019; Obolenskaya & Hills, 2019), social inequality (Raudenbush & Eschmann, 2015) and income inequality (Tao et al., 2019). Although inequality and social topics have always been a part of most discussions about economy, many sociology researchers recognize E. Durkheim (1858 – 1917) as one of the first to attempt to explain income inequality (Riley, 2014). However, it is important to identify not only the magnitude of income inequality, but also tendencies of its growth and its determining factors, which could help in identifying instruments of economic policy that are more effective in reducing increased unevenness of income distribution. There is almost unanimous agreement amongst economists that income inequality has increased over the last few decades, but their views on the solution to this problem differs widely (Cingano, 2014; Stone, Trisi, Sherman, & Debot, 2015). As a result, there is currently considerable debate as to what level of income inequality could be tolerated and which instruments are the most effective in reducing income inequality.

Some authors argue that economic inequality often leads to social exclusion and inaccessible health services and can thus make a significant contribution to increasing inequality of opportunity and risk of poverty (Wright, 2000). In other studies, income inequality is identified as a negative factor, which may be associated with increased crime rates, less trust in other community members and civic activity, or even a lack of loyalty to their country (Hanson, 2013; Lancee & Van de Werfhorst, 2012). If the trends were to grow, high income inequality countries, either democratic or autocratic, might have to face social and political uneasiness that could affect the relative stability of Europe as a whole, or even of the world. For this reason, the central government tends to respond to growing income inequality by avoiding conflict with the protesters and redistributing parts of tax revenue through social programs, which are relevant in the country at the time. In addition, International Monetary Fund (IMF) economists claims that higher income inequality reduces investment and economic growth and can be one of the reasons of economic, policy and financial instability (Dabla-Norris, Kochhar, Suphaphiphat, Ricka, & Tsounta, 2015).

When analyzing the causes of economic inequality, it is also important to consider possible types of income, as the level of income inequality may also be affected. In most cases, income is defined in three different ways: I) income from economic activities; II) gross income; III) disposable income. Income from economic activities includes activities related to the employment relationship, freelancers or business. In addition to the aforementioned income

from an individual economic activity, gross income also includes benefits, such as pensions, health or unemployment benefits, and other allowances that are awarded when the income is below a certain level or equal to basic universal income, such as allowances for minor children in the family. Meanwhile, an individual's disposable income is the balance of income after taxes on gross income. From the point of view of economic well-being, this is one of the most important indicator of individuals' actual income and the existing inequality in the society in question (Weinberg, 2004).

1.2. Measures of income inequality

As there is no universal definition of income inequality, there are many indicators, which measure income inequality. However, each of these indicators has its own advantages and disadvantages and none of them are suitable for all cases (Hoeller, Joumard, Bloch, & Pisu, 2012). In the last few decades, the most common indicators of income inequality are the Gini index and percentile ratios (e.g. deciles / quantiles or quartiles ratios). In the first case, income inequality is determined by measuring the total income distribution, while the second group of indicators reveals income inequality only at certain points of income distribution.

1.2.1. Gini index

Gini index, as a measure of the inequality in the distribution of population income (expenditure), was proposed by the Italian statistician Corrado Gini (Ceriani & Verme, 2012). Values for this index ranges from 0 to 1, where a zero value represents complete income equality, i.e. income of all population is the same, and a unit represents total income inequality, i.e. all income is generated by one person. Marginal values of the Gini coefficient are not specific to any of the countries in the world, and as researchers notes, index values tend to fluctuate between 0.20 and 0.65 (Gasparini & Lustig, 2011). Africa, Latin America and some Asian countries currently have the highest income inequality, where Gini index values range from 0.5 to 0.7. Meanwhile, in most of the highly-developed countries, the rate varies from 0.2 to 0.35. Although there is no consensus on the generally acceptable range of the Gini index, economists agree that a value of less than 0.25 indicates that income inequalities are less significant in terms of poverty and social exclusion. Gini coefficient values which are greater than 0.25 but less than 0.35 correspond to the average income inequality and indicate that income distribution has a significant impact on the risk of poverty and social exclusion. When this indicator is above 0.35, the importance of structural reforms to reduce income inequality is underlined. In addition, a European Commission study proposed that a high level of income inequality can have a significant negative impact on public budgets and, subsequently, on rising public debt (Larch, 2012). Therefore, it is easier to ensure fiscal discipline in countries with relatively more equal societies.

Since the Gini index is a relative size, one of its key benefits is the ability to compare

the extent of income inequality across countries and regions at different times. On the other hand, evaluating individuals and households produces different results, showing different levels of income inequality.

1.2.2. Lorenz curve

The Lorenz curve is used in practice in order to visually assess the inequality of income distribution between different entities and to calculate the Gini index (Gastwirth, 1972). Thus, Lorenz curve is a graphical representation of the distribution of income in a given society.

The distribution of the population is represented by the horizontal axis, while the distribution of income corresponds to the vertical axis. In the case of an ideally distributed income, the Lorenz curve would show absolute equality and be inclined at 45°. The further the Lorenz curve moves away from the 45° diagonal, the greater the income inequality it displays. The coefficient can be calculated by dividing the area of the bounded figure of absolute equality and the Lorenz curve by the lower rectangle bounded by the line of absolute equality, i.e. $y = x$, $y = 1$ and $x = 1$, area.

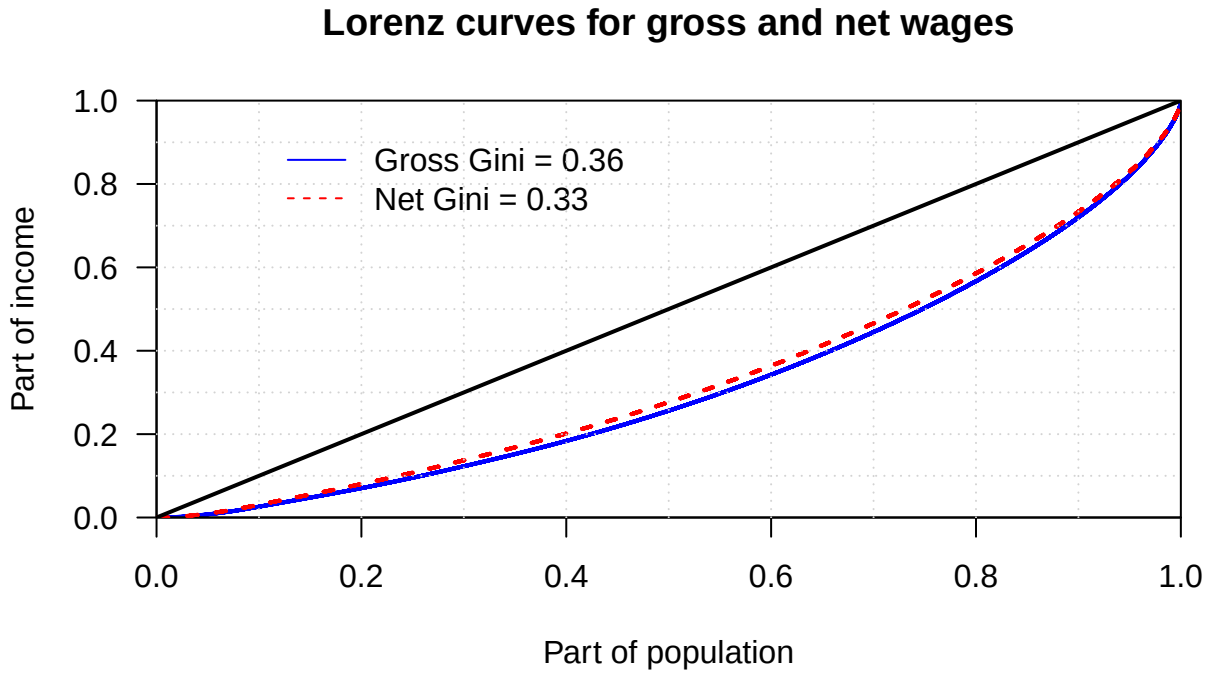

Figure 1. Lorenz curves for gross and net wages in Lithuania, according by data of SODRA September, 2019.

Here Figure 1 presents Lorenz curve for gross and net wages in Lithuania, according to the SODRA (State Social Insurance) data in September, 2019. This graph can be interpreted, for example, in the following way: 50 percent of the population earns slightly more than 25

percent of total employment-related income in case of the wages before taxes. The obtained Gini coefficient in gross wages employment-related cases in Lithuania is 0.36. Since it is generally accepted that the tax system usually contributes to reducing income inequality (Agnello & Sousa, 2014), in this case income inequality of net salaries (employment-related income exclusing taxes) is more important. Thus, under the current tax regime in Lithuania, employment-related income Lorenz curve has shifted closer to the line of equality. This suggests that inequality of net salaries is lower than the gross income inequality in case of employment-related incomes. In this case, the Gini coefficient is approximately 0.33. Therefore, calculations confirm that net income inequality is lower than gross income inequality.

1.2.3. Other indicators

Also a relatively common using measure of income inequality is decile ratios (Sila & Dugain, 2019). In this case, the income of individuals, ordered by ascending or descending, is divided into ten equal parts. In this method, for example, the first decile corresponds to 10 percent of the lowest income earners. The decile ratio is calculated by dividing the tenth decile by the first decile. Thus, it shows the relative size of the tenth and the first deciles. Higher values of this indicator represent a wider gap between the highest and lowest income individuals, which is in line with the higher income inequality. Some researchers and international organisations are prone to distinguishing this indicator, due to its ability to identify income differences at the margins of income distribution (Cingano, 2014). However, estimating by decile ratios, weights of observations are not taken into account. Therefore, individual data is needed to accurately calculate the extent of inequality. Furthermore, this indicator does not provide an opportunity to measure differences in the middle of income distribution.

Equally important and frequently used indicators for measuring income inequality are quartiles and quantiles. The estimation of income inequality is almost identical to the way of the decile ratios, but in calculating the quartiles ratio, individual‘s income is divided into four equal parts and the quantiles ratio is divided into five equal parts. These indicators are often measured by national statistics departments while performing surveys of households or by international organizations in cross-national researches.

Summarizing the indicators for measuring income inequality, it can be stated that the choice of measure often depends on the purpose of the study. In addition, because inequality assessment indices are constructed differently, they can be used to measure different aspects, such as regional or country income inequality.

1.3. Factors determining income inequality and its changes

After reflections on the concept of income inequality, its impact on societies and indicators, the factors that determine its occurrence and change should be examined in more

detail. Scientists in their studies identify different causes of income inequality, but most agree that inequality and poverty are influenced by both endogenous and exogenous factors.

1.3.1. The endogenous factors of income inequality

Previous studies have shown different endogenous causes of income inequality, however, the intellectual and physical characteristics of human beings play a major role in uneven income distribution (Charles-Coll, 2011). Research has shown that respondents with higher IQ's had higher career opportunities. In addition, disability has been identified as a major contributor to poverty and inequality in low and middle income countries (Banks, Kuper, & Polack, 2017). Meanwhile, many societies, including countries such as the United States, still have gender and racial equality problems, which lead to lower levels of employment and lower income for women and foreigners (Charles-Coll, 2011). The endogenous determinants of income inequality will not be explored further in this thesis, though literature analysis suggests that at the micro level, income inequality is largely influenced by endogenous causes, which are more or less reflected in country income inequality indicators.

1.3.2. The exogenous factors of income inequality

Most often the exogenous factors of income inequality are not distinguished between different groups. However, throughout the past century some economists have tried to summarize and group together the determinants of income inequality into more tightly defined groups. For example, International Monetary Fund experts have defined economic, educational and environmental groups (Dabla-Norris et al., 2015), while other researchers have identified economic and interdependent factors of globalization (Bakija, Cole, Heim, et al., 2012). The scientific literature on income inequality also suggests stricter proposals, distinguishing all factors to five groups: economic development, demographic, political, cultural - environmental and macroeconomic factors (Kaasa, 2005).

Economic development. Economic development and its quantitative variables as national income (usually measured as GDP per capita), economic growth or technological development are the most frequently discussed factors that effects income inequality. One of the first studies on the relationship between economic growth and income inequality is based on the inverted U hypothesis: as the GDP grows, income inequality initially increases, but then starts to decrease (Kuznets, 1955, 1963). Thus, the author constructed a hypothesis, which states that in countries with lower levels of development, income inequality is rising due to the early stages of transition, and at some point inequality starts to decline. At the end of the last century, there were many attempts to test the Kuznets hypothesis on the basis of different growth and income data and most of these practical studies confirmed theory (Barro, 1999; Nielsen & Alderson, 1997).

On the other hand, there is a different set of ideas that was followed by some of the most prominent economists of this time, such as Paul Krugman and Joseph Stiglitz, who were influenced by neo-Keynesianism. The latter strongly criticizes the idea that income inequality tends to decrease in countries with high levels of development (Stiglitz, 2016). Hence, Stiglitz takes a different view, based on trends of income inequality and the income of affluent populations, which have increased dramatically in recent decades. For example, in the US, over the last three decades, the income of the 1% richest people has risen 1.42 times and the 0.1% even 2.36 times. In addition, average household income increased by only 9%, while productivity increased by three-quarters. As the Nobel laureate in economics notes, wealth in the US is more concentrated than income, with the richest 1% of Americans owning about 35% of the country's wealth.

Thus, the Kuznets hypothesis cannot explain the paradoxes that have arisen in high-growth countries over the last few decades, where income inequality has grown faster than anywhere else. Some critics of the neoclassical economy emphasize the imperfection of marginal productivity theory as a possible cause of such inequality growth (Stiglitz, 2016; Syll, 2014). The ideas of neoclassical economy maintain that, due to competition, everyone participating in the production process earns a remuneration equal to her or his marginal productivity. Part of the Keynes followers, called New-Keynesians, advocate for the one percent and criticize high progressive tax rates by arguing that taxing high incomes would deprive them of the 'just deserts' for their contribution to society, and, even more importantly, it would discourage them from expressing their skills (Mankiw, 2013). Although both freshwater and saltwater economists admit there is a bit of truth in these statements, the latter emphasize that inequality from the growth perspective is also an outcome of exploitation, discrimination, and rising of monopoly powers. Furthermore, in conclusions of previous studies arguing, that excess profits and rent-seeking have significantly influence income distribution in high development countries (Bakija et al., 2012; Bivens & Mishel, 2013). In his article, Stiglitz concludes that perceptions of the relationship between economic development or economic growth and income inequality need to change, and GDP is not an appropriate measure of humanity's well-being (Stiglitz, 2009). As a more effective tool, the scientist proposes active economic policy: bigger investments in public goods, stronger workers rights, more progressive taxation system and transfer policies.

Demography. Another group of factors that influence income distribution are demographic parameters: urbanization, structure of household, population age structure, and education level (Kaasa, 2005). Previous studies have produced contradictory findings regarding the influence of urbanisation (Crenshaw, 1993; Litwin, 1998) and population age structure (Deaton & Paxson, 1997; Higgins & Williamson, 2002) on income inequality. Such controversies can be explained: although urbanisation is associated with better possibilities for advanced social organisation, it does not evaluate unstable growth concepts, discussed in

the previous chapter. In the case of age structure, it should be admitted that older people usually earn higher wages due to their huge experience in a particular field or sector, and that decreases income inequality. On the other hand, changes in family planning habits in countries with a higher level of economic development lead to lower birth rates and an aging population, which result in higher expenditures for social protection.

Over the last few decades, the link between educational attainment and income inequality or risk of poverty has been debated extensively. For example, a cross-sectional study of 50 countries concluded that countries with a lower average year of schooling noted with higher income inequality (Sylwester, 2002), while economic inequality was lower in states with more average years of education (Partridge, Partridge, & Rickman, 1998).

More recent research into these interactions revealed that the growing trend towards migration can dramatically increase the number of people at risk of poverty. At the same time, the reciprocity between life expectancy, household structure and educational attainment suggests that policy makers will need to take into account the interaction between different demographic factors and their combined impact on income inequality in Europe over the coming decades, in order to avoid greater social exclusion (Guerin, 2013).

Policy. Policy factors are inseparable from the countries‘ economic situation, and therefore affect the distribution of income of the population. Most commonly used indicators to quantify the influence of government on a country's economy are share of government expenditure, tax revenue and social spending in the GDP (Kaasa, 2005; Lazutka, 2003).

A panel analysis of the effect of government social spending on income inequality in OECD countries showed that increases in public social spending result in a decrease in the Gini index. In addition, secondary school enrollment rate, unemployment rate in the civilian labor force and population growth rate resulted in significantly bigger influence effects on income inequality (Ulu, 2018). Other analysis of the relationship between public sector size (share of tax revenue in GDP) and income inequality concluded that higher income inequality characterises autocracy or limited democracy countries, while fully institutionalized democracies reduced the income inequality (Lee, 2005).

However, according to some economists, the level of democracy does not always mean lower income inequality: it is also important which welfare state model is used by economic policy makers (Lazutka, 2003). There are four standard social welfare models: the Nordic countries (Denmark, Sweden, Finland and Norway) attributed to the institutional model of social welfare, the countries of the former British Empire (United Kingdom, United States of America, Ireland, Australia) adapted to the Anglo-Saxon model of the welfare state, countries of the Southern Europe where the Catholic Church retained its influence (Italy, Greece) belong to the Catholic model, while countries in the center of the Western Europe (Germany, Austria, France, Netherlands, Belgium) are considered to be the continental Europe model states. According to extensive studies of welfare states, the highest income,

social and health inequalities occur in the Anglo-Saxon group with the lowest shares of tax revenue and social security expenditure in the GDP (Eikemo, Bambra, Joyce, & Dahl, 2008; Lazutka, 2003). The Continental model, meanwhile, has the lowest income-related health inequalities, while the Scandinavian countries with the highest public budgets relative to the economy and the highest shares of social protection expenditure in the GDP recorded the lowest income inequality.

Cultural and environment. Some authors also highlight cultural and environmental factors that influence income inequality in one way or another. In this context, cultural and environmental factors such as abundance of natural resources, shadow economy and corruption are seen as indirectly contributing to income inequality.

Studies of corruption and shadow economy in Latin American and Asian countries have shown that these factors have a significant impact on economic inequality (Dobson & Ramlogan-Dobson, 2012; Kar & Saha, 2012). Contrarily, informal economy can distort official statistics and does not have a dramatic impact on income inequality in developing countries (Dobson & Ramlogan-Dobson, 2012). The effect on the response is negative (i.e. lower inequality) and significant when the share of shadow economy in the GDP interacts with the corruption index. One of the reasons can be that the unfold of the shadow economy within Asia might have raised earnings among people which otherwise would not have changed (Kar & Saha, 2012).

The relationship between natural resources and their abundance, land concentration and income inequality is another object of discussion (Kaasa, 2005). Some studies claim that the abundance of natural resources, which is usually associated with high land concentration of ownership and rent, increases income inequality (Gupta, Davoodi, & Alonso-Terme, 2002). Furthermore, cross-country research concluded that there is a significant negative relationship between income inequality and amount of leisure time or participation in cultural and sporting activities (Veal, 2016). Thus, people living in countries with lower decile ratios have more leisure time and higher levels of cultural and sporting participation. In addition, it was also confirmed that in countries with lower economic inequality, increased leisure time and higher levels of participation in cultural and sporting leisure activities are a part of all income and socio-economic groups.

Macroeconomic. Last but not least is the group of factors that include macroeconomic variables such as inflation, unemployment or international trade (export / import).

There is no consensus on the impact of inflation on income inequality and people's well-being. Some argue that higher inflation in Europe is associated with lower nominal income (e.g. pensions) for the population and is therefore linked to higher income inequality (Kaasa, 2005). Panel data study of 46 developing countries revealed a non-linear relationship between inflation and income inequality, while inflation showed a positive effect on economic

inequality (Nantob et al., 2015). However, increasing inflation can be a reason of the decreasing real value of private debt, which can reduce inequality. When summarizing the relationship between inflation and economic inequality, it should be noted that although economists agree that inflation has a complex effect on income inequality, both short-run and long-run, this interaction depends on the initial inflation rate (Balcilar, Chang, Gupta, & Miller, 2018).

The relationship between unemployment rate and income inequality is also not fully explored or completely clear. Most studies analyse country-level models, and therefore need further detailed analysis, including a cross-country element (Kaasa, 2005). At the end of the twentieth century, it is assumed and most studies have shown that unemployment rate increases income inequality, although in some cases there was no statistically significant relationship (Gustafsson & Johansson, 1999).

The impact of export and import on income inequality is not widely studied, therefore it requires further analysis. One of few dynamical studies of international trade in 65 countries found that general trade has not significantly influenced income inequality, but trading with high developed countries did have impact. It has also been found that trading with high income countries increases economic inequality between countries, both in terms of import and export (Meschi & Vivarelli, 2009).

1.4. Current situation in Lithuania

In Lithuania, income inequality has increased in recent years, as provided in Figure 2. Two-fifths of people of retirement age and one-third of single parents raising two or more children were at the risk of poverty. The emerging risks were identified by researchers (Bonoli, 2007) as low levels of education and skills, incompatibility between family and work, and being a single mother or father, as the major challenges for developing economies, including Lithuania as well. In addition, atypical employment contracts and part-time employment has become a new form of social risk in recent decades (Skučienė, Lazutka, Čižauskaitė, & Markevičiūtė, 2018). These groups are not protected by traditional social security systems, therefore measures have to be taken in order to reduce inequality of opportunity and social exclusion. One of the proposed solutions of this problem is an active social investment policy, which could serve as a tool in order to prevent the emergence of new social risks and the spread of old ones, and to provide better living opportunities for individuals in terms of their future life path. Previous studies highlighted investment in human capital as a preventive policy and emphasized the need for long-term investment in education to reduce inequality of opportunity (Esping-Andersen et al., 2002). However, as some Lithuania researchers claims, the situation has not changed in the period of economic growth and the ability of social security to protect the most vulnerable populations has not been revealed, therefore income inequality has increased in most population groups, compared to the situation in

2007 (Skučienė et al., 2018).

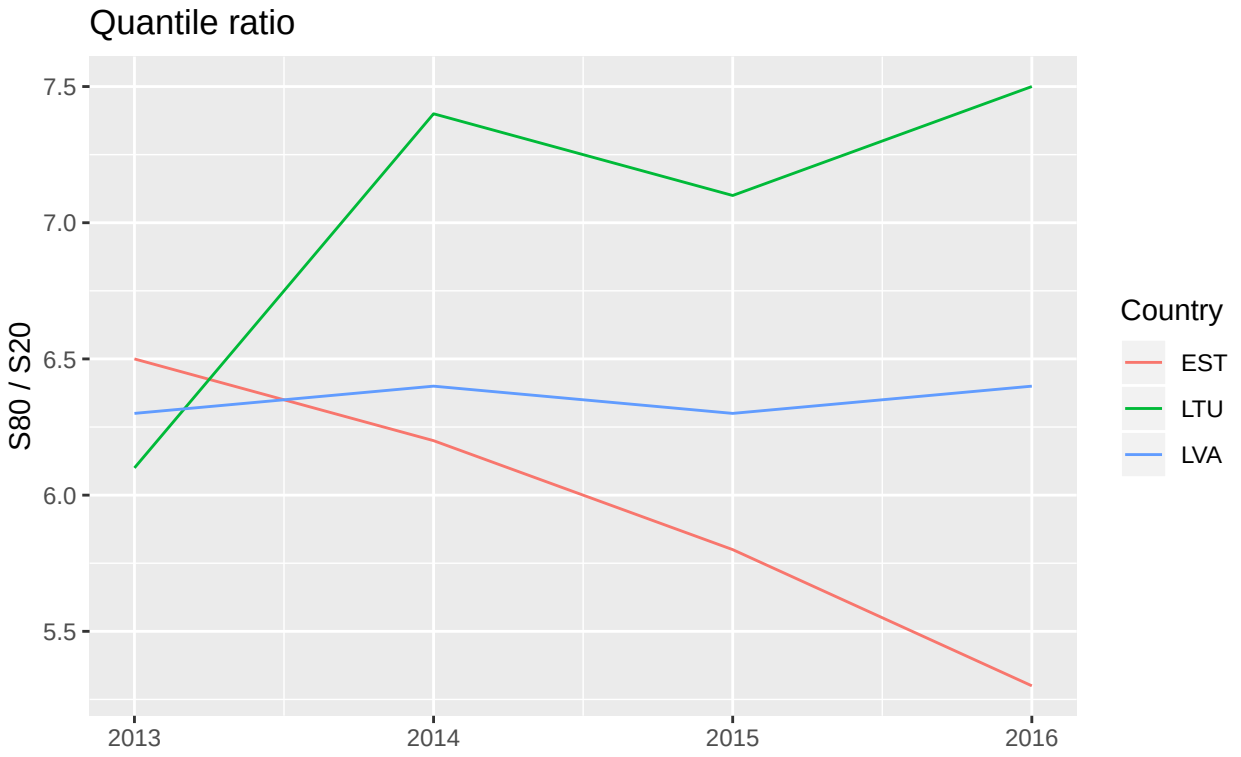

Figure 2. Ratio of the average income of the 20% richest to the 20% poorest in Baltic countries.

Some Lithuanian economists and sociologists point to inequality and corruption in the healthcare system (Lazutka, Poviliūnas, & Žalimienė, 2018). According to the authors, healthcare inequality was caused by indadequate financing - in Lithuania's healthcare system, expenditures were barely half of the EU average, meanwhile in 2018 the country's GDP reached three quarters of the EU average. Other research revealed inequality in the distribution of national income between labor and capital (Lazutka, Juška, & Navickė, 2018). Researchers agree that the neoliberal economic development model, which has been applied in the country since the establishment of independence, has boosted economic growth; however, they emphasize that it is based on low labor-related income, low capital taxation and the export of cheap labor to major EU countries, especially United Kingdom. Such operation of the neoliberal economic development model has led to severe income and wealth inequalities and a socio-demographic crisis. It is argued that changing the current situation requires reviewing and modifying the existing neoliberal development model in economic and fiscal policy making.

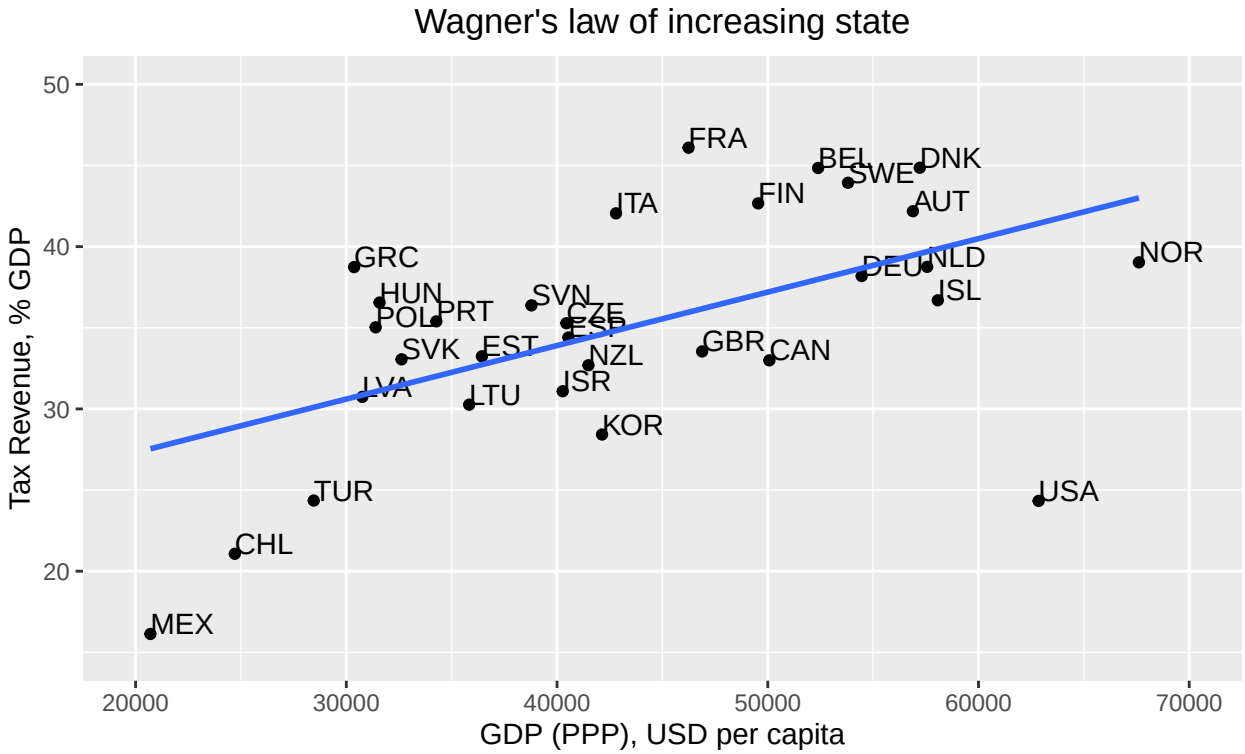

Figure 3. Trend of Wagner's law of increasing state in OECD countries: Real GDP, USD per capita versus Tax revenue, % of GDP.

Prof. R. Kuodis, at the annual SIGNALS conference organized by the Supreme Audit Institution National Audit Office of Lithuania in discussion on „What changes are needed to reduce income inequality in Lithuania?", stated that one of the biggest problem in Lithuania is the lower than usually general government revenue in comparison to the level of economic development of the country, i.e. GDP. According to Kuodis, the gap from Wagner's law, which states that for any country, that public expenditure increases constantly as income growth expands, is approximately 2 billion euros, that could potentially solve the problems of risk of poverty and inadequate financing of public sector. Here Figure 3, presented above describes Wagner's law of increasing state. Additionally, economist identifies „animal farms" which are related to personal income tax priviledges for various occupational groups and value added tax plundering as the two biggest reasons of low revenue of government budget, which occur on a national level. Dr. J. Navickė points out that data scarcity is one of the problems in assessing income inequality: the extent of wealth inequality in the country is unknown, and income inequality is underestimated due to the large-scale shadow economy, which is concentrated not only in the lowest income strata. According to researcher, income inequality has increased in the country during the economic growth period, and demographic factors, especially the changes in population structure, may contribute to even greater income and wealth inequalities in the future. However, as most participants of this discussion have noted, generous social benefits and extremely high progressive taxes can reduce motivation and incentives, and a balance is needed in this regard. National Audit Office delegate M.

Macijauskas highlighted the importance of impact assessment analysis: Ministry of Social Security and Labour is developing a scoreboard that could show how the risk of poverty is changing across regions by applying specific measures. Meanwhile, R. Kuodis concluded the debate by stating that Europe does not have a full employment strategy, giving an example with ten dogs and nine bones. According to professor, in order to pass large-scale inequalities and the risk of poverty, it requires initially tackling relatively high levels of unemployment and, ideally, the number of vacancies should be around the same as the number of unemployed people (NAO, 2019).

1.5. Machine learning algorithms and income inequality

Statistical learning includes a large set of tools to better understand the data. These measures are usually classified as supervised or unsupervised. The first one, supervised learning, typically involves classical statistical methods (e.g., least squares), designed to describe the interaction of one or more variables with a dependent variable, or to predict response values according to the available data. Thus, the abundance of parametric models and the possibilities of their interpretation enable researchers to perform both qualitative and quantitative studies, the inferences of which usually reflect the relationships between one or more variables and their directions of interaction. Meanwhile, in a case of unsupervised statistical learning, there are inputs, but no supervising output. Therefore, in this case, the researcher has more limitations in interpreting the results that were obtained (Hastie, Tibshirani, & Friedman, 2009; James, Witten, Hastie, & Tibshirani, 2013). However, the combination of classical and modern methods is increasingly being used to reduce the disadvantages of supervised statistical learning and to reap the benefits of unsupervised learning.

Combinations of cluster analysis (e.g., k-means) and parametric models have rarely been found in previous studies of income inequality, however, unsupervised learning is becoming an increasingly popular approach among researchers, used in order to distinguish between groups of countries or individuals based on certain indicators. One of such studies had the aim to group developing countries into welfare states according to their economic, social and health indicators at different times, i.e. it was investigated only based on values of indicators from 1990 and 2000 (Abu Sharkh & Gough, 2010). As the researchers noted, countries‘ dependence to relevant cluster remained unchanged in most cases; however, the key factors behind the country's move to other clusters were the HIV-AIDS pandemic in Africa and the growing role of remittances in some countries. Another article examines health inequalities and their consequences, in this case, mortality rates for adults and children under five at countries‘ level (Ruger & Kim, 2006). Interestingly, in both studies, the researchers ultimately chose the number of K based on their knowledge of the field under study. Thus, the optimal number of K does not yet guarantee the logical clustering of countries into groups, therefore clustering of countries requires knowledge of the field under study (Abu Sharkh &

Gough, 2010; Ruger & Kim, 2006).

1.6. Generalization of the theoretical part

An overview of the concept of income inequality, review of its impact on societies and economies, appreciation of quantitative measurement indicators, analysis of previous studies in income inequality and current situation in Lithuania leads to several conclusions:

1. Income inequality can be described as an uneven distribution of financial goods across individuals, households or countries.
2. Although the terms of social inequality, health inequality and education inequality are frequently used in the scientific literature, they can all be described as inequalities of opportunity, which is the outcome of income and wealth inequality.
3. High income inequality can be the cause of large-scale social problems, such as rising crime, lower public confidence or even a lack of loyalty to the country, therefore both democratic and autocratic governments tend to address this issue sooner or later. On the other hand, irresponsible economic policy, when facing of large-scale income inequality, can become the cause of public financial instability.
4. Measures of income inequality and their uses depend on the purpose and population of the study.
5. The complexity of the estimation of factors that influence income inequality is determined by their interaction and that they belong to different groups of factors. Researchers distinguish between different sets of factors, but in most cases studies do not include the whole spectrum of factors. A detailed literature review allowed to group the factors into five separate groups: economic development, demography, policy, cultural-environmental and macroeconomic.
6. According to the OECD, over the last decade, income inequality has increased in Lithuania. Furthermore, the risk of poverty threatens retired people, single parents, and other vulnerable social groups. Even in a period of economic upturn, the government has failed to increase expenditure of social and health care, and did not allocate additional assigns for social investment policy. Furthermore, in Lithuania, there is a tax revenue gap, which, compared to the country's economy, is approximately 2 billion euros. The major reasons for the lack of public financing are personal income tax privilege for part of population and value added tax plundering in national level.
7. Although machine learning algorithms are not commonly used in income inequality and other social researches, the combination of their methods and supervised statistical learning practices can lead to new results in cross-national studies.

Thus, the following disadvantage can be emphasized from previous studies: most parts of the studies do not include all groups of the factors determining income inequality, and

the studies which do incorporate them assess all countries at once, rather than using two stage methodology; where in the first stage, countries are clustering into different groups using machine learning algorithms, and in the second stage, there are building parametric models in order to estimate factors and their effects on income inequality in different groups of countries.

2. Notions

This section includes methods and models that were used thesis. Descriptions of the panel models (Cameron & Trivedi, 2005), LASSO regression (James et al., 2013), partial selection methods (Belloni, Chernozhukov, & Hansen, 2014; Chernozhukov, Hansen, & Spindler, 2016) and cluster analysis (Čekanavičius & Murauskas, 2002; James et al., 2013) are based on the literature presented in the brackets.

2.1. Panel data regression

Panel data is obtained by observing the same individuals at multiple time points, therefore such data includes dimensions of cross-sectional and time series. Usually many individuals and few moments of time are observed. Panel data have not only cross-sectional but also time series properties, therefore questions that could not be answered using only one dimension of panel data can be addressed. There are two types of panel data: balanced when all objects are observed at the same time intervals and unbalanced when not all objects are observed at the same time intervals.

2.1.1. Regressors

Let $\mathbf{X} = (x_{1it}^T, x_{2it}^T, ..., x_{pit}^T)$ is regressor matrix without intercept. Here p is a number of regressors, i is ith individual and t is tth year. There are three types of regressors depending on the structure of the panel data:

1. Varying in time and in individual objects (two-ways effects): x_{it}.
2. Varying in only individual objects (one-way individual effects): $x_{it} = x_i$, $\forall t$.
3. Varying in only time (one-way time effects): $x_{it} = x_t$, $\forall i$.

2.1.2. Assumptions

Using of panel data regression model requires data to meet some following assumptions:

1. Conditional mean is equal to zero: $\mathbb{E}[\varepsilon_i|\mathbf{X}] = 0$, $i = 1, 2, ..., N$. Therefore, the unconditional mean is also equal to zero and model residuals ε_i are not correlated with regressors $\mathbf{X}$.
2. Variance of model residuals is homogeneous: $\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma^2$.
3. Model residuals are not autocorrelated: $\text{Corr}[\varepsilon_t, \varepsilon_s|\mathbf{X}] = 0$, $t \neq s$.

Hence, model residuals are normally distributed with mean of zero and constant variance: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

2.1.3. Pooled panel

All observed objects are homogeneous, therefore the intercept β_0 is the same for all of them. Thus, all of coefficients are not dependent from individual or time:

$$Y_{it} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{jit} + \varepsilon_{it}, \quad i = 1, 2, ..., N, \quad t = 1, 2, ..., T. \tag{Eq. 2.1}$$

Estimates of model can be obtained by the ordinary least squares:

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \beta_0 - \beta x_{it})^2 = \min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \varepsilon_{it}^2. \tag{Eq. 2.2}$$

2.1.4. Fixed effects

Due to the heterogeneity of cross-sectional objects each individual has a group error μ_i, an unobserved and time-independent factor. Therefore, residuals ε_{it} consists of two components μ_i and e_{it}:

$$Y_{it} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{jit} + \mu_i + e_{it}, \quad i = 1, 2, ..., N, \quad t = 1, 2, ..., T. \tag{Eq. 2.3}$$

Here e_{it} is idiosyncratic errors has mean of zero and are independent of regressors x_{it} and individual errors μ_i. Since fixed effects model takes into account individuals heterogeneity, (Eq. 2.3) can be rewritten:

$$Y_{it} = \beta_{0i} + \sum_{j=1}^{p} \beta_j x_{jit} + e_{it}, \quad i = 1, 2, ..., N, \quad t = 1, 2, ..., T. \tag{Eq. 2.4}$$

Two other cases are possible: $\varepsilon_{it} = \lambda_t + e_{it}$ (time effects) and $\varepsilon_{it} = \mu_i + \lambda_t + e_{it}$ (individual and time effects).

If μ_i correlates with regressors, estimates obtained by the ordinary least squares method may be inconsistent. Therefore, the estimates of parameters are evaluated by the least squares method using transformed variables i.e. entering dummy variables. Let d_{ij} be a dummy variable:

$$d_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{Eq. 2.5}$$

Then fixed effects model can be rewritten as least squares dummy variables estimator regression model:

$$Y_{it} = \beta_{0i} + \sum_{j=1}^{N-1} \mu_i d_{ij} + \sum_{j=1}^{p} \beta_j x_{jit} + e_{it}, \quad i = 1, 2, ..., N, \quad t = 1, 2, ..., T. \tag{Eq. 2.6}$$

Thus, the model coefficients can be obtained by the least squares method.

If the number of individuals is large, estimating the coefficients of (Eq. 2.6) by the OLS may lead to calculation problems. After fixing the ith group and averaging in respect to time t, (Eq. 2.3) can be rewritten as following:

$$\overline{Y_i} = \beta_0 + \sum_{j=1}^{p} \beta_j \overline{x_{ji}} + \mu_i + \overline{e_i}, \quad i = 1, 2, ..., N. \qquad \text{(Eq. 2.7)}$$

Here $\overline{Y_i} = \sum_{t=1}^{T} Y_{it}/T$, $\sum_{j=1}^{p} \overline{x_{ji}} = \sum_{t=1}^{T} \sum_{j=1}^{p} x_{jit}/T$, $\overline{e_i} = e_{it}/T$ Subtracting (Eq. 2.7) from (Eq 2.3) gives:

$$Y_{it} - \overline{Y_i} = \sum_{j=1}^{p} \beta_j (x_{jit} - \overline{x_{ji}}) + e_{it} - \overline{e_i}. \qquad \text{(Eq. 2.8)}$$

Since errors $e_{it} - \overline{e_i}$ are not correlated with regressors $x_{it} - \overline{x_i}$, parameters can be estimated by OLS. Obtained coefficients are called within (fixed effects) estimators. Parameters estimates of within model are unbiased and consistent. In addition, coefficients of fixed (individual) effects can be obtained as follows:

$$\hat{\mu_i} = \overline{Y_i} - \overline{x_i'} + \hat{\beta}. \qquad \text{(Eq. 2.9)}$$

2.1.5. Durbin-Watson test

Durbin-Watson test is used to determine whether model residuals are autocorrelated. Errors of regression model are related by the following relationship:

$$e_{it} = \rho e_{it-1} + u_{it}, \qquad \text{(Eq. 2.10)}$$

here $u_{it} \sim \mathcal{N}(0, \sigma^2)$. Hence, the null hypothesis $H_0 : \rho = 0$ states that errors are not correlated, and alternative hypothesis $H_1 : \rho \neq 0$ states that errors are correlated. Then Durbin-Watson test statistic:

$$d = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} (e_{it} - e_{it-1})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2}, \qquad \text{(Eq. 2.11)}$$

here T is the number of observations. Since $d \simeq 2(1 - \hat{\rho})$, where $\hat{\rho}$ is the sample autocorrelation, $d = 2$ indicates no autocorrelation. Durbin-Watson statistic less than 2 indicates positive serial correlation, while test statistic bigger than 2 specify that there is evidence of negative serial correlation.

2.1.6. Levene's test

Levene's test is used to assess the equality of variances for a variable calculated for two or more groups. The null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ stands for homogeneity of variance, while the alternative stands for opposite claim $H_1 : \sigma_i^2 \neq \sigma_j^2$ for at least one pair (i, j). Then Levene's test statistic for variance of variable Y with sample of size N divided into k subgroups and N_i is the sample size of the ith subgroup, is defined by:

$$W = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k} N_i(\overline{Z_{i.}} - \overline{Z_{..}})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \overline{Z_{i.}})^2}, \qquad \text{(Eq. 2.12)}$$

here Z_{ij} can be defined by three different ways. Definition for the mean of the ith subgroup:

$$Z_{ij} = \left|Y_{ij} - \overline{Y_{i.}}\right| \qquad \text{(Eq. 2.13)}$$

In addition, $\overline{Z_{i.}}$ are the group means of the Z_{ij} and $\overline{Z_{..}}$ is the overall mean of the Z_{ij}. The other two forms of Z_{ij} allows to test data that are not normal distributed, therefore this test has enough of robustness.

2.1.7. Pesaran's CD test

Considering to (Eq. 2.4), where x_{jit} is jth regressor, β_j is parameters to be estimated and β_{0i} represents time invariant individual parameters. The null hypothesis states that e_{it} is independent and identically distributed (i.i.d.) over time and across cross sections. Meanwhile, under the alternative e_{it} is correlated across cross-sectional units, but the assumption of no autocorrelation remains. Hence, the null $H_0 : \rho_{ij} = \rho_{ji} = \text{Corr}[e_{it}, e_{jt}] = 0$ for $i \neq j$, while the alternative $H_1 : \rho_{ij} = \rho_{ji} \neq 0$ for some $i \neq j$. Here ρ_{ij} is the product-moment correlation coefficient of the disturbances, which can be define as follows:

$$\rho_{ij} = \rho_{ji} = \frac{\sum_{t=1}^{T} e_{it}e_{jt}}{\sqrt{\sum_{t=1}^{T} e_{it}^2}\sqrt{\sum_{t=1}^{T} e_{jt}^2}}. \qquad \text{(Eq. 2.14)}$$

Pesaran's proposed alternative of the LM test:

$$CD = \sqrt{\frac{2T}{N(N-1)}}\Big(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\widehat{\rho_{ij}}\Big). \qquad \text{(Eq. 2.15)}$$

CD statistic has mean at exactly zero for fixed values of T and N, under a wide range of panel-data models. There is modified version of (Eq. 2.15) test for unbalanced panels:

$$CD = \sqrt{\frac{2}{N(N-1)}}\Big(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sqrt{T_{ij}}\widehat{\rho_{ij}}\Big), \qquad \text{(Eq. 2.16)}$$

here $T_{ij} = \#(T_i \cap T_j)$ (i.e., the number of common time-series observations between units i and j),

$$\widehat{\rho_{ij}} = \widehat{\rho_{ji}} = \frac{\sum_{t\in T_i\cap T_j}(\widehat{e_{it}} - \overline{\widehat{e_i}})(\widehat{e_{jt}} - \overline{\widehat{e_j}})}{\sqrt{\sum_{t\in T_i\cap T_j}(\widehat{e_{it}} - \overline{\widehat{e_i}})^2}\sqrt{\sum_{t\in T_i\cap T_j}(\widehat{e_{jt}} - \overline{\widehat{e_j}})^2}}, \qquad \text{(Eq. 2.17)}$$

$$\overline{\widehat{e_i}} = \frac{\sum_{t\in T_i\cap T_j}\widehat{e_{it}}}{\#(T_i \cap T_j)}. \qquad \text{(Eq. 2.18)}$$

2.2. LASSO regression

LASSO (Least Absolute Shrinkage and Selection Operator) is a technique that constrains or regularizes the coefficient estimates. It is important that shrinking the coefficient estimates can significantly reduce their variance and help to facilitate interpretation of model. LASSO regression performs $L1$ regularization, which adds a penalty equal to the sum of absolute values of the parameters. In this case, some coefficients can become zero, therefore this makes easier to interpret models. The aim of this algorithm is to minimize the sum of squares with constraint $\sum_{j=1}^{p} |\beta_j| \leq s$:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + |\beta_j|\,, \qquad \text{(Eq. 2.19)}$$

here n is the number of observations and p is a number of predictors. Hence, in case of the LASSO, a penalty shrinks model coefficients towards zero when the tuning parameter λ is sufficiently large. Therefore, selection of the variable depends on the value of λ. Minimization problem of the LASSO that performs $L1$ regularization (Eq. 2.19) can be rewritten in a following way:

$$\min_{\beta} \Big\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \Big\}, \quad \lambda \sum_{j=1}^{p} |\beta_j| \leq s. \qquad \text{(Eq. 2.20)}$$

By performing the LASSO, there are trying to find the set of parameter estimates that lead to the smallest RSS, subject to the constraint that there is a budget s for how large $\sum_{j=1}^{p} |\beta_j|$ can be. Thus, when s is extremely large, then this budget is not sufficiently restrictive, therefore the coefficient estimates can be far away from zero. In other words, if s is large such that the solution of least squares belongs to the budget, then solution of the (Eq. 2.20) leads to the least squares solution. In contrary, if s is small, then $\sum_{j=1}^{p} |\beta_j|$ must be small that do not avoid violating the budget constraint.

2.2.1. Tuning parameter and its selection

Since tuning parameter λ controls the strength of the $L1$ penalty, λ is the amount of shrinkage: when λ is equal to zero, no coefficients are eliminated and the estimate is same as the obtained by ordinary least squares; meanwhile, as λ increases, more coefficients shrunks towards to zero. However, it is important to note that in this case there is a bias-variance trade-off: as λ increases, bias increases, while as λ decreases, variance increases.

Cross-validation is one of the ways to select the best model according to the optimal λ, with which RSS are smallest. In the first stage, choosing a grid of λ values, and computing the cross validation error for each value of λ. In the second stage the tuning parameter value for which the cross-validation error is smallest should be selected. Finally, the model is re-fit using of the selected value of the tuning parameter.

2.2.2. Inference of the target regressors

Let consider exogenous model with a target regressor such as policy or other factor whose regression coefficient α need to estimate:

$$Y_i = \alpha d_i + \sum_{j=1}^{p} \beta_j x_{ij} + \zeta_i, \quad \mathbb{E}[\zeta_i | d_i, x_i] = 0, \quad i = 1, 2, ..., n. \tag{Eq. 2.21}$$

$$d_i = \sum_{j=1}^{p} \gamma_j x_{ij} + \nu_i, \quad \mathbb{E}[\nu_i | x_i] = 0, \quad i = 1, 2, ..., n. \tag{Eq. 2.22}$$

Here p is a number of predictors can be small, approximately equal to number of observations n or much higher than n.

Then, double selection method is based on the idea is to select variables by LASSO of the dependent variable Y_i on the control variables x_i and the target variable d_i on the control variables x_i. The algorithm of the method is implemented in three steps:

1. Selecting controls x_{ij} that predict outcome Y_i by LASSO;
2. Selecting controls x_{ij} that predict target d_i by LASSO;
3. Fitting OLS of outcome on target variable d_i, and the union of controls x_{ij} selected in steps 1 and 2.

In case of the partialling-out firstly, the effect of the regressors x_{ij} on the response Y_i and the target variable d_i is taken out by LASSO, and then regressing residuals from step 1 $\widetilde{Y_i}$ on residuals from step 2 $\widetilde{d_i}$ using OLS:

1. Partialling out from Y_i the effect of all x_{ij} that are significant predictors of Y_i and obtaining model residuals $\widetilde{Y_i}$;
2. Partialling out from d_i the effect of all x_{ij} that are significant predictors of d_i and obtaining model residuals $\widetilde{d_i}$;
3. Regressing residuals of the first model $\widetilde{Y_i}$ on residuals of the second model $\widetilde{d_i}$ using OLS.

The resulting estimator for α is normally distributed which allows inference on the target d_i effect. In addition, it should be noted that the partial regression approaches represent a special case of the orthogonal estimating equations approach.

2.3. Cluster analysis

By applying cluster analysis, the selected quantitative measure identifies the similarities between objects, and by the chosen method similar objects are clustered. Thus, the aim of cluster analysis is to divide objects such that differences between clusters are minimized and between clusters as large as possible. The most commonly used similarity measures are metric distance measures, correlations and associative coefficients.

The only metric distance measure used in this work is the squared Euclidean distance. The numerical non-negative function $d(X, Y)$ of two objects X and Y satisfies conditions of metrics:

1. symmetry: $d(X, Y) = d(Y, X)$;
2. triangle inequality: $d(X, Y) \leq d(X, Z) + d(Y, Z)$;
3. separability of non-identical objects: if $X \neq Y$, then $d(X, Y) \neq 0$;
4. inseparability of identical objects: if $d(X, Y) = 0$, then X and Y are identical.

The Euclidean distance for objects X and Y is then defined in a following way:

$$d(X, Y) = \|X - Y\| = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}, \qquad \text{(Eq. 2.23)}$$

here m is number of observations. Then the squared Euclidean distance is:

$$d(X, Y) = \|X - Y\|^2 = \sum_{i=1}^{m}(x_i - y_i)^2. \qquad \text{(Eq. 2.24)}$$

2.4. K-means clustering

One of the method, which can divide data into K distinct, non-overlapping clusters is K-means clustering. In order to perform K-means, firstly number of clusters K should be specified. After number of clusters were chosen, each observation is assigned to exactly one of the clusters by K-means algorithm. This procedure includes mathematical optimization problem.

If $C_1, ..., C_K$ are the sets, where containing the indices of the observations in each cluster. Then these sets must satisfy two properties:

- $C_1 \cup C_2 \cup ... \cup C_K = \{1, ..., n\}$, i.e. each observation belongs to at least one of the K clusters.
- $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$, i.e. no observation belongs to more than one cluster.

Let the ith observation be in the kth cluster. Then this belonging can be denoted as $i \in C_k$. Since aim of K-means clustering is minimize within-cluster variation, solving problem can be written:

$$\min_{C_1,...,C_K} \Big\{ \sum_{k=1}^{K} W(C_k) \Big\}, \qquad \text{(Eq. 2.25)}$$

here measure $W(C_k)$ denotes the within-cluster variation for cluster C_k by which the observations belonging to cluster differs from each other. If this measure is squared Euclidean distance, $W(C_k)$ can be defined as:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \qquad \text{(Eq. 2.26)}$$

here $|C_k|$ is the number of observations in the kth cluster. Then the optimization problem for the K-means can be defined by (Eq. 2.26) interposing to (Eq. 2.25):

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \qquad \text{(Eq. 2.27)}$$

Algorithm for K-means clustering problem defined by (Eq. 2.27) includes three steps:

1. When the researcher chooses the number of clusters K, the observations are dividing into them.
2. The distance of each object to the cluster centers is calculating using the squared Euclidean distance. The object is assigned to the nearest cluster and the centers of the clusters are recalculated.
3. Step 2 is repeating until the redistribution to clusters is completed.

However, the pre-selection of the number of clusters may lead to several problems, which includes that some clusters might little differs from each other; outliers can form a separate cluster; specifying the initial number of clusters, the structure of data is imposed. Therefore, using K-means clustering can produce unexpected results.

3. Econometric analysis

3.1. Methodology

3.1.1. Data

The practical part of the work uses annual data from 24 OECD countries from the period of 2007 - 2016. In order to involve more countries, unbalanced panel data, (not all countries have the same number of observations) were used. In this case, the number of observations ranges from 2 to 10. OECD countries which has 1 or less observations have been excluded from data. In addition, Ireland has been excluded from the study, since the country's GDP and ratios of GDP are overestimated or underestimated, therefore any conclusions on the statistical significance of the variables may not be realistic.

3.1.2. Dependent variable

In this work income is defined as the disposable income of a household in a given year. Disposable income includes wages, self-employment and capital income, as well as government cash transfers. In addition, income taxes and social security contributions paid by households are subtracted. Dependent variable of this research is Gini index. This coefficient is based on the cumulative ratio of population against cumulative ratio of its received income. Values of this indicator ranges between 0 and 1 or 0% and 100%, where 0 or 0% corresponds perfect equality, while 1 or 100% indicates perfect inequality.

3.1.3. Regressors

Five groups of determinants of income inequality were identified based on the theoretical literature analysis of income inequality. Groups of factors includes economic growth or economic development, demography, policy, culture and environment and macroeconomic. Regression analysis includes 5 independent variables that represent each of the groups of factors:

- **GDP growth (Economic growth / Economic development)**. Gross domestic product (GDP) is the mostly used measure of the value added created through the production of goods and services in a country during particular period. It can be said that GDP measures the income earned from production, or the total amount spent on final goods and services (less imports). This indicator is based on real GDP (also called GDP at constant prices). Thus, despite GDP reflects total income (expenditure) and can be as indicator of output or economic development, this measure does not take into account to the distribution of that income (expenditure), therefore its growth may not necessarily lead to decreasing income inequality or economic welfare of society. GDP growth is measured as percentage change from the previous year.

- **Fertility rate (Demography)**. The total fertility rate in a certain year is determined as the total number of children that would be born to each woman if she were to live to the end of her child-bearing years and give birth to children in respect with the prevalenting age-specific fertility rates. This index is calculated by suming the age-specific fertility rates as defined over five-year intervals. With assume that there is no net migration and unchanged mortality, a total fertility rate of 2.1 children per woman assures stable population. Hence, considering to the prevailing tendencies and their causes, fertility rate is one of the most important indicator of demography. Fertility rate is measured as number of children per woman.

- **Tax revenue (Policy)**. Tax revenue is defined as the revenues collected from taxes on income and profits, social and health security contributions, value added taxes, and other taxes. Total tax revenue as a percentage of GDP represents the share of a country's output that later becomes assignation of health, social and other services. It can be considered as one of the measure to which the government controls the country's resources. Hence, this indicator is widely related with making of economic policy.

- **Culture expenditure (Culture and Environment)**. General government spending indicates the size of government across OECD countries. The different values of this indicator suggests that countries have different priorities and different economic policy is being executed across countries. Cultural expenditure used in this context includes government expenditure on recreation, culture and religion. This indicator is measured as percentage of the GDP.

- **Unemployment rate (Macroeconomic)**. Unemployed people are defined as who are registered by the state as unemployed. They should be able for work and they have taken active steps to find work during the last four weeks. Unemployment rate is the number of unemployed as a percentage of the labour force. Thus, this indicator is closely related to the condition of the country and is one of the main ones representing the situation in the labor market.

3.1.4. Additional variables

In order to obtain more robust clustering results and to highlight the influence of each of the factor groups, five additional variables corresponding to each of the factor groups were used for clustering with the above variables:

- **Labour productivity (Economic growth / Economic development)**. Labour productivity growth is one of the main aspect of economic activity and a major driver of changes in living standards. Gross Domestic Product (GDP) per capita growth can be defined as labour productivity growth, measured as GDP per hour worked, and changes in the labour utilization rate, measured as changes in hours worked per

capita. High labour productivity growth may suggests higher capital utilization, lower employment of low-productivity workers, overall efficiency gains and innovation. This indicator is measured as percentage change from the previous year.

- **Adult education level (Demography)**. Adult education level can be defined as defined by the highest level of education completed by the 25-64 year-old population. Tertiary education level is measured as a percentage of same age population.
- **Social expenditure (Policy)**. General government spending on social protection includes cash benefits, direct provision of goods and services and tax benefits with social purposes. Benefits usually targeting low-income households, the elderly, disabled, sick, unemployed or young persons. This indicator is measured as a percentage of GDP.
- **Environmental tax (Culture and Environment)**. Environmentally related taxation can be divided into different domain: energy products (including fuel); motor vehicles and transport services; measured or estimated emissions to air and water, ozone depleting substances and other. Hence, this measure can represent governments attention to the environmental protection. Environmental taxes are measured as a percentage of GDP.
- **Inflation (Macroeconomic)**. Inflation measured by consumer price index (CPI) is defined as the change in the prices of a basket of goods and services that are typically purchased by specific groups of households. Inflation is measured in terms of the annual growth rate.

3.1.5. Models and methods

Practical study uses pooled and one-way fixed individual and -time effects panel regression models, which are initially constructed for non-clustered data. Economic model for income inequality problem:

$$GINI = f(GDP + FERT + TAX + CULT + UNEMP), \qquad \text{(Eq. 3.1)}$$

where $GINI$ - Gini index, GDP - GDP growth, $FERT$ - fertility rate, TAX - tax revenue, $CULT$ - culture expenditure, $UNEMP$ - unemployment rate. Then econometric model can be written as following:

$$GINI_{it} = \beta_0 + \beta_1 GDP_{it} + \beta_2 FERT_{it} + \beta_3 TAX_{it} + \beta_4 CULT_{it} + \beta_5 UNEMP_{it} + \epsilon_{it}, \qquad \text{(Eq. 3.2)}$$

where i - ith country and t - tth year: $GINI$ - Gini index, GDP - GDP growth, $FERT$ - fertility rate, TAX - tax revenue, $CULT$ - culture expenditure, $UNEMP$ - unemployment rate, ϵ - model residuals. In addition, $i = 1, 2, ..., 24$ and $t = 1, 2, ..., 10$.

Durbin-Watson and Levene's tests are used to test assumptions of panel models. In order to achieve the purpose of thesis, the data is clustered by the K-means method. A total of

11 variables are used for clustering the data: dependent variable, regressors, and additional variables that are used only in clustering. Since it is important to distinguish each country by its characterized indicators, the averages of all variables were calculated for each country before clustering. After clustering, pooled and fixed effects models for grouped data are constructed. In order to control the results of these models, to improve their interpretation and to reduce the standard errors of the model parameters, LASSO regression and double selection and partialling-out methods were used.

3.2. Modelling income inequality

3.2.1. Exploratory analysis

Before modelling any type of data, it is useful to understand its characteristics and trend using data visualization tools. In the case of panel data, this is particularly important, since not only individuals (countries), but also time (years) is included in models. Thus, in this case the variability of the dependent variable for both across countries and across time are important. Here Figure 4 box plots represents heterogeneity of income inequality across 24 OECD countries for the period under 2007 - 2016.

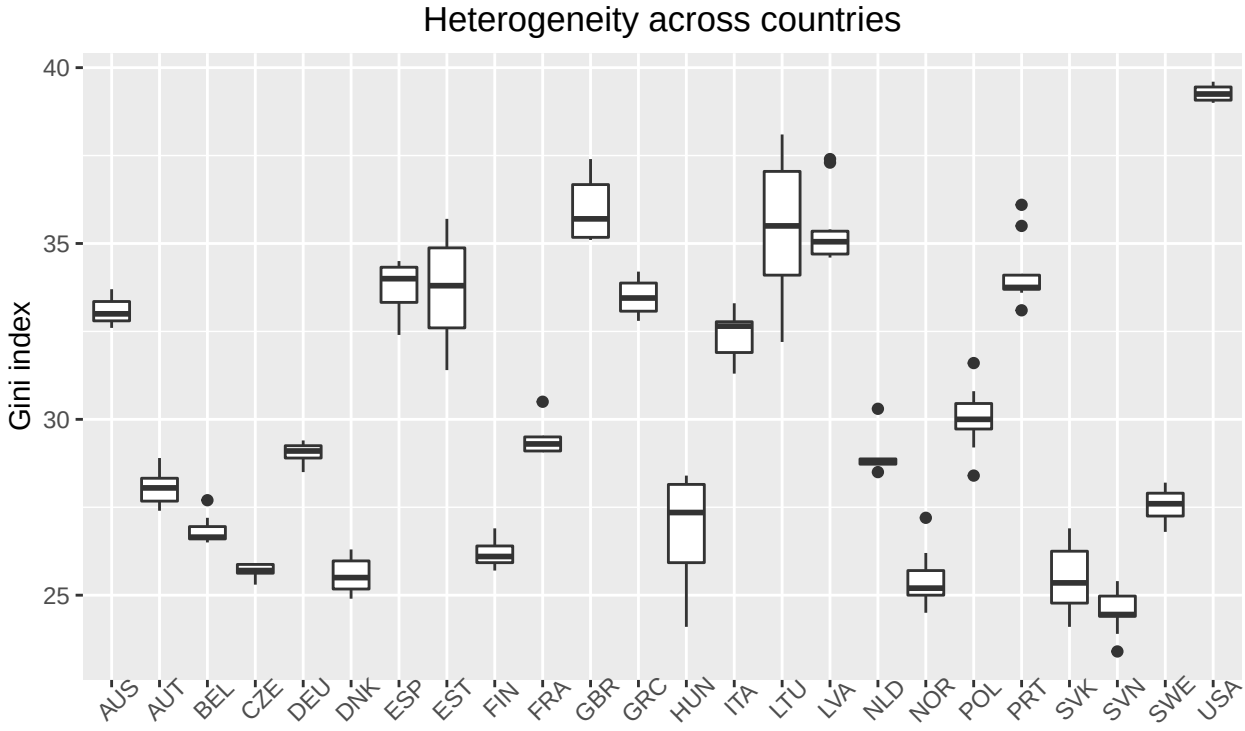

Figure 4. Box plots of Gini index: heterogeneity of income inequality across OECD countries.

During the period under review, the largest variability of income inequality were found in Lithuania, Estonia and Hungary, while the smallest fluctuations of the Gini index values were

recorded in the United States, Czech Republic and Germany. It is worth noting that the levels of income inequality in these countries were significantly different over the period under review, with Gini in the United States ranging from 39.0 and 39.6, in Lithuania between 32.2 and 38.1, while in the Czech Republic between 25.3 and 25.9 and in Hungary between 24.1 and 28.4.

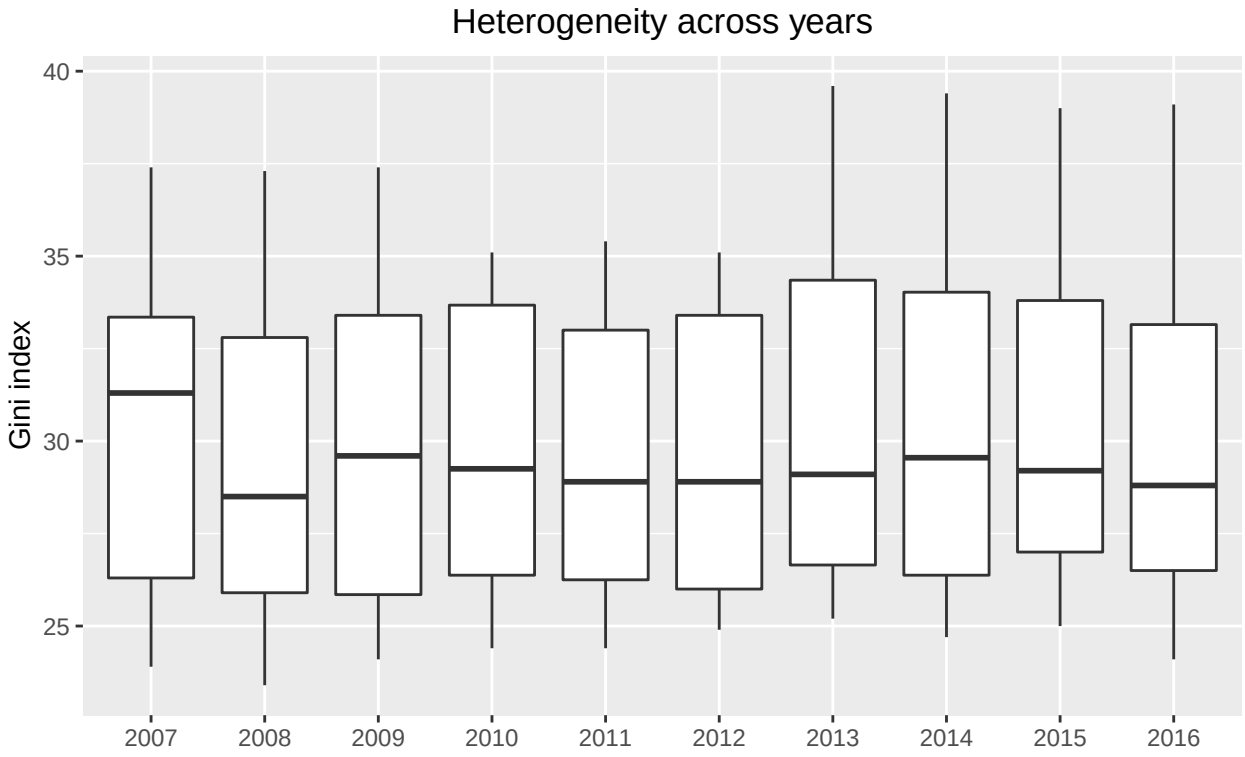

Figure 5. Box plots of Gini index: heterogeneity of income inequality across period under 2007 - 2016.

Figure 5 presented box plots represents heterogeneity of income inequality in OECD countries across years in the period of 2007 - 2016. The largest differences in income inequality between OECD countries was observed after the Great Recession (2013 - 2016), while the largest median of income inequality was recorded in 2007. Hence, box plots of heterogeneity across countries and across years showed that need to consider to both individuals and time effects in this case.

Since linear panel data models will be used in the modelling, it is also important to check the correlations between the variables used in the study. Figure 6 provides the linear correlations between the variables included in the panel models and the LASSO regression. In addition, fields who crossed out represents not significant associations between paired variables, according by Kendall's tau correlation coefficient. Regardless to heterogeneity across individuals (countries) and time (years) it can be stated, there is a direct correlation between the Gini index and unemployment rate, while public spending on cultural services and tax revenue have a negative correlation with income inequality. In this case, the cor-

relation between Gini coefficient and GDP growth and fertility rate is not significant with significance level $\alpha = 0.05$.

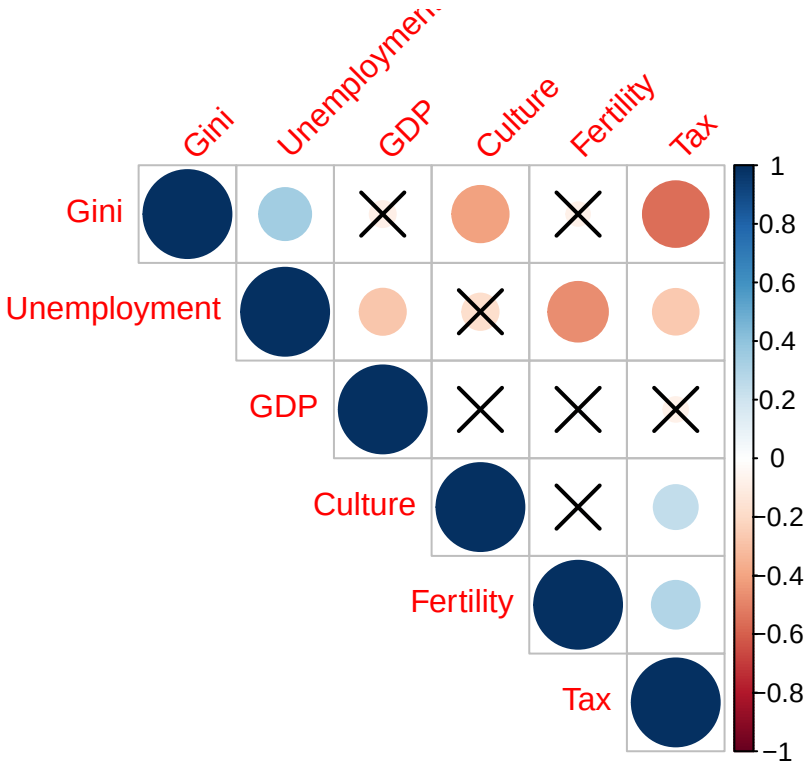

Figure 6. Linear correlation between the variables included in the panel models and the LASSO regression.

Thus, it is important to consider heterogeneity across countries and across years, while there is no detected strong correlation between regressors.

3.2.2. Models of income inequality for non-clustered data

Firstly, pooled, fixed-individual effects and fixed-time effects models for non-clustered and unbalanced data were fitted. Since the data is non-clustered, in this case the number of countries n is 24 and the number of observations per country T ranges from 2 to 10. The total number of observations N for all three models is 197.

Estimates of model parameters, their standard errors, t-test statistic values and its corresponding p-values are presented in Table 1. In pooled model only GDP growth is statistically insignificant regressor, while the only one culture expenditure is significant in fixed-individual effects model. For the time-fixed effects model, all five variables were statistically significant. In all cases R^2 is lesser than 0.50, thus it can be stated that possibilities of the explanation of the models are very limited. In addition, as shown in Figure 4 and Figure 5, the variability of income inequality were significant across countries and across years, therefore model on non-clustered data may hide the important determinants of income inequality across different countries.

Table 1. Parameters estimates and their standard errors in case of the pooled panel, fixed-individual effects and fixed-time effects model for non-clustered data.

Variable	**Pooled**	**Fixed-individual effects**	**Fixed-time effects**
GDP growth	-0.079	0.019	-0.298***
	(0.07)	(0.02)	(0.10)
Fertility rate	3.917***	1.235	4.246***
	(1.24)	(1.02)	(1.26)
Tax revenue	-0.363***	-0.081	-0.400***
	(0.04)	(0.05)	(0.04)
Culture expenditure	-2.543***	1.340***	-2.255***
	(0.57)	(0.40)	(0.57)
Unemoployment rate	0.211***	0.050*	0.200***
	(0.06)	(0.03)	(0.06)
R^2	0.454	0.090	0.489

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

In order to check for panel model assumptions, Durbin-Watson, Levene's and Pesaran's CD tests has been completed. Obtained tests results provided in Table 2.

Table 2. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross dependence in panels in case of all three panel models for non-clustered data.

Models / Tests	**Durbin-Watson test**	**Levene's test**	**Pesaran's CD test**
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	0.284***	3.484***	5.961***
Fixed-individual effects	1.207***	3.804***	0.975
Fixed-time effects	0.306***	3.863***	1.478

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Durbin-Watson and Levene's tests showed that there are evidence of autocorrelation and heteroscedasticity in the residuals of all three models. Meanwhile, Pesaran's CD test indicates that there is no cross-dependence in case of the both fixed effects models. Scatter of residuals plots provided in subsection named A1 confirms Levene's test results about heterogeneity of variance in residuals, while Q-Q Normal plots indicates that errors are not normally distributed.

Since residuals of the models are autocorrelated and heteroscedastic, Heteroscedasticity

and Autocorrelation Consistent (HAC) covariance matrix estimation has been done. Recalculated robust standard errors are provided in Table 3.

Table 3. Parameters estimates and their robust standard errors in case of the pooled panel, fixed-individual effects and fixed-time effects model for non-clustered data.

Variable	**Pooled**	**Fixed-individual effects**	**Fixed-time effects**
GDP growth	-0.079	0.019	-0.298*
	(0.09)	(0.04)	(0.17)
Fertility rate	3.917	1.235	4.246
	(2.60)	(1.11)	(2.86)
Tax revenue	-0.363***	-0.081	-0.400***
	(0.10)	(0.07)	(0.09)
Culture expenditure	-2.543***	1.340***	-2.255**
	(0.96)	(0.50)	(1.02)
Unemoployment rate	0.211***	0.050	0.200**
	(0.07)	(0.05)	(0.09)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Heteroscedasticity and autocorrelation consistent standard errors showed that fertility rate becomes insignificant in case of the pooled and fixed-time effects models. In addition, GDP growth is no longer significant regressor in fixed-time effects model after standard erros were recalculated. Thus, despite no changes were detected in case of the fixed-individual effects, all three models does not met the assumptions. Therefore, in order to achieve the purpose of thesis, only grouped data will be modeled in further analysis.

3.2.3. Clustering

In order to uniform the scale of variables measurement, first of all, standardization of data were done. A detailed literature analysis has shown that there is no reason to set more than 7 centers. On the other hand, the most common grouping of countries are at least into 4 or 5 clusters, therefore cases with centers of 5, 6 and 7 were analyzed before choosing the final number of centers.

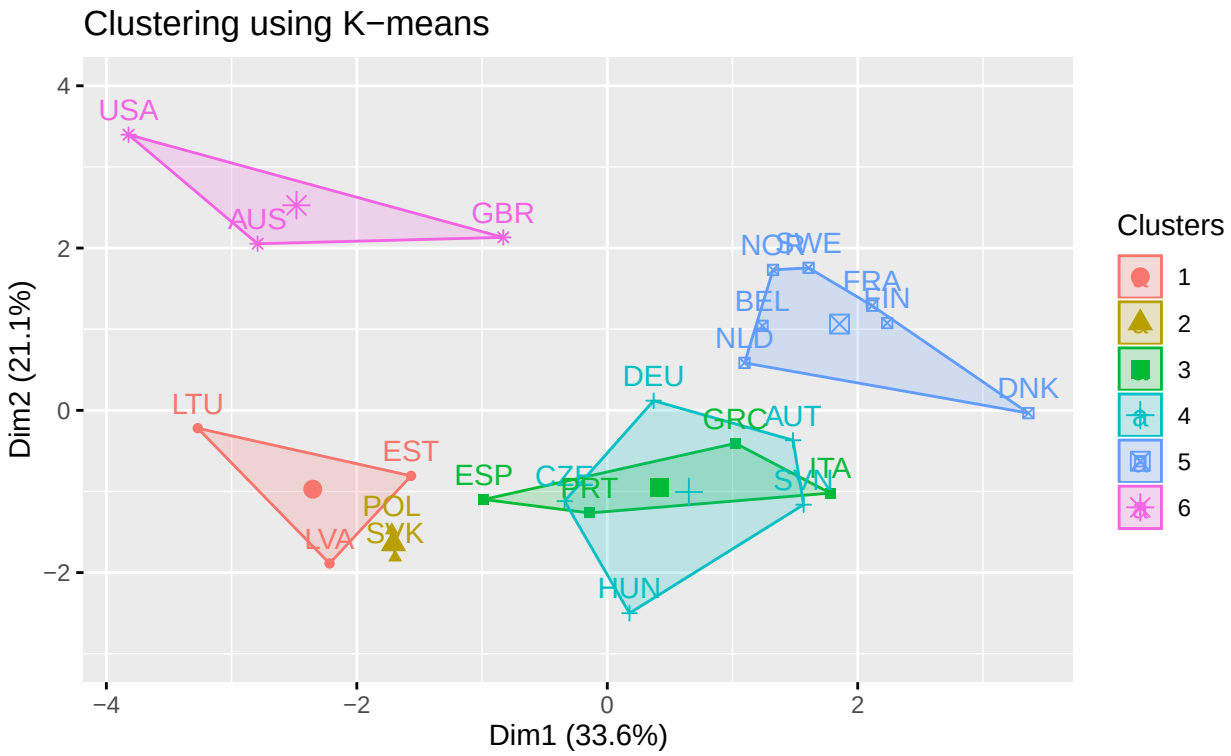

Figure 7. Clustering OECD countries into 6 different groups using K-means.

After selecting number of clusters $K = 6$, the results of country clustering are similar to the grouping of European Union countries were proposed before in the context of social welfare state models (Lazutka, 2003). Thus, clustering results provided in Figure 7 showed that the first group contains Baltic countries (Estonia, Latvia and Lithuania), while in the second cluster two Middle Europe countries (Poland and Slovak Republic). The third group of countries includes Southern European (Greece, Italy, Portugal and Spain), which are characterized by the conservative social welfare state model with the features of clientelism. In addition, Central European countries (Austria and Germany), which belongs to the Conservative or Corporatist social welfare state model, fell into the fourth cluster. On the other hand, due to the relatively low averages of Gini index and other indicators similarity to the Austria's and Germany's, this group also includes three Central European countries (Czech Republic, Hungary and Slovenia). Meanwhile, the fifth cluster includes Nordic countries (Denmark, Finland, Norway and Sweden), operating on the institutional model of the social welfare state, and continental European countries (Belgium, France and the Netherlands). Finally, the last cluster of countries includes only Anglo-Saxon countries (Australia, Great Britain and the United States).

3.2.4. Models of income inequality for clustered data

After data clustering, panel models are fitted for each group of countries. In order to improve interpretation of the panel models, LASSO regression with tuning parameter λ, selected by cross validation was performed. To reduce the standard errors of the parameter

estimates, double selection and partialling-out methods were applied for each target variable.

Cluster 1: Baltics. Table 4 provided below presents parameters estimates and their standard errors obtained in case of panel models and LASSO variables selection methods for the first cluster which includes Estonia, Latvia, and Lithuania. This cluster has a total of 24 observations, and number of observations per country ranges from 4 to 10.

Table 4. Parameter estimates and their standard errors for cluster 1 using panel regression models and variable selection methods.

Variable	**Pooled**	**FE-ind**	**FE-time**	**LASSO(λ)**	**Partial**	**Double**
GDP growth	-0.155**	-0.134	-0.146	-0.063	-0.050	-0.050
	(0.06)	(0.08)	(0.36)		(0.06)	(0.04)
Fertility rate	0.170	1.023	-10.369	0.000	2.895	2.895
	(2.82)	(3.77)	(7.85)		(3.25)	(2.31)
Tax revenue	-0.794***	-0.734**	-1.801*	-0.405	-0.649***	-0.649***
	(0.22)	(0.33)	(0.91)		(0.24)	(0.22)
Culture expenditure	0.261	1.744	1.277	0.000	-0.001	-0.001
	(0.84)	(4.25)	(1.94)		(1.04)	(0.99)
Unemoployment rate	-0.347***	-0.316**	-1.093	-0.153	-0.345***	-0.345***
	(0.10)	(0.13)	(0.76)		(0.10)	(0.09)
R^2	0.521	0.442	0.523			

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Thus, in this case, tax revenue and unemployment rate are statistically significant regressors in fixed-individual effects model with $\alpha = 0.05$. Pooled panel model offers same significant regressors as fixed-individual effects model, however also includes GDP growth. Furthermore, signs of all values of significant regressors are negative, therefore as GDP growth, tax revenue and unemployment rate increases, income inequality in terms of Gini index decreases. LASSO shrunks parameter estimates of fertility rates and culture expenditure to zero. On the other hand, only tax revenue and unemployment rate are significant in case of the double selection and partialling-out. Thus, in the case of the cluster 1, pooled panel model and the LASSO regularization gives identical results in terms of regressors significance, while double and partial methods confirmed significance of unemployment and tax revenue.

In order to be able to draw conclusions about the results of panel models, its residuals must met assumptions. Obtained tests results of autocorrelation, heteroscedasticity and cross-dependence provided in Table 5.

Table 5. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross-dependence in panels for cluster 1.

Models / Tests	**Durbin-Watson test**	**Levene's test**	**Pesaran's CD test**
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	1.062^{***}	1.525	-2.043^{**}
Fixed-individual effects	1.055^{***}	1.163	-2.073^{**}
Fixed-time effects	1.195^{***}	0.535	-2.869^{***}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Since the p-values for all models are lesser than 0.05 in case of tests of autocorrelation and cross-dependence in panels, it can be argued that the null hypothesis should be rejected, which results in autocorrelation of residuals and cross-dependence in panels. On the other hand, Levene's test results indicates that clustering has solved the problem of heteroscedasticity.

Results of heteroscedasticity and autocorrelation consistent standard errors are provided in the subsection A.2. placed in the Appendix. Thus, Table 16 provides that robust standard errors has not changed significance level of regressors for fixed-individual effects model, while GDP growth becomes insignificant in pooled model. Furthermore, tax revenue and unemployment rate becomes significant in fixed-time effects model. Hence, in case of the first cluster, autocorrelation significantly affects results of pooled and fixed-time effects models.

Cluster 2: Part of Eastern Europe. Table 6 provided below presents parameters estimates and their standard errors in case of panel models and LASSO variables selection methods for Poland and Slovak Republic. The second cluster has a total of 20 observations, and number of observations per country is equal to 10.

There is no one statistically significant regressor in both fixed-individual and fixed-time effects models in case of the second cluster. Meanwhile, in the pooled panel model, fertility rate and culture expenditure are statistically significant with $\alpha = 0.05$. The signs of corresponding estimates are negative and positive. Thus, as fertility rate increases, income inequality decreases and in reversely, as culture expenditure increases, Gini coefficient increases. LASSO regression method in this case shrunks estimates of GDP growth and unemployment rate to zero, while fertility rate, tax revenue and culture expenditure were left as significant effects. In addition, in case of the double selection and partialling-out only culture expenditure and fertility rate are significant with slightly lower standard errors comparing with panel models.

Table 6. Parameter estimates and their standard errors for cluster 2 using panel regression models and variable selection methods.

Variable	Pooled	FE-ind	FE-time	LASSO(λ)	Partial	Double
GDP growth	-0.047	-0.003	-0.125	0.000	-0.028	-0.032
	(0.09)	(0.07)	(0.18)		(0.07)	(0.08)
Fertility rate	-12.129**	-0.869	3.093	-3.772	-10.915***	-11.393***
	(4.15)	(5.24)	(12.99)		(3.57)	(3.86)
Tax revenue	0.193	0.027	0.885	0.026	0.157	0.184
	(0.19)	(0.17)	(0.61)		(0.16)	(0.24)
Culture expenditure	11.570***	3.711	-4.653	9.644	11.379***	11.379***
	(1.86)	(3.18)	(9.83)		(2.04)	(1.96)
Unemoployment rate	0.027	0.240	-0.996	0.000	0.038	0.036
	(0.15)	(0.15)	(0.64)		(0.14)	(0.15)
R^2	0.872	0.355	0.940			

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

The results of the Levene's test provided in Table 7 indicates that the problem of heteroscedasticity has been solved. On the other hand, the Durbin-Watson test shows that there is evidence to reject the null hypothesis, therefore the residuals of the pooled and fixed-individual effects models are autocorrelated. Furthermore, cross-dependence in panels were not detected in case of the pooled and fixed-individual effects.

Table 7. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross-dependence in panels for cluster 2.

Models / Tests	Durbin-Watson test	Levene's test	Pesaran's CD test
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	1.435**	2.184	-1.361
Fixed-individual effects	1.359**	1.745	0.769
Fixed-time effects	1.693	0.000	-3.162***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Results of robust standard errors provided in Table 17 indicates that there is no changes in case of the pooled model. On the other hand, in fixed-individual effects model culture expenditure has become significant, while in fixed-time effects model unemployment rate becomes significant with at least $\alpha = 0.01$.

Cluster 3: South Europe. Table 8 provided below presents parameters estimates and their standard errors in case of panel models and LASSO variables selection methods for the cluster of South Europe countries which includes Greece, Italy, Portugal and Spain. This cluster contains total of 40 observations, and number of observations per country is equal to 10.

In this case tax revenue is significant with at least level of significance $\alpha = 0.05$ in pooled and fixed-time effects models. Meanwhile, in the fixed-individual effects model, only unemployment rate is statistically significant. Tax revenue has negatively relationship with income inequality, i.e. as consolidations of government tax revenue increases, income inequality decreases, while unemployment rate has positive relationship. LASSO regularization shrunks coefficients of GDP growth and culture expenditure to zero. On the other hand, only fertility rate and tax revenue are significant using the double selection and partialling-out methods. In addition, these methods gives more accurate estimates of parameters for fertility rate and identical accuracy as in case of the tax revenue.

Table 8. Parameter estimates and their standard errors for cluster 3 using panel regression models and variable selection methods.

Variable	**Pooled**	**FE-ind**	**FE-time**	**LASSO(λ)**	**Partial**	**Double**
GDP growth	0.059	0.081*	0.045	0.000	0.027	0.028
	(0.04)	(0.04)	(0.09)		(0.04)	(0.04)
Fertility rate	-3.997*	4.007	-3.405	-1.671	-4.700**	-4.774***
	(2.19)	(3.39)	(3.56)		(2.13)	(1.76)
Tax revenue	-0.122***	-0.108	-0.130***	-0.060	-0.124***	-0.124***
	(0.03)	(0.07)	(0.04)		(0.03)	(0.03)
Culture expenditure	-0.472	-0.544	-0.321	0.000	-0.289	-0.292
	(0.41)	(0.86)	(0.60)		(0.40)	(0.28)
Unemoployment rate	0.030	0.119***	0.015	0.007	0.018	0.018
	(0.02)	(0.04)	(0.03)		(0.02)	(0.03)
R^2	0.531	0.323	0.518			

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Once again, results of Levene's test provided that there is no evidence to reject the null hypothesis, therefore there is no heteroscedasticity in case of the Southern European countries cluster. Pesaran's CD test indicates no cross-dependence in panels for all three panel models, while obtained Durbin-Watson test results confirmed autocorrelation of the models residuals.

Table 9. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross-dependence in panels for cluster 3.

Models / Tests	**Durbin-Watson test**	**Levene's test**	**Pesaran's CD test**
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	1.062^{***}	0.815	-0.435
Fixed-individual effects	1.015^{***}	1.726	-0.947
Fixed-time effects	0.957^{***}	1.325	-1.102

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 18, which represents results of robust standard errors in case of the group of South European countries showed that autocorrelation does not significantly affected fixed-time effects model results. Meanwhile, repaired standard errors led that fertility rate becomes significant in pooled panel and tax revenue becomes significant in fixed-individual effects model.

Cluster 4: Central Europe. Table 10 provided below presents parameters estimates and their standard errors in case of panel models and LASSO variables selection methods for Austria, Czech Republic, Germany, Hungary and Slovenia. The fourth cluster has a total of 47 observations, and number of observations per country ranges between 7 and 10.

Table 10. Parameter estimates and their standard errors for cluster 4 using panel regression models and variable selection methods.

Variable	**Pooled**	**FE-ind**	**FE-time**	**LASSO(λ)**	**Partial**	**Double**
GDP growth	0.049	0.080^{**}	-0.026	0.000	0.094	0.094
	(0.08)	(0.03)	(0.17)		(0.08)	(0.09)
Fertility rate	-5.415^{*}	2.760	-13.238^{***}	0.000	-3.374	-3.602
	(2.84)	(1.84)	(3.21)		(2.59)	(2.77)
Tax revenue	0.186^{**}	-0.287^{*}	0.087	0.000	0.246^{***}	0.246^{***}
	(0.08)	(0.15)	(0.08)		(0.08)	(0.04)
Culture expenditure	-0.629	1.594^{***}	-0.446	0.000	-0.888	-0.884
	(0.59)	(0.38)	(0.53)		(0.59)	(0.84)
Unemoployment rate	-0.222	0.133^{*}	-0.436^{***}	0.000	-0.201^{*}	-0.205^{*}
	(0.14)	(0.07)	(0.14)		(0.11)	(0.12)
R^2	0.332	0.536	0.558			

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Tax revenue is statistically significant in pooled model in case of the, while GDP growth and culture expenditure are statistically significant regressors in fixed-individual effects model. On the other hand, fertility rate and unemployment rate are significant in fixed-time effects model. LASSO shrunks all estimates of parameters, while double and partial methods states that only tax revenue is significant for this cluster. Since, there is a lot of contradictions in case of panel models and LASSO regularization, it is difficult to draw conclusions in case of this cluster. Such contradictions can be explained by the results of clustering: Central European countries as the Czech Republic, Hungary and Slovenia got into one group with other European countries as Germany and Austria, which proceeds on different social models.

Table 11, which provides tests of autocorrelation, heteroscedasticity and cross-dependence in panels results, confirms that clustering did not help to avoid heterogeneity across different countries. In order to get more validity in results and homogeneity in clusters, fourth cluster could be divided into two parts.

Table 11. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross-dependence in panels for cluster 4.

Models / Tests	**Durbin-Watson test**	**Levene's test**	**Pesaran's CD test**
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	0.969^{***}	4.151^{***}	1.688^{*}
Fixed-individual effects	1.814	4.357^{***}	1.614
Fixed-time effects	0.966^{***}	0.372	-1.858^{*}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Results of heteroscedasticity and autocorrelation consistent standard errors provided in Table 19 showed that autocorrelation does no significant affected results of fixed-time effects model. On the other hand, unemployment becomes significant in case of pooled panel, and GDP growth becomes insignificant in fixed-individual effects model. Hence, heteroscedasticity of residuals of pooled and fixed-individual effects models significantly change estimation of parameters.

Cluster 5: North and Continental Europe. Table 12 provided below presents parameters estimates and their standard errors in case of panel models and LASSO variables selection methods for Belgium, Denmark, Finland, France, Netherlands, Norway and Sweden. This cluster contains a total of 50 observations, and number of observations per country ranges from 4 to 10.

Table 12. Parameter estimates and their standard errors for cluster 5 using panel regression models and variable selection methods.

Variable	Pooled	FE-ind	FE-time	LASSO(λ)	Partial	Double
GDP growth	0.229^{***}	0.081^{*}	0.186	0.118	0.162^{**}	0.165^{***}
	(0.07)	(0.04)	(0.11)		(0.07)	(0.06)
Fertility rate	4.858^{***}	-1.910	7.658^{***}	2.372	3.069^{**}	3.177^{*}
	(1.46)	(1.36)	(1.47)		(1.45)	(1.73)
Tax revenue	-0.397^{***}	-0.222^{***}	-0.431^{***}	-0.259	-0.310^{***}	-0.310^{***}
	(0.06)	(0.07)	(0.06)		(0.07)	(0.05)
Culture expenditure	1.483^{*}	-0.353	1.481^{*}	0.000	0.246	0.300
	(0.79)	(0.88)	(0.76)		(0.74)	(0.87)
Unemoployment rate	0.659^{***}	-0.077	0.629^{***}	0.497	0.544^{***}	0.544^{***}
	(0.08)	(0.14)	(0.08)		(0.09)	(0.08)
R^2	0.644	0.326	0.749			

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Hence, in case of the fifth cluster, all regressors but culture expenditure are statisically significant with $\alpha = 0.05$ in pooled panel model. Fertility rate, tax revenue and unemployment rate are significant in fixed-time effects model, while only tax revenue is significant in fixed-individual effects model. The LASSO regression with penalty λ selected by cross-validation shrunks only estimate of culture expenditure to zero. On the other hand, using partialling-out method, statistically significantly regressors are GDP growth, fertility rate, tax revenue and unemployment rate, while in case of double selection there is no evidence to reject the null hypothesis that parameter estimate of fertility rate significantly differs from zero. In addition, using double and partial regression methods has not lead to lower standard errors of parameters estimates for all cases, except tax revenue.

Table 13. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross-dependence in panels for cluster 5.

Models / Tests	Durbin-Watson test	Levene's test	Pesaran's CD test
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	1.172^{***}	1.899	1.354
Fixed-individual effects	1.902	0.862	-0.818
Fixed-time effects	1.116^{***}	2.452^{**}	-1.763^{*}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 13 indicates that there is no evidence to reject the null hypothesis of Durbin-Watson test in case of the fixed-individual effects model. As in the case of the first, second, and third clusters, the null hypothesis of error homoscedasticity were not rejected. In addition, results of Pesaran's CD test showed no cross-dependence in panels.

In this case robust standard errors provided in Table 20 indicates that there is no changes in case of the fixed-time effects model, therefore autocorrelation and heteroscedasticity does not significantly affected results of estimation. In fixed-individual effects model GDP growth has become significant, while in pooled model fertility rate becomes insignificant after consideration to robust standard errors.

Cluster 6: Anglo-Saxon. Table 14 provided below presents parameters estimates and their standard errors in case of panel models and LASSO variables selection methods for Australia, Great Britain, United States. Group of Anglo-Saxon countries contains total of 16 observations, and number of observations per country ranges from 2 to 10.

Only culture expenditure and GDP growth are statistically significant regressors in case of the pooled model. Meanwhile, there were not found statistically significant regressors in both fixed effects models. LASSO regression with penalty λ shrunks estimates of fertility rate, tax revenue and unemployment rate to zero. On the other hand, partial and double regression methods indicates that only GDP growth is statistically significant. Furthermore, these methods did not provided more accurate estimates of parameters than pooled panel model.

Table 14. Parameter estimates and their standard errors for cluster 6 using panel regression models and variable selection methods.

Variable	**Pooled**	**FE-ind**	**FE-time**	**LASSO(λ)**	**Partial**	**Double**
GDP growth	-0.684***	-0.116	-1.643	-0.175	-0.547***	-0.657***
	(0.20)	(0.18)	(1.65)		(0.19)	(0.20)
Fertility rate	12.948	-9.808	60.727	0.000	7.701	7.701
	(8.65)	(6.40)	(65.92)		(9.57)	(11.94)
Tax revenue	0.370*	-0.173	0.016	0.000	0.173	0.173
	(0.20)	(0.51)	(0.35)		(0.24)	(0.34)
Culture expenditure	-11.185***	6.226	-3.466	-3.408	-6.473**	-6.473*
	(2.57)	(3.68)	(10.84)		(3.08)	(3.82)
Unemoployment rate	-0.066	-0.199	-0.196	0.000	0.288	0.338
	(0.33)	(0.18)	(1.51)		(0.36)	(0.38)
R^2	0.766	0.581	0.986			

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

The results of the Durbin-Watson and Levene test, presented in Table 15, indicates that after clustering, autocorrelation and heteroskedasticity problems of model residuals were solved in group of Anglo-Saxon countries. In addition, cross-dependence in panels in case of the pooled and fixed-time effects models was detected.

Table 15. Durbin-Watson test for autocorrelation, Levene's test for homogeneity of variance and Pesaran's CD test for cross-dependence in panels for cluster 6.

Models / Tests	**Durbin-Watson test**	**Levene's test**	**Pesaran's CD test**
The null hypothesis	$H_0 : \rho = 0$	$H_0 : \sigma_i^2 = const$	$H_0 : \rho_{ij} = \rho_{ji} = 0$
Pooled	1.707^{*}	1.146	2.406^{**}
Fixed-individual effects	1.743	2.224	-0.056
Fixed-time effects	1.776	0.552	-2.402^{**}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

As would be expected, since there were no evidence to reject the null hypothesis of homoscedasticity of residuals and has no detected evidence of autocorrelation, using robust standard errors has not led to significant changes for both fixed effects model in case of the cluster of Anglo-Saxon countries. On the other hand, fertility rate has become significant with at least $\alpha = 0.05$ in pooled panel.

CONCLUSIONS

1. Panel models for non-clustered countries has not allowed to identify significant relationships between income inequality and socio-economic indicators, and their residual errors do not satisfy the assumptions of panel regression analysis.
2. Using of the K-means on the standartized data with K = 6 resulted in countries being grouped according to their geographical location and the social models operating in them. Cluster analysis led to solve the problem of heteroscedasticity and in the case of some clusters helped to avoid autocorrelation of errors.
3. Partial variable selection methods based on LASSO regression not in all cases led to lower standard errors of parameter estimates, but confirmed a statistically significant relationship between GDP growth and income inequality in North European and Anglo-Saxon countries. Meanwhile, higher tax revenue is associated with lower income inequality in all clusters, with the exception of some Central European and Anglo-Saxon countries.
4. Higher fertility rate is related with lower income inequality in case of the some Central European countries and South European countries, where an emphasis on family policy. Meanwhile, the significant link between unemployment and income inequality identified in the North and some Continental European countries indicates that rising unemployment rates in these countries lead to increasing in income inequality.

Based on the modelling results obtained during this work, it is proposed to launch a broader discussion on one of the objectives to achieve a higher tax revenue share in the GDP in Lithuania. Econometric analysis of OECD countries has shown that the link between GDP growth and income inequality is ambiguous and in most cases insignificant, suggesting that GDP growth can not lead to public welfare until the government do not pursue inclusive economic policy. The proposed debate should include the questions of response of the governments to economic growth and changing demography, and financing of public sector. The literature of income inequality suggests that inadequate public financing and inefficiencies have led to a significant proportion of the population in some OECD countries not having equal access to services of health care and social protection. Children in these countries face educational inequality due to an inefficient net of educational institutions and lack of financing for some key areas such as pre-school education. Such tendencies have led to an increase in social exclusion between urban and rural populations, and unequal conditions of the starting point are destroying the economic potential of the country and may lead to a political or economic instability. Therefore, in order to reduce income inequality and social exclusion, the ideas of a supply-side economy which is based on the assumptions of market self-regulation should be refused, and a more prominent role of government in the national economy at all stages of the economic cycle should be provided.

References

Abu Sharkh, M., & Gough, I. (2010). Global welfare regimes: a cluster analysis. *Global Social Policy*, *10*(1), 27–58.

Agnello, L., & Sousa, R. M. (2014). How does fiscal consolidation impact on income inequality? *Review of Income and Wealth*, *60*(4), 702–726.

Allub, L., & Erosa, A. (2019). Financial frictions, occupational choice and economic inequality. *Journal of Monetary Economics*, *107*, 63–76.

Bakija, J., Cole, A., Heim, B. T., et al. (2012). *Jobs and income growth of top earners and the causes of changing income inequality: Evidence from us tax return data.* Williams College, Williamstown, MA.

Balcilar, M., Chang, S., Gupta, R., & Miller, S. M. (2018). The relationship between the inflation rate and inequality across us states: a semiparametric approach. *Quality & Quantity*, *52*(5), 2413–2425.

Banks, L. M., Kuper, H., & Polack, S. (2017). Poverty and disability in low-and middle-income countries: A systematic review. *PloS one*, *12*(12), e0189996.

Barro, R. J. (1999). *Inequality, growth, and investment* (Tech. Rep.). National bureau of economic research.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Bivens, J., & Mishel, L. (2013). The pay of corporate executives and financial professionals as evidence of rents in top 1 percent incomes. *Journal of Economic Perspectives*, *27*(3), 57–78.

Bonoli, G. (2007). New social risks and the politics of post-industrial social policies. In *The politics of post-industrial welfare states* (pp. 21–44). Routledge.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications.* Cambridge university press.

Ceriani, L., & Verme, P. (2012). The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, *10*(3), 421–443.

Charles-Coll, J. A. (2011). Understanding income inequality: concept, causes and measurement. *International Journal of Economics and Management Sciences*, *1*(3), 17–28.

Chernozhukov, V., Hansen, C., & Spindler, M. (2016). hdm: High-dimensional metrics. *arXiv preprint arXiv:1608.00354*.

Cingano, F. (2014). Trends in income inequality and its impact on economic growth.

Crenshaw, E. M. (1993). Polity, economy and technoecology: Alternative explanations for income inequality. *Social forces*, *71*(3), 807–816.

Dabla-Norris, M. E., Kochhar, M. K., Suphaphiphat, M. N., Ricka, M. F., & Tsounta, E. (2015). *Causes and consequences of income inequality: A global perspective.* International Monetary Fund.

Deaton, A. S., & Paxson, C. H. (1997). The effects of economic and population growth on national saving and inequality. *Demography*, *34*(1), 97–114.

Dobson, S., & Ramlogan-Dobson, C. (2012). Inequality, corruption and the informal sector. *Economics Letters*, *115*(1), 104–107.

Eikemo, T. A., Bambra, C., Joyce, K., & Dahl, E. (2008). Welfare state regimes and income-related health inequalities: a comparison of 23 european countries. *The European Journal of Public Health*, *18*(6), 593–599.

Čekanavičius, V., & Murauskas, G. (2002). *Statistika ir jos taikymai, ii knyga.* TEV.

Esping-Andersen, G., et al. (2002). A child-centred social investment strategy. *Why we need a new welfare state*, 26–67.

Gasparini, L., & Lustig, N. (2011). *The rise and fall of income inequality in latin america* (Tech. Rep.). Documento de Trabajo.

Gastwirth, J. L. (1972). The estimation of the lorenz curve and gini index. *The review of economics and statistics*, 306–316.

Guerin, B. (2013). Demography & inequality.

Gupta, S., Davoodi, H., & Alonso-Terme, R. (2002). Does corruption affect income inequality and poverty? *Economics of governance*, *3*(1), 23–45.

Gustafsson, B., & Johansson, M. (1999). In search of smoking guns: What makes income inequality vary over time in different countries? *American sociological review*, 585–605.

Hanson, J. K. (2013). Loyalty and acquiescence: Authoritarian regimes and inequality outcomes. In *2010 apsa annual meeting paper.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Higgins, M., & Williamson, J. G. (2002). Explaining inequality the world round. *Japanese Journal of Southeast Asian Studies*, *40*(3), 268–302.

Hoeller, P., Joumard, I., Bloch, D., & Pisu, M. (2012). Less income inequality and more growth–are they compatible?: Part 1: mapping income inequality across the oecd.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kaasa, A. (2005). Factors of income inequality and their influence mechanisms: a theoretical overview. *University of Tartu Faculty of Economics and Business Administration Working Paper*(40).

Kar, S., & Saha, S. (2012). Corruption, shadow economy and income inequality: evidence from asia.

Kuznets, S. (1955). Economic growth and income inequality. *The American economic review*, *45*(1), 1–28.

Kuznets, S. (1963). Quantitative aspects of the economic growth of nations: Viii. distribution of income by size. *Economic development and cultural change*, *11*(2, Part 2), 1–80.

Lancee, B., & Van de Werfhorst, H. G. (2012). Income inequality and participation: A comparison of 24 european countries. *Social science research*, *41*(5), 1166–1178.

Larch, M. (2012). Fiscal performance and income inequality: Are unequal societies more deficit-prone? some cross-country evidence. *Kyklos*, *65*(1), 53–80.

Lazutka, R. (2003). Gyventojų pajamų nelygybė. *Filosofija. Sociologija*(2), 22–29.

Lazutka, R., Juška, A., & Navickė, J. (2018). Labour and capital under a neoliberal economic model: Economic growth and demographic crisis in lithuania. *Europe-Asia Studies*, *70*(9), 1433–1449.

Lazutka, R., Poviliūnas, A., & Žalimienė, L. (2018). Espn thematic report on inequalities in access to healthcare: Lithuania.

Lee, C.-S. (2005). Income inequality, democracy, and public sector size. *American Sociological Review*, *70*(1), 158–181.

Litwin, C. (1998). Trade and income distribution in developing countries. *rapport nr.: Working Papers in Economics*(1998).

Mankiw, N. G. (2013). Defending the one percent. *Journal of Economic Perspectives*, *27*(3), 21–34.

Meschi, E., & Vivarelli, M. (2009). Trade and income inequality in developing countries. *World development*, *37*(2), 287–302.

Nantob, N., et al. (2015). Income inequality and inflation in developing countries: An empirical investigation. *Economics Bulletin*, *35*(4), 2888–2902.

NAO. (2019). *National Audit Office of Lithuania Conference SIGNALS, Discussion: What changes are needed to reduce income inequality in Lithuania?* https://www.lrt.lt/mediateka/irasas/2000085681/signals-2019-diskusija-kokiu-pokyciu-reikia-kad-pajamu-nelygybe-lietuvoje-mazetu. ([Online; accessed 05-January-2020])

Nielsen, F., & Alderson, A. S. (1997). The kuznets curve and the great u-turn: income inequality in us counties, 1970 to 1990. *American Sociological Review*, 12–33.

Obolenskaya, P., & Hills, J. (2019). Flat-lining or seething beneath the surface? two decades of changing economic inequality in the uk. *Oxford Review of Economic Policy*, *35*(3), 467–489.

OECD. (2005). *Society at a Glance 2016: OECD Social Indicators.* https://www.oecd-ilibrary.org/docserver/soc_glance-2016-16-en.pdf?expires=1578224964&id=id&accname=guest&checksum=7313528931576DF841A586C93DA40E55. ([Online; accessed 05-January-2020])

Partridge, J. S., Partridge, M. D., & Rickman, D. S. (1998). State patterns in family income inequality. *Contemporary Economic Policy*, *16*(3), 277–294.

Raudenbush, S. W., & Eschmann, R. D. (2015). Does schooling increase or reduce social

inequality? *Annual Review of Sociology*, *41*, 443–470.

Riley, A. (2014). *The social thought of emile durkheim*. Sage Publications.

Ruger, J. P., & Kim, H.-J. (2006). Global health inequalities: an international comparison. *Journal of epidemiology & community health*, *60*(11), 928–936.

Sila, U., & Dugain, V. (2019). Income, wealth and earnings inequality in australia.

Skučienė, D., Lazutka, R., Čižauskaitė, A., & Markevičiūtė, J. (2018). *Socialinių išmokų vaidmuo mažinant skurdą ir pajamų nelygybę „naujųjų" socialinės rizikos grupių gyvenimo kelyje*. Vilniaus universiteto leidykla.

Stiglitz, J. E. (2009). Gdp fetishism. *The Economists' Voice*, *6*(8).

Stiglitz, J. E. (2016). Inequality and economic growth. In *Rethinking capitalism* (pp. 134–155).

Stone, C., Trisi, D., Sherman, A., & Debot, B. (2015). A guide to statistics on historical trends in income inequality. *Center on Budget and Policy Priorities*, *26*.

Syll, L. P. (2014). Piketty and the limits of marginal productivity theory. *Real-world economics review*, *69*(1), 36–43.

Sylwester, K. (2002). Can education expenditures reduce income inequality? *Economics of education review*, *21*(1), 43–52.

Tao, Y., Wu, X., Zhou, T., Yan, W., Huang, Y., Yu, H., … Yakovenko, V. M. (2019). Exponential structure of income inequality: evidence from 67 countries. *Journal of Economic Interaction and Coordination*, *14*(2), 345–376.

Ulu, M. I. (2018). The effect of government social spending on income inequality in oecd: a panel data analysis. *Uluslararası Ekonomi Siyaset İnsan ve Toplum Bilimleri Dergisi*, *1*(3), 184–202.

Veal, A. J. (2016). Leisure, income inequality and the veblen effect: cross-national analysis of leisure time and sport and cultural activity. *Leisure Studies*, *35*(2), 215–240.

Weinberg, D. H. (2004). Income data quality issues in the annual social and economic supplement to the current population survey. *Washington: US Census Bureau*, *126*.

Wright, E. O. (2000). Real utopian proposals for reducing income and wealth inequality. *To appear in Contemporary Sociology*.

A. Appendix

A.1. Residual analysis of models for non-clustered data

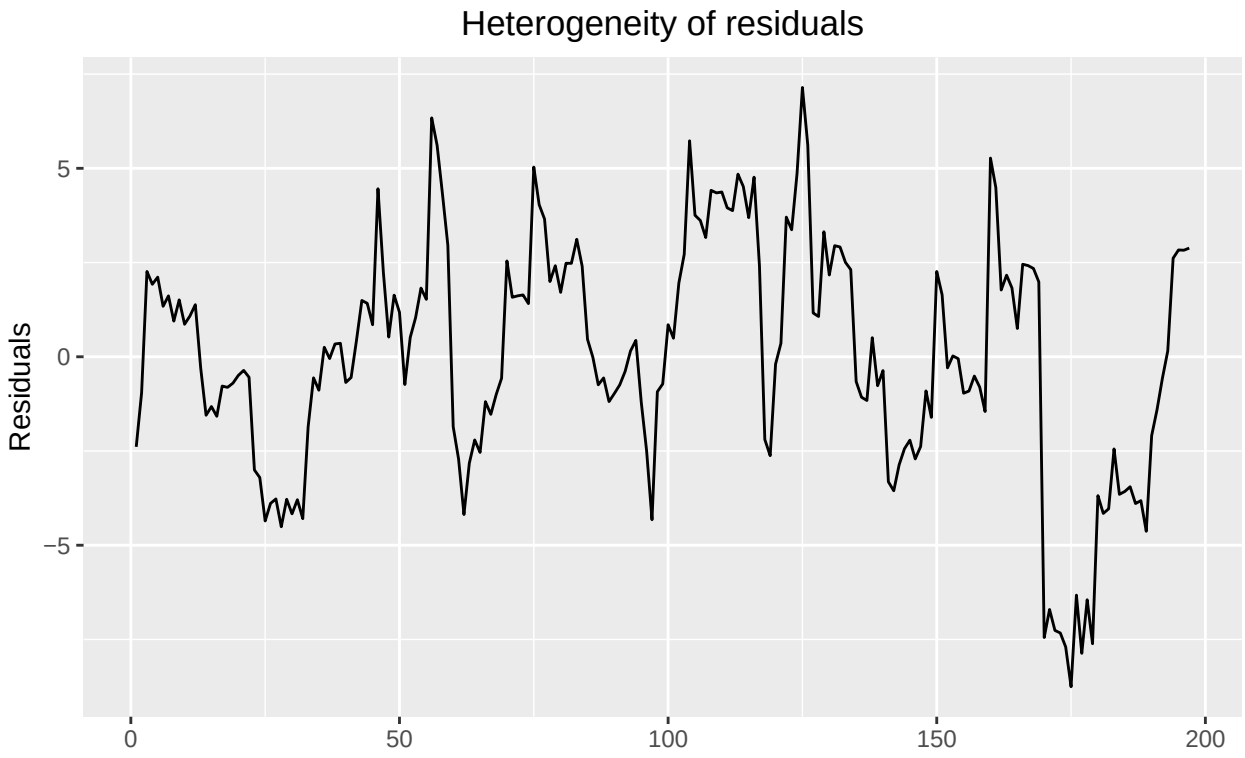

Figure 8. Residuals of all observations in case of the pooled model.

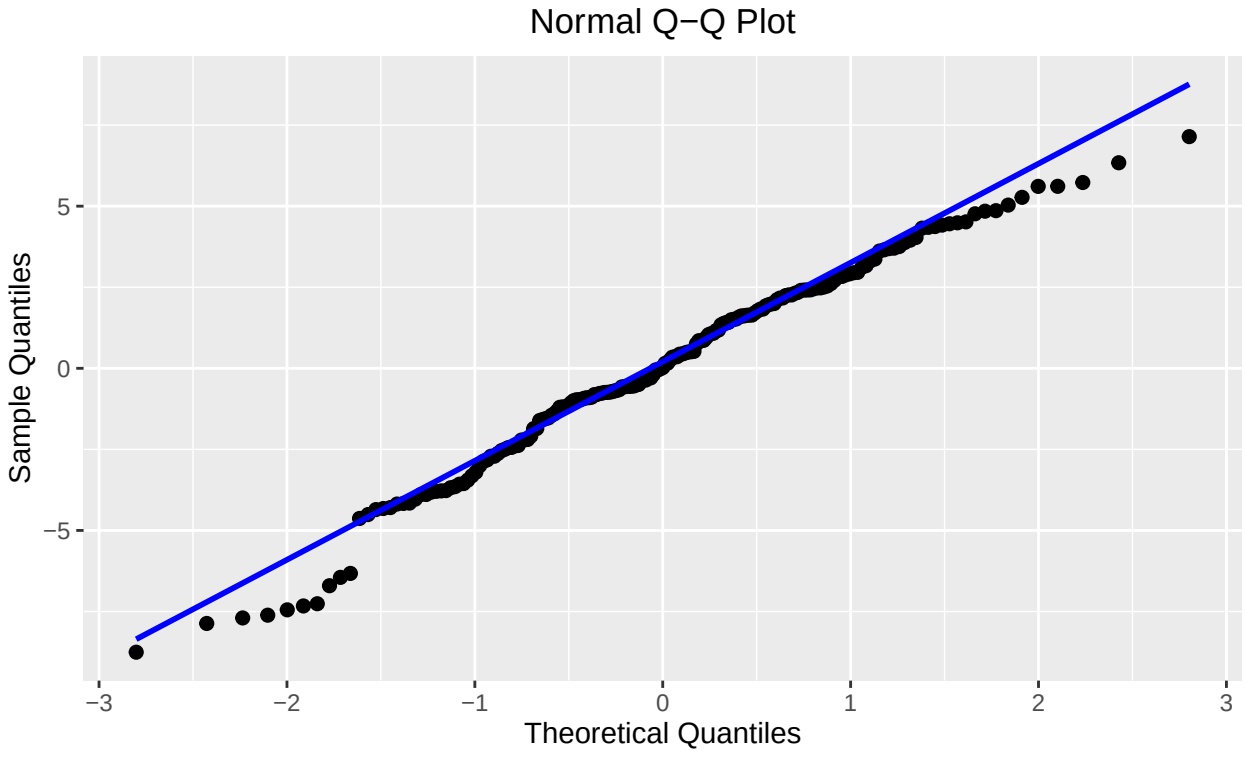

Figure 9. Normal Q-Q plot in case of the pooled model.

Figure 10. Residuals of all observations in case of the fixed-individual effects model.

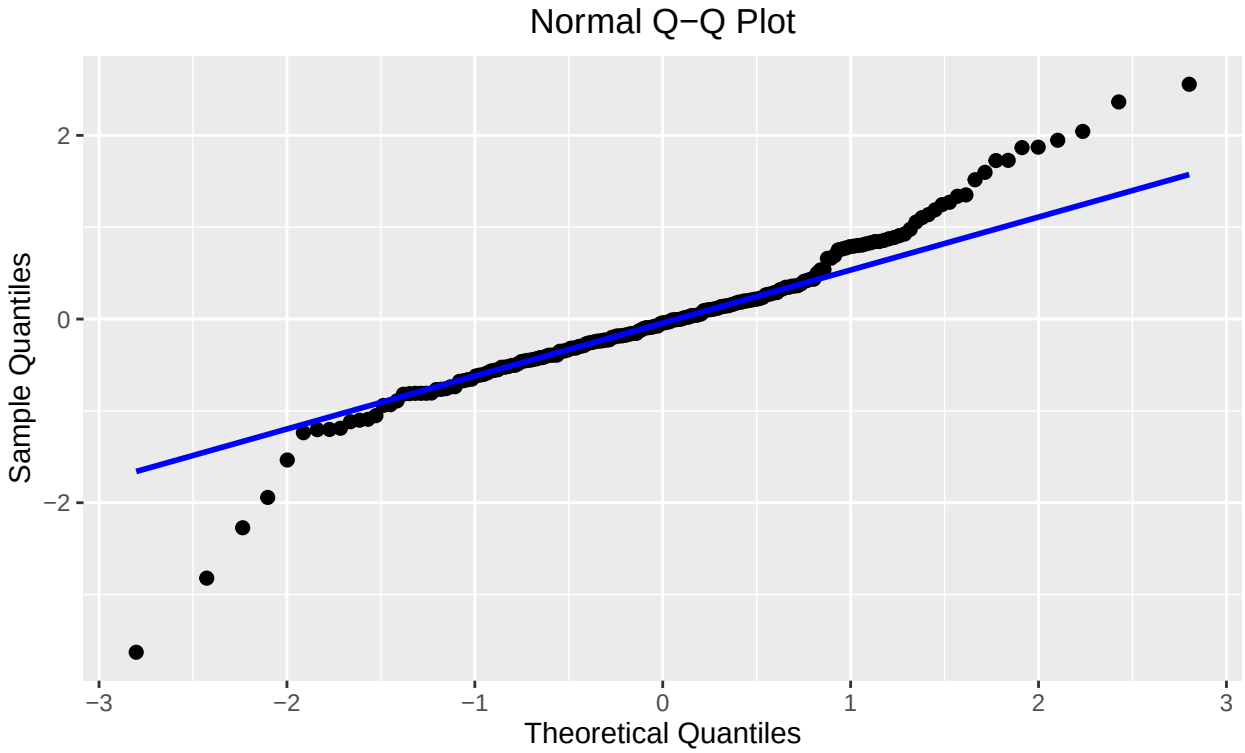

Figure 11. Normal Q-Q plot in case of the fixed-individual effects model.

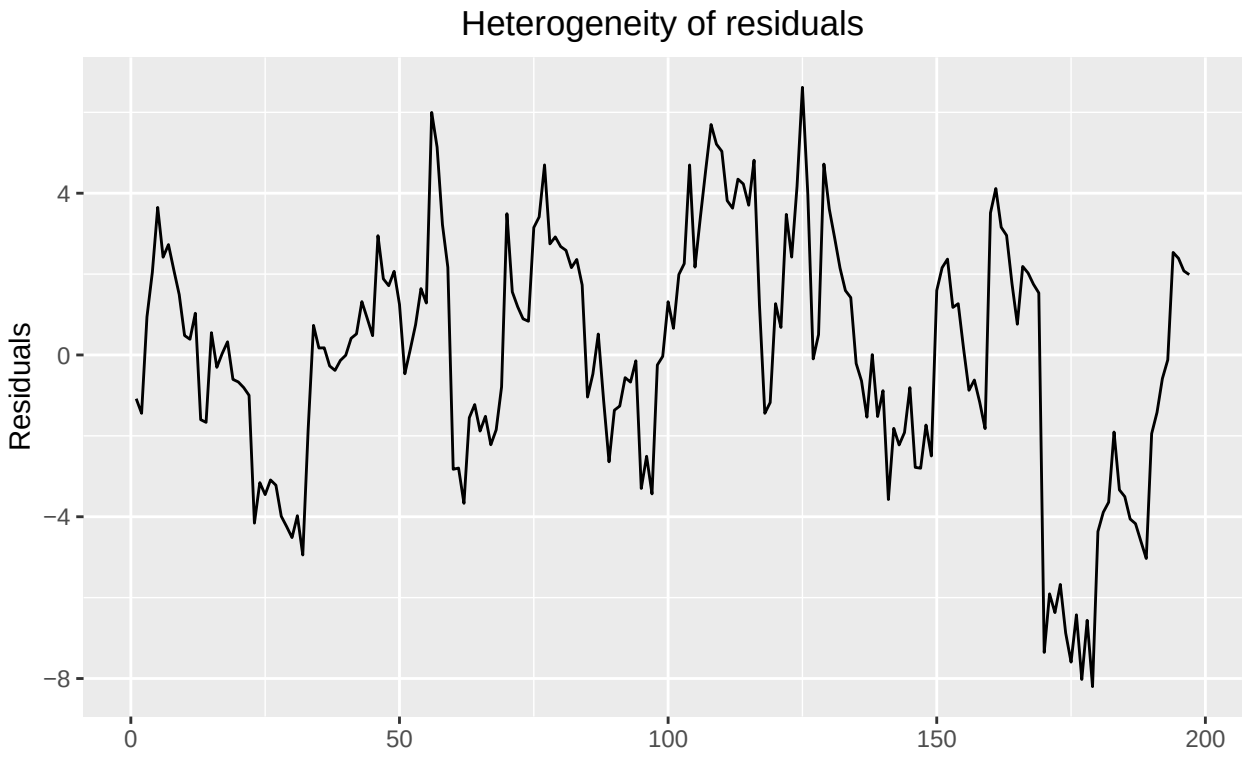

Figure 12. Residuals of all observations in case of the fixed-time effects model.

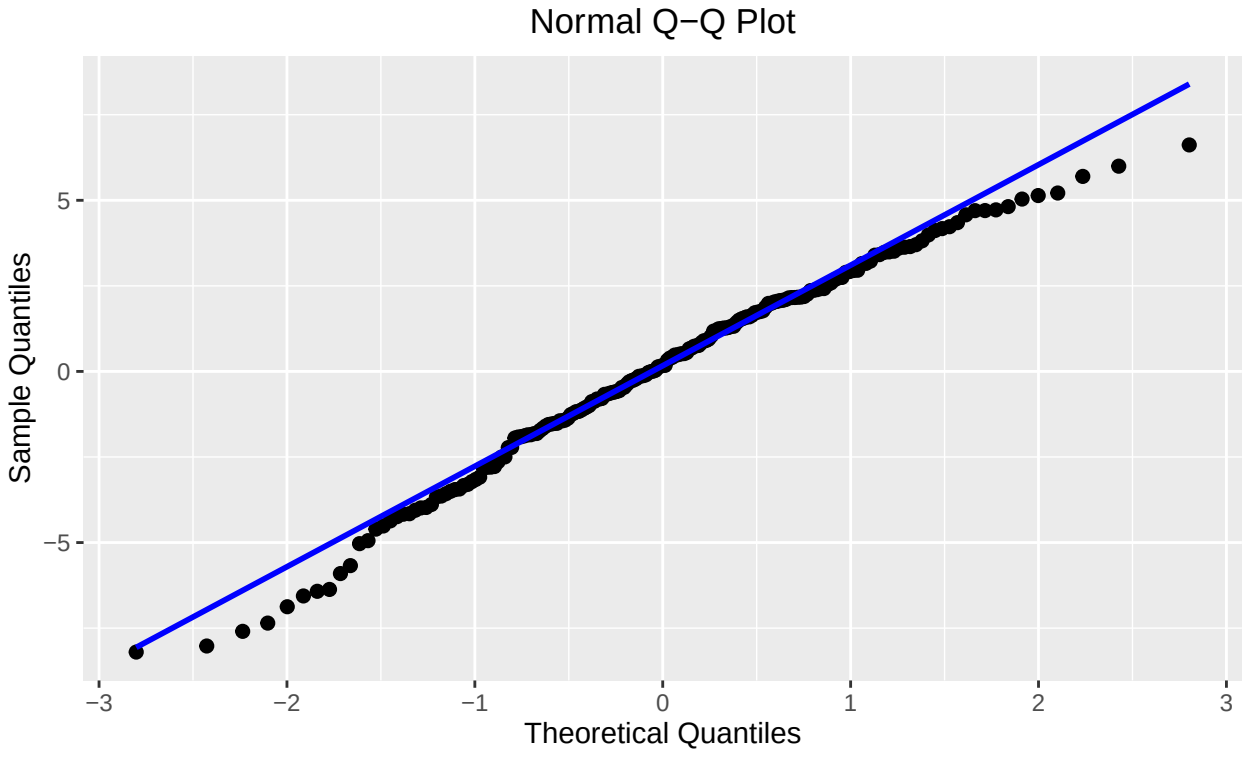

Figure 13. Normal Q-Q Plot in case of the fixed-time effects model.

A.2. Robust standard errors of models for clustered data

Table 16. Parameters estimates and their robust standard errors for the cluster 1.

Variable	Pooled	Fixed-individual effects	Fixed-time effects
GDP growth	-0.155*	-0.134	-0.146
	(0.08)	(0.11)	(0.18)
Fertility rate	0.170	1.023	-10.369
	(2.84)	(4.47)	(6.39)
Tax revenue	-0.794***	-0.734**	-1.801***
	(0.18)	(0.35)	(0.49)
Culture expenditure	0.261	1.744	1.277
	(0.77)	(2.78)	(0.81)
Unemoployment rate	-0.347***	-0.316**	-1.093***
	(0.10)	(0.15)	(0.39)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 17. Parameters estimates and their robust standard errors for the cluster 2.

Variable	Pooled	Fixed-individual effects	Fixed-time effects
GDP growth	-0.047	-0.003	-0.125
	(0.08)	(0.06)	(0.15)
Fertility rate	-12.129**	-0.869	3.093
	(5.00)	(3.10)	(5.12)
Tax revenue	0.193	0.027	0.885
	(0.27)	(0.23)	(0.85)
Culture expenditure	11.570***	3.711**	-4.653
	(1.73)	(1.70)	(10.42)
Unemoployment rate	0.027	0.240	-0.996***
	(0.14)	(0.17)	(0.36)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 18. Parameters estimates and their robust standard errors for the cluster 3.

Variable	Pooled	Fixed-individual effects	Fixed-time effects
GDP growth	0.059	0.081*	0.045
	(0.04)	(0.04)	(0.06)
Fertility rate	-3.997**	4.007*	-3.405
	(1.99)	(2.32)	(4.34)
Tax revenue	-0.122***	-0.108**	-0.130***
	(0.04)	(0.05)	(0.04)
Culture expenditure	-0.472	-0.544	-0.321
	(0.44)	(0.61)	(0.59)
Unemoployment rate	0.030	0.119***	0.015
	(0.03)	(0.02)	(0.03)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 19. Parameters estimates and their robust standard errors for the cluster 4.

Variable	Pooled	Fixed-individual effects	Fixed-time effects
GDP growth	0.049	0.080	-0.026
	(0.07)	(0.05)	(0.17)
Fertility rate	-5.415*	2.760*	-13.238***
	(3.29)	(1.62)	(3.33)
Tax revenue	0.186***	-0.287*	0.087
	(0.05)	(0.15)	(0.11)
Culture expenditure	-0.629	1.594***	-0.446
	(1.02)	(0.55)	(0.60)
Unemoployment rate	-0.222**	0.133*	-0.436***
	(0.10)	(0.07)	(0.15)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 20. Parameters estimates and their robust standard errors for the cluster 5.

Variable	Pooled	Fixed-individual effects	Fixed-time effects
GDP growth	0.229***	0.081***	0.186*
	(0.05)	(0.03)	(0.10)
Fertility rate	4.858*	-1.910	7.658***
	(2.71)	(1.45)	(2.15)
Tax revenue	-0.397***	-0.222***	-0.431***
	(0.09)	(0.07)	(0.09)
Culture expenditure	1.483	-0.353	1.481*
	(1.17)	(0.74)	(0.78)
Unemoployment rate	0.659***	-0.077	0.629***
	(0.11)	(0.19)	(0.09)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 21. Parameters estimates and their robust standard errors for the cluster 6.

Variable	Pooled	Fixed-individual effects	Fixed-time effects
GDP growth	-0.684***	-0.116	-1.643
	(0.14)	(0.22)	(1.55)
Fertility rate	12.948**	-9.808*	60.727
	(5.85)	(5.15)	(82.52)
Tax revenue	0.370*	-0.173	0.016
	(0.20)	(0.39)	(0.54)
Culture expenditure	-11.185***	6.226	-3.466
	(2.01)	(5.20)	(13.64)
Unemoployment rate	-0.066	-0.199	-0.196
	(0.40)	(0.19)	(0.99)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$