CHARLES UNIVERSITY IN PRAGUE

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Bachelor thesis

2016 **Martin Stárek**

CHARLES UNIVERSITY IN PRAGUE

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Martin Stárek

The Effect of Globalization on the Income Inequality

Bachelor thesis

Prague 2016

Author: Martin Stárek
Supervisor: doc. Ing. Tomáš Cahlík, CSc.

Academic Year: 2015/2016

Bibliographic note

STÁREK, Martin. *The Effect of Globalization on the Income Inequality*. 58 p. Bachelor thesis. Charles University, Faculty of Social Sciences, Institut of Economic Studies. Supervisor doc. Ing. Tomáš Cahlík, CSc.

Abstract

In this thesis, we explore the effect of globalization on the income inequality. We examine some features of methodology used in the majority of research on this topic that can have significant impact on results but they are not addressed in the publicly available research. Firstly, we proposed a new method of normalization that creates more stable data and created a new simple index of globalization using this method. This index then yielded more consistent results than the standard globalization indices. Secondly, we found out the most significant variable in a composite index can have no economic or logical interpretation. This was the case with the effect of mobile cellular subscriptions per 100 people on the income inequality. This means results of composite indeces should be interpreted carefully and a better analysis is probably estimating effects of all underlying variables individually. Moreover, we found that underlying variables in a composite globalization index can have opposite effects on the income inequality. The effects than cancel out, at least partly, and this can lead to smaller, statistically less significant results. Nevertheless, the overall effect of globalization on the income inequality, though statistically not significant, appeared to be negative. This is the case especially in developing countries. The effect is higher (closer to 0 or positive) for developed countries but the sign is ambiguous. Finally, we found the evidence of Kuznets curve and also found out taxation schemes have probably significant impact on the post-tax income inequality, though we were not directly controlling for it (we dealt with it by fixed effect estimation).

Abstrakt

Tato práce zkoumá efekt globalizace na přijímovou nerovnost. Zaměřujeme se na některé prvky metodologie použité ve většině prací na toto téma, které mohou mít významný vliv na výsledky, ale které nejsou ve veřejně dostupném výzkumu adresovány. Zaprvé jsme navrhli nový způsob normalizace, který vytváří stabilnější data, a za použití této metody jsme vytvořili nový jednoduchý globalizační index. Tento index vykazoval konzistentnější výsledky než standardní globalizační veličiny. Zadruhé, zjistili jsme, že nejvýznamnější veličina v kompozitním indexu nemusí mít žádné ekonomické nebo logické vysvětlení. To byl případ efektu počtu přihlášených mobilních účtů na 100 obyvatel na přijímovou nerovnost. To znamená, že výsledky kompozitních indexů by měly být interpretovány opatrně a lepší přístup k analýze je pravděpodobně odhad efektů všech dílčích veličin individuálně. Navíc jsme zjistili, že dílčí veličiny kompozitního globalizačního indexu mohou mít opačný směr efektů na přijímovou nerovnost. Tyto efekty se pak vzájemně odečtou, přinejmenším částečně, což vede k menším a statisticky méně významným výsledkům. Nicméně, celkový efekt globalizace na přijímovou nerovnost, ačkoli není statisticky významný, je nejspíše záporný. To platí zejména v případě rozvíjejících se zemích. Pro rozvinuté země je efekt větší (blížší k 0 nebo kladný), ale jeho znaménko je nejasné. Nakonec jsme našli důkazy o existenci Kuznětsovy křivky a také o tom, že způsoby danění mají pravděpodobně významný vliv na podaňovou nerovnost, ačkoli jsme příslušnou veličinu do našich modelů nezahrnuli (vyřešili jsme je pomocí metody fixních efektů).

Klíčová slova

globalizace, přijímová nerovnost, Gini koeficient, KOF index, DHL index, Kuznetsova křivka, panelová data

Keywords

Globalization, Income Inequality, Gini Coefficient, KOF Index, DHL Index, Kuznets Curve, Panel Data

Range of thesis: 79 127 characters

Declaration of Authorship

1. The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

2. The author hereby declares that all the sources and literature used have been properly cited.

3. The author hereby declares that the thesis has not been used to obtain a different or the same degree.

Prague 26.7.2016 Martin Stárek

Acknowledgments

I would like to express my gratitude to doc. Ing. Tomáš Cahlík, CSc. for the leading of this thesis, his help and useful comments on the topic. I am also very grateful to Mgr. Tomáš Křehlík M.A. for his time spent on consulting some issues of this thesis.

Institute of Economic Studies
Bachelor thesis propsal

Description of the Thesis:

This thesis will focus mainly on the within-country income inequality in the world. The thesis will attempt to answer questions like:

1) What is the effect of globalization on the income inequality?
2) Does it depend on the level of development of the country?
3) Is the relationship robust or sensitive to measures of the key variables?
4) How do underlying variables influence the effect of composite globalization indices?
5) Is there some evidence of the Kuznets curve?

We will also analyse globalization indices and methodology used in their composition.

The measures of inequality that will be taken into account are mainly Gini coefficients, Theil Index (a wage inequality measure in the manufacturing sector), and Ratio 90:10. Globalization will be then measured by both standard and widely-used KOF globalization index and a recent one, DHL Index.

The outcome of this thesis will be analysis of both the effect of globalization on the income inequality and the methodology used in this field of research, mainly concerning globalization measures.

Main sources of data will be the World Data Bank, the Standardized World Income Inequality Database and individual releases of KOF and DHL globalization indices. Missing data will be treated by linear normalization.

Preliminary content:

1) Introduction, problem setting
2) Inequality measures
3) Globalization measures
4) Normalization
5) Methodology
6) Base model (i.e. for all countries)
7) Different levels of development
8) Conclusion

Literature overview:

BERGH, Andreas and NILSSON, Therese, 2010. Do liberalization and globalization increase income inequality? European Journal of Political Economy. Vol. 26. pp. 488-505.

DREHER, Alex, 2006. Does globalization affect growth? Evidence from a new index of globalization. Applied Economics. Vol. 38, Issue 20. pp. 1091–1110.

DREHER, Alex and GASTON, Noel, 2008. Has Globalization Increased Inequality? Review of International Economics. Vol. 16, Issue 3. pp. 516-536.

GALBRAITH, James K. and KUM, Hyunsub, 2002. Inequality and Economic Growth: Data Comparisons and Econometric Tests. University of Texas Inequality Project Working Paper Number 21.

GHEMAWAT, Pankaj and ALTMAN, Steven A., 2014. DHL Global Connectedness Index.

KUZNETS, Simon, 1955. Economic Growth and Income Inequality. The American Economic Review. Vol. 45, Issue 1.

SOLT, Frederic, 2014. The Standardized World Income Inequality Database. Working paper. SWIID Version 5.0.

WOOLDRIDGE, Jeffrey M. 2013. Introductory Econometrics: A Modern Approach. 5th edition. South-Western, Mason, OH: Cengage Learning. 910 p. ISBN 1-111-53104-8.

Contents

Introduction

Globalization and income inequality are two modern phenomena that are nowadays being often discussed and a great attention is paid to them. Possibly because of their both social as well as welfare effects. Too big income inequality can have negative social impacts so income equality, a situation when everyone has roughly the same income, is something modern nations often want to achieve. This is of course impossible in most cases but at least no too large differences are desired. For instance, in September 2015 more than 150 world leaders adopted the United Nations' 2030 Agenda for Sustainable Development, including the Sustainable Development Goals (United Nations, 2015). The first goal on the list is to "*end poverty in all its forms everywhere*" and the tenth is to "*reduce inequality within and among countries*". This illustrates the close attention paid to the inequalities in the world.

On the other hand, globalization is also something nations want to achieve most time. For example, countries create and involve in many trade unions and other partnerships. We know organisations such as the European Union, the North America Free Trade Agreement or the Organization of Petroleum Exporting Countries. Countries involve in such international interaction because it is often believed that states can benefit from it. But involving in international trade is only a part of the complex process of globalization. There is also social, informational and political dimension of globalization. More precise definitions follow in next sections.

However, globalization can have many consequences and also negative ones such as displacing local cultures or destroying environment by the large-scale transportation. Of course there are also negative economic consequences of globalization. One of them much discussed in the recent years, at least on the academic field, is the income inequality. More precisely, the within-countries income inequality. There are for example more than 100 published research papers that explore this relationship using the KOF Index of globalization (Potrafke, 2014). However, researches in this field often have insignificant or even contradictory results.

In this thesis we explore the effect of globalization on the income inequality using the most recent publicly available panel data on around 120 countries all around the world and the time periods from 1995 up to 2012. A very important part of this thesis is also developing a new normalization method for measuring globalization and analysing its performance. Finally, we add variables that detect for the Kuznets curve which we will discuss later.

1 Literature Review

The effect of globalization on the income inequality is of a big academic interest. The main findings of surveying empirical literature of more than 100 studies exploring the relationship between globalization and income inequality using the KOF Index of globalization are that globalization increases economic growth on one hand, but on the other hand it increases within-country inequalities (Potrafke, 2014).

Also direct empirical studies confirm that globalization magnifies the income inequality (Dreher and Gaston, 2008). They used the Theil index, a household income inequality dataset and different dimensions of the KOF Index of globalization in an econometric analysis with a battery of robustness tests and found out they are positively related, especially in the case of OECD countries. On the other hand, no robust effect of globalization on the income inequality was found in poorer countries. Moreover, no evidence of the Kuznets curve was found either in developed or developing countries.

Another empirical study using the KOF Index for globalization and the Gini coefficient for inequality was not able to find any robust relationship concluding the effects of globalization on the income inequality differ because of some fixed characteristics of each country such as their institutional setting. But they say a slight evidence show, contrary to the previously discussed study, that increased globalization leads to higher inequality in less developed countries (Atif, Srivastav, Sauytbekova and Arachchige, 2012).

Then, Lall, Jaumotte, Papageorgiou, Topalova et al. (2007) linked a set of inequality variables with trade and financial globalization (measured mainly as foreign direct investments), controlling for variables such as technology progress measured as the share of information and communications technology capital in the total capital stock. They found out the technological progress has much higher and significant effect on the income inequality than globalization. Moreover, they argue that the overall influence of globalization is a result of offsetting effects of trade and financial globalization.

The main paper we want to compare our results with is the one published in the European Journal of Political Economy with the name *Do liberalization and globalization increase income inequality?* (Bergh and Nilsson, 2010). In this study, using KOF Index for globalization and Gini coefficients for income inequality, authors find that freedom to trade internationally has a robust positive effect on the income inequality. Next, economic freedom (economic globalization) has a positive effect on the inequality mainly in the rich countries, whereas social globalization is more significant in poorer countries. Finally, political globalization turned out to have no effect on the income inequality. They tested this results with a battery of sensitivity tests and controlling for GDP, human capital and dependency ratio in the baseline analysis.

On the other hand, other authors arrived to the opposite results. It was showed that economic freedom (most importantly in terms of trade liberalization) can have actually negative effect on the inequality measured by the Gini coefficient (Berggren, 1999).

We can see there is more evidence that the relationship between globalization and income inequality is positive, but the empirical evidence is sometimes contradictory. There are many reasons for this. First of all, the main problem is both the data availability and quality because both key variables, globalization and income inequality, are hard to measure. Next, when it comes to the income inequality, it does not vary a lot. In particular, both pre-tax and post-tax Gini coefficients series have both low variance and a high serial correlation. And finally, the major problem is that different forces or types of globalization have probably different effects on the inequality. In addition, the effects can vary depending on the development levels of the countries.

2 Measuring Inequality

2.1 Gini Coefficient

Gini coefficient is a measure of income inequality in a population and it ranks between 0 and 1 (alternatively, between 0 and 100). Gini of 0 means everyone in the population has the same income and so there is a perfect income equality in the population. The theoretical maximum of 1 can be achieved in an infinite population where all the income is concentrated to only one person, representing perfect income inequality. In particular, it measures how much does the income distribution in the population deviates from the perfectly equal distribution.

The basis for the Gini coefficient calculation is the Lorenz curve which is used in economics as well as in other fields of study to describe inequality. *"It can be expressed as follows:*

$$L(y) = \int_0^y x \, dF(x)/\mu$$

where $F(y)$ is the cumulative distribution function of ordered plants, and μ is the average plant size" (Damgaard and Weiner, 2000). In our case, ordered plants are individuals in the population ordered by their income which is referred to as the plant size. In the case when there is a perfect equality in the population, the Lorenz curve is a straight line, called the equality line. When inequality is present, the Lorenz curve is bent downwards.

"The total amount of inequality can be summarized by the Gini coefficient, which is the ratio between the area enclosed by the line of equality and the Lorenz curve, and the total triangular area under the line of equality. The Gini coefficient is most easily calculated from unordered plant size data as the "relative mean difference," i.e., the mean of the difference between every possible pair of individuals, divided by the mean size

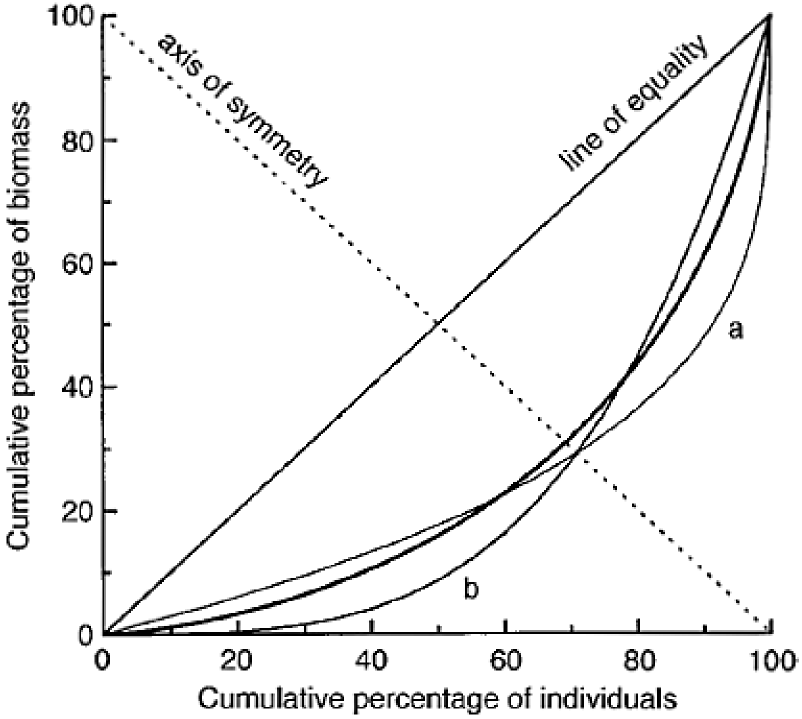
$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2\mu}$$

Alternatively, if the data is ordered by increasing plant size

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)x_i'}{n^2\mu}$$

where x_i' is the size of plant i" (Damgaard and Weiner, 2000).

Figure 1: Three Lorenz curves. A symmetric case (bold line) and two asymmetric cases (*a* and *b*). Note: In our case the vertical axis is the cumulative percentage of income.



Source: Damgaard and Weiner (2000)

When it comes to data on the Gini coefficient, we face many problems. First of all, there is no united institution that would regularly compute the coefficients. Moreover, there is no united methodology. This makes the data very hard both to collect and compare. The leading source of income inequality data is nowadays the Standardised World Income Inequality Database. It accumulates data from the leading available sources and makes them broadly comparable (Solt, 2014). A great advantage of this database is that it contains data on both pre-tax and post-tax Gini coefficients. The data used in our analysis are from the Standardized World Income Inequality Database.

2.2 Theil Index

The University of Texas Inequality Project (UTIP) compiles a within-country inequality statistics based on the Theil Index and the United Nations Industrial Development Organisation's (UNIDO) Industrial Statistics. This statistics is a measure of inequality in the manufacturing pay and according to the authors *"these (measures) provide indicators of inequality that are more stable, more reliable, and more comparable across countries (than standard income inequality measures)"* (Galbraith and Kum, 2002). They argue that pay is a substantial part of income and that *"the manufacturing pay has been measured with reasonable accuracy as a matter of official routine in most countries around the world for nearly 40 years."* In other words, this statistics measures the wage differential in the manufacturing sector arguing this is a kind of stable proxy variable in determining the within-country income inequality.

The mathematical computation of this statistics is based on the *"between-groups component of the Theil's T statistic, an entropy measure whose functional form is defined as*

$$T = \sum \left(\frac{Y_i}{Y}\right) T_i + \sum \frac{Y_i}{Y} log\left(\frac{Y_i/_Y}{N_i/_N}\right) = T^W + T^B$$

where T^W and T^B indicate within-group and between-group inequality measures respectively. N and Y stand for total employment and total pay respectively, and subscript i denotes group identity. We capture T^B as our inequality measure, where groups are defined as categories within the UNIDO industrial classification codes. Theil (1972) has shown that T^B is a consistent lower-bound inequality measure, where the within-groups component is unobserved" (Galbraith and Kum, 2002).

We will refer to this variable as to the Theil Index and use it as our second within-country income inequality measure. The interpretation of the Theil Index is similar to the Gini coefficient, value of zero means perfect equality. However, there is no upper bound, but higher values indicates higher inequality. As the authors recommend, due to its skewed distribution, we will use the logarithmic transformation of this variable which will be denoted by lTheil. However, the data are available up to the year 2008 only so our models with lTheil as the dependent variable use periods from 1995 to 2008 only.

2.3 Ratio 90:10

As the third measure of within-country income inequality we employ the Ratio 90:10. It is the ratio of income share held by the highest 10% to the income share held by the lowest 10%. In a perfectly equal population, when both the highest and the lowest

10% of population hold exactly 10% of income, this should be 1 which is the theoretical minimum of this variable. Higher values denote higher inequality which is when either the highest 10% of population holds more than 10% of income or the lowest 10% of population holds less than 10% of income (or both at the same time). The data are from the Poverty and Equity Database of the World Data Bank (The World Bank DataBank, 2016).

3 Measuring Globalization

3.1 Definition of Globalization

It is very important to define precisely all the concepts used in any research. Globalization is one of the trickiest phenomena to define and different authors follow different approaches. We include four of them.

"Globalization is a process of growing interaction and interdependence between economies, societies and nations across large distances" (Vujakovic, 2010).

"Globalization is a transformation in the spatial organization of social relations and transactions – assessed in terms of their extensity, intensity, velocity and impact – generating transcontinental or interregional flows" (Ghemawat and Altman, 2014).

"Globalization is a process of creating networks of connections among actors at multicontinental distances, mediated through a variety of flows including people, information and ideas, capital, and goods" (Dreher, 2006).

"It is a process that erodes national boundaries, integrates national economies, cultures, technologies and governance, and produces complex relations of mutual interdependence" (Dreher, 2006).

We will not involve much in the philosophical debate about the precise nature of globalization. However, for our next discussion it is very important to mention that all of these definitions view globalization in some sense as the actual level of connectedness between countries, not as their ranking in terms of the connectedness (meaning co definition involves the order of countries).

An important point to mention is also that any composite index of globalization cannot fulfil this definitions perfectly. For example the KOF Index contains also data on McDonald's restaurants and Ikea shops which favours the "western style" of globalization which is generally insufficient. Any index of globalization serves rather as a proxy to measure globalization, never perfectly capturing it. The core of this problem lies in the definition of globalization which is mathematically very vague. In order to be able to precisely measure globalization we would need to define a set of observable variables and declare them as the globalization. The problem is that abstract terms defining globalization cannot be easily observed. Based on the nowadays used definitions of globalization we cannot do much about this, but we should be aware of it and take it into our considerations.

3.2 KOF Index

To measure the level of globalization we will use also multiple variables. One of the most popular ones, used in the majority of research, is the KOF Index of globalization, provided by the Swiss Federal Institute of Technology Zurich (Dreher, 2006). This index provides data on three dimensions of globalization (economic, social and political), as well as on the overall level of globalization. It covers a number of countries and data are available from 1970. All dimensions as well as subdimensions of the overall index, their units and weights are attached in Appendix 9. To make underlying variables comparable it uses the panel percentile normalization. We will discuss its properties later.

3.3 DHL Index

As the second measure of globalization we use the DHL Connectedness Index developed by Ghemawat and Altman (2014). This is much more modern and recent indicator in comparison with the KOF Index as it is computed only since 2005. It consists of only 2 dimensions, the breadth and the depth of globalization which build up the overall globalization index.

When compared to the KOF Index, the DHL Index consists of less variables and does not include as outdated variables as number of international letters per capita (which are included in the KOF Index). On the other hand, it does not include any variable measuring the political dimension of globalization. Detailed information on the index's structure, its components and weights is attached in Appendix 10. To normalize the different data into one comparable index, the method of panel normalization is used, too, as in the case of the KOF Index.

3.4 Normalization

3.4.1 The Need for Normalization

When creating a composite index, underlying variables need to be normalized in order that we can compare their effects with each other. Without normalization this is not directly possible because different variables have different units, distributions, ranges etc. Also when estimating the effects of underlying variables separately, we want to be able to say that for example the globalization variable Trade has relatively larger effect on income inequality than the globalization variable Migrants because we treat globalization as one complex phenomenon of interest that however consists of more partial variables.

In order to do that, we need to transform values of a variable with both arbitrary units and values into one standardized scale, for example from 0 to 100. There are more ways how to do that. We will discuss some them in the following sections, because they can have significant influence on the results and the performance of the regression analysis.

3.4.2 Linear Panel Normalization

This is the case when the transformation is linear and across the whole panel. Usually the normalization is of the form:

$$V_{it,norm} = \frac{V_{it} - V_{min}}{V_{max} - V_{min}} \times 100 \, ,$$

where $V_{it,norm}$ denotes the normalized value of variable V for the individual i in period t, V_{it} denotes the original value of variable V for individual i in period t and V_{min} and V_{max} denote the minimal and maximal value, respectively, of the variable V across all individuals and all time periods. Similar normalization is used for example in early publications of the KOF Index of globalization (they multiply the result by 10 only, but for our discussion this does not matter). Let us now briefly discuss the properties of this normalization.

Firstly, values of each variable range from 0 to 100. The value 0 means that the original value was the minimum across all i and t and the value 100 means it was the maximum across all i and t. All other values range in between.

Secondly, and this is the main problem of every panel normalization, the values are recalculated each year, also retrospectively. If the source variable (the already transformed variable) is published for the period from 1970 up to the year of publication, but we are interested in the years 1990 to 2000 only, then the data differ depending on the year of publication because generally different minima and maxima are used for the normalization. It basically means that data for a fixed period in the past differ systematically based on from which year we “view” them.

This is very important and could be a serious problem, especially because many variables that build up globalization in either index are trending. So there is a significant probability that in each additional year there is for example a new maximum for some variable, leading to a reduction of all past values because of the term $V_{max} - V_{min}$ in the denominator.

This phenomenon generally violates the first assumption of a regression analysis of constant parameters (see FD.1 or FE.1 in the following sections). More specifically, with a single concrete dataset the assumption is of course fulfilled but results for the same period with different datasets systematically differ which we think is undesirable.

Mathematically expressed, if we assume the true relationships (the true beta coefficients) are constant but values of the variable x_1, for example, are changing systematically based on the "viewing year", we get:

$$y = \beta_0 + (a\beta_1)\left(\frac{x_1}{a}\right) + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

$$\text{where } a = \frac{V_{max,new} - V_{min}}{V_{max,old} - V_{min}} > 1 \text{ for } V_{max,new} > V_{max,old}$$

So the estimated and reported coefficient for x_1, $a\beta_1$, gets systematically larger.

In the case when x_1 is a complex index variable, when a single underlying variable reaches new maximum in the new year (other maxima and minima remaining constant), the effect on the reported coefficient depends on the weight the particular variable has in the index and on the direction of the effect it has on the regressand.

Of course, this can be fixed by normalizing the variables directly in each specific case, but firstly this would be very demanding, especially for as large amount of variables as in the case of the KOF Index. But, most importantly, the same analysis of different periods would still not be directly comparable. Anyway, this is not the case as all publicly available and broadly used indices of globalization normalize their variables across the whole panel.

3.4.3 Percentile Panel Normalization

Other way how to compare effects of different variables is a percentile panel normalization. This method transforms values of a variable into percentiles of its distribution. This assigns each value a score from 0 to 100 according to how many percent of all scores (across all individuals as well as years) were lower. This method is used in the recent KOF Index editions as well as in the DHL Index. However, this approach actually causes even more problems, especially in interpretation of results.

First of all, the normalized score depends exclusively on how many countries are below the country of interest and not on the actual level of globalization. This point collides with the definition of globalization as the “level of connectedness” and not as the rank of countries. For example, let us consider a country whose value of an underlying variable is somewhere in the middle. If its value increases, potentially significantly and potentially having significant impact on the income inequality, its normalized score can remain constant depending on whether the change leads to skipping another country or not and this is undesirable based on the cited definitions of globalization.

Second, the problem of recalculating values retrospectively remains. As an example, let us consider the KOF Index for all countries in the period 1990 to 2006. Each year there is a new edition of the index. Let us focus on eight editions from 2009 to 2014. When we calculate the average KOF indices across all countries and the whole period 1990 to 2006, we get the following values:

Table 1: Average KOF indices for 1990 to 2006.

Year of edition	2009	2010	2011	2012	2013	2014	2015	2016
Average KOF 1990 to 2006	55.08	49.85	49.34	49.11	48.80	48.39	48.17	47.99
Difference	-	-5.23	-0.51	-0.23	-0.31	-0.41	-0.22	-0.18

Source: Author

We can see the average KOF Index in the fixed period 1990 to 2006 decreases steadily depending on the year of index edition. This is caused by the fact that each year a new set of values is added and those new values are, on average, higher (probably as many of the underlying variables are trending). So the normalized scores for past values get lower, as a lower percentage of all values is below their levels.

The effect on the reported coefficient is however not that straightforward. For a composite index, the total change of its reported coefficient depends on the individual changes of partial components that caused this decrease. In any case, this pattern is undesirable and can lead to misleading results.

3.4.4 Linear Normalization with Respect to Years

Solution to problems caused by the panel normalization, either linear or percentile, is to use another method to normalize variables. In this thesis we will use and discuss linear normalization with respect to years. That means the transformation is exactly the same as in the case of linear panel normalization but maxima and minima used in this transformation are considered across all countries in the given year only. That causes the normalized scores to be constant across time, regardless of for how many time periods we carry out the normalization or from which year we "view" a fixed past period.

Then, normalized value of 100 means the corresponding individual country has the highest level of the variable across all countries in the particular year and the normalized value of 0 means the lowest level in the particular year.

Let us now discuss some examples on how those methods of normalization behave. For example, what if there are Olympic Games in China in year t, causing the variable Tourists to rise extremely? In both panel normalizations, this reduces scores on Tourists for all other countries in all periods. In our linear normalization with respect to years only, this affects scores on Tourists for all other countries in the year t only, leaving other years unchanged.

The problem is that for a country on the opposite part of the world (Chile, for example), Olympic Games in China could be an exogenous change leaving everything in Chile unaffected. And still, its globalization score decreases. The difference between panel and year-to-year normalization is that in the panel normalization this event affects all periods whereas in the linear normalization with respect to years it affects only one period, because why should the globalization level in 2020 be influenced by the fact that Olympic Games attracted a lot of people in 2010? Or, most importantly, why should be the globalization level of country A in the year 1970 be different when measuring it in 2010 or in 2016? We assume the globalization levels can be measured and that they are constant numbers. This is fulfilled by the linear normalization with respect to years.

Based on this discussion, we believe linear normalization with respect to years has better conditions to measure globalization in a regression analysis and we will compare its performance with standard globalization measures.

3.4.5 Maximal Globalization

However, all three methods of normalization suffer from the fact that the normalized value of 100 can mean actually very low level of globalization because it does not reflect the actual level of the corresponding maximum which can be arbitrarily low or different each time in the year-to-year normalization.

One way to solve this would be to somehow normalize also the maxima (and minima). This would mean to define something like a maximal globalization. We would normalize maxima so that they would be the same, this maximal value would have to be the maximal value possible for the particular variable. Then the normalized score of 100 would correspond to the maximal possible value of the variable, leading to comparability of maxima across time. For variables measured in percent, this would basically mean no normalization.

Let us have a look on the variable Migrants, representing the percentage share of foreigners in the population. In this case the normalized score of 100 would then denote 100% share of foreigners in the population. This is of course nonsense and no country could achieve such a maximum (naturally because then there would be no original population). Finally, maximal globalization has no interpretation or practical meaning because for example there is no upper boundary of how many tourists can visit a country.

On the other hand defining minimal globalization is easy as all variables can be 0 without any problems and interpretation of this is simple that the country is isolated.

This consideration shows there is nothing like a maximal globalization (though globalization indices can theoretically achieve their maximum). The most easily and straightforward interpretable method of normalization is therefore in comparison with other countries regardless of the actual level, as is used in all mentioned methods. This is another point that verifies the normalization used in our thesis. But it is an example why globalization is so badly mathematically operable.

3.5 New Measure of Globalization

In this section we propose an alternative simple measure of globalization that is based on the linear normalization with respect to years. We employ a measure that covers three main aspects of globalization, each represented by two straightforward variables. All data in this section are from the World Development Indicators Database of the World Data Bank (The World Bank DataBank, 2016).

Economic globalization

Trade: This variable measures the sum of exports and imports of goods and services as a percentage of GDP. It is traditionally the most frequently used and simplest measure of globalization.

FDI: The sum of inflows and outflows of foreign direct investment measured as percentage of GDP.

Social globalization

Tourists: This measures expenditures by international inbound tourists plus expenditures by international outbound tourists as percentage of GDP.

Migrants: This variable measures the international migrant stock in percent of population. The data are for one in five years only so we used linear interpolation to predict the most likely values in the middle years.

Informational (technological) globalization

Phone: The number of mobile cellular subscriptions per 100 people. This captures the communication technology availability.

Internet: The last globalization variable shows the number of internet users per 100 people.

To be able to compare effects of variables between each other we use the year-to-year linear normalization. That is, in each year, we find the minimum and maximum for each variable, subtract the minimum from the value that is being normalized and divide the result by the difference of maximum minus minimum, as described in the above section.

In this thesis we firstly estimate the effect of each of this variables individually to see their effect in the composite index, but for the purpose of comparison with established globalization indices we then compile them into a single index, too. We use a simple average with equal weights, that is adding all 6 variables together and dividing the result by 6. This leads to a single index variable describing globalization, let us name it the G Index, with values between 0 and 100, higher values denoting higher level of globalization. Correlations between the G Index, 2016 KOF Index and 2014 DHL Index in the years 2005 and 2010 are shown in Appendix 1. We can see that they are all highly positively correlated.

4 Control Variables

The selection of control variables in our thesis is based on both the empirical literature and theory. We have basically 4 control variables and variables on GDP both actual and squared to test for presence of the Kuznets curve as proposed by the American economist Simon Kuznets (1955). All data are also from the World Development Indicators Database of the World Data Bank (The World Bank DataBank, 2016), except for data on Education which are from the Barro-Lee Database (Barro and Lee, 2016).

Agriculture: It shows employment in the agriculture as percent of total employment. This corrects for the distribution of sectorial share of employment as it is a very important factor that affects the income. It is believed that lower employment in agriculture (shifting it to other sectors) could reduce the income inequality by increasing the income of low-earning people (Lall, Jaumotte, Papageorgiou, Topalova et al., 2007).

Dependency ratio: It is the ratio of people younger than 15 or older than 64 to the working age population expressed in percent. This corrects for the age distribution as the income differs for different stages in life. This is often used in the corresponding research and higher shares of young or old people is expected to rise the total within country income inequality (Bergh and Nilsson, 2010).

Education and Education squared: The traditional variable in this manner is the labour force with tertiary education in percent. However, publicly available series on this variable surprisingly suffer from a significant lack of data. Thus instead we employ the variable on average years of tertiary education in the population above 25 years. This has the advantage of reflecting also the quality of the education in terms of years at school. The data are also available for one in five years only so we also used linear interpolation for the middle years. Controlling for educational distributions is very important as the income is often positively influenced by education. However, the effect on the income inequality can be ambiguous. More skilled workers may reach the wage premium and thus increasing the inequality. But if everyone in the population is skilled and has the higher wage, then the inequality is lower (Bergh and Nilsson, 2010). We therefore include also the variable Education squared and expect it to be negative, whereas the level of Education we expect to be positive.

Technology: In the research paper provided by Lall, Jaumotte, Papageorgiou, Topalova et al. (2007) this turned out to be one of the most significant variables affecting the income inequality. To measure the role of technology in the economy we use the same variable, high-technology exports as percent of manufactured exports. This is connected with the demand for skilled labour enabling corresponding workers to reach a wage premium and thus increasing the income inequality (Lall, Jaumotte, Papageorgiou, Topalova et al., 2007).

GDP and GDP squared: We also control for any effects on the income inequality driven by the national income. Bergh and Nilsson (2010) found in their analysis that the level of GDP has a significant positive effect on the income inequality. Moreover, Simon Kuznets (1955) proposed a theory that the relationship between level of development of a country and its income inequality should be a curve similar to inverted U. This means income inequality should be lower for low- and high-income countries and higher for countries in between. We therefore include both GDP and GDP square to test for this. The variables are natural logarithms of real GDP per capita (with 2005 as the base year).

Development: This is a dummy variable with value 1 if the country is developed or 0 if the country is developing. We divide countries into two groups according to their level of development to be able to estimate the effect of globalization on different income levels. For this purpose we used categories set by The World Bank DataBank (2016). As developed countries we treat the ones in categories high income and upper middle income and as developing we treat countries in categories low income and lower middle income. We have 131 developed and 83 developing countries in our sample.

4.1 Missing Data

Our dataset has a lot of missing values which is the cost of using publicly available data. It is important to fix this problem in order to use the dataset in our regression analysis as we are left with around 100 complete observations without any data completion which is very insufficient.

We use a simple method of linear interpolation in order to fix the problem of missing data. It is based on connecting existing data points for each country by a linear curve, replenishing missing values as points on that curve. This may be a good approximation as most of the variables we use are trending. Moreover, linear interpolation is used in the completion of both the KOF and DHL indices of globalization so we will stick to this method.

In the case when available data cover longer period that 1995 to 2012, we interpolated the whole dataset (using also the data we don’t directly need) so we did not lose any available information.

Overview of number of original and interpolated values can be seen in Appendix 11. High numbers of interpolated values for variables Migrants and Agriculture is because they are only available in 5 years intervals. Values for other variables miss more or less randomly. Together we interpolated less than 12% of values in the whole dataset. If we don’t count biased cases of Migrants and Agriculture, we interpolated even less than 4% of the dataset. The benefit of this is around 10 times more observations.

5 Methodology

In this analysis we use panel data to discover and quantify the relationships between variables. This kind of data is very useful because first it contains a lot of information and second it allows us to overcome problems caused by some kind of endogeneity, in particular the unobserved fixed effects.

There are many methods of estimation in the framework of panel data regression analysis. We will have a look at those we employ in our analysis. All methodology in this section is based on the textbook by Wooldridge (2012).

5.1 First Differences Estimation

We can write a general unobserved effects model as

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \;\; i = 1, \dots, N, t = 1, \dots, T \qquad [1]$$

where the subscript i denotes the i-th cross-sectional unit, subscript t denotes the time period, $\beta_1, \dots, \beta_k$ are the parameters to estimate and a_i is the unobserved, individual-specific time-constant effect. Now, when we subtract the equation for the time $t-1$ from the equation for the time t, we get the first-differenced equation

$$\Delta y_{it} = \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}, \;\; i = 1, \dots, N, t = 2, \dots, T \qquad [2]$$

where we don't have the unobserved effect a_i anymore so we have dealt with the endogeneity problem so we can use pooled OLS estimation. When $T = 2$, after first differencing we are left with a single cross-section. When $T > 2$, the assumptions needed are as follows:

$FD.1$: We can write the population model as in [1], where the β_j are the parameters to estimate and a_i is the unobserved effect.

$FD.2$: We have a random sample from the cross section.

$FD.3$: Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables.

$FD.4$: For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero: $E(u_{it}|\boldsymbol{X}_i, a_i) = 0$.

$FD.5$: The variance of the differenced errors, conditional on all explanatory variables, is constant: $Var(\Delta u_{it}|\boldsymbol{X}_i) = \sigma_u^2, t = 2, \dots, T$.

$FD.6$: For all $t \neq s$, the differences in the idiosyncratic errors are uncorrelated (conditional on all explanatory variables): $Cov(\Delta u_{it}, \Delta u_{is}|\boldsymbol{X}_i) = 0,\ t \neq s$.

$FD.7$: Conditional on $\boldsymbol{X}_i$, the Δu_{it} are independent and identically distributed normal random variables.

Under the first four assumptions the first-difference are unbiased and consistent with a fixed T as $N \to \infty$. Moreover, for consistency it is sufficient to assume that Δx_{itj} are uncorrelated with Δu_{it} for all $t = 2, \dots, T$ and $j = 1, \dots, k$. Under the first six assumptions, pooled OLS standard errors and test statistics are asymptotically valid. Under the full set of FD assumptions the FD estimates are normally distributed and t and F statistics are exactly t and F distributed (Wooldridge, 2012).

The first difference estimator is most powerful when original errors follow random walk or a very high serial correlation of the first order because then the differenced errors are independent and identically distributed.

5.2 Fixed Effects Estimation

This is an advanced panel data estimation method which also leads to the elimination of unobserved effects a_i that are assumed to be correlated with the independent variables. It is based on the time demeaning transformation. More precisely, we average the equation [1] across time and get

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \cdots + \beta_k \bar{x}_{ik} + a_i + \bar{u}_i, \qquad [3]$$

where

$$\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$$

and so on. Now we subtract the time-averaged equation, [3], from the original equation, [1], to get the time-demeaned equation

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{i1} + \cdots + \beta_k \ddot{x}_{ik} + \ddot{u}_{it}, \qquad t = 1, \dots, T,$$

where we see the unobserved fixed effect disappeared, because it is fixed over time so it is differenced away after the time-demeaning. Now we use pooled OLS on this equation because the unobserved errors are not correlated across time anymore. Let us now list the assumptions needed for this method.

$FE.1$: For each i, the model is as in the equation [1] where the β_j are the parameters to estimate and a_i is the unobserved effect.

$FE.2$: We have a random sample from the cross section.

$FE.3$: Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables.

$FE.4$: For each t, the expected value of the idiosyncratic error (u_i) given the explanatory variables in all time periods and the unobserved effect is zero: $E(u_{it}|\boldsymbol{X}_i, a_i) = 0$.

$FE.5$: $Var(u_{it}|\boldsymbol{X}_i, a_i) = Var(u_{it}) = \sigma_u^2$, for all $t = 1, \ldots, T$.

$FE.6$: For all $t \neq s$, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and a_i): $Cov(u_{it}, u_{is}|\boldsymbol{X}_i, a_i) = 0$.

$FE.7$: Conditional on $\boldsymbol{X}_i$ and a_i, the u_{it} are independent and identically distributed as $Normal(0, \sigma_u^2)$.

Under the first four assumptions – which are identical to the assumptions for the first-differencing estimator – the fixed effects estimator is unbiased. Under these same assumptions, the FE estimator is consistent with fixed T as $N \rightarrow \infty$. Under all seven assumptions the FE estimator is normally distributed, and t and F statistics have exact t and F distributions so we can employ the standard inference (Wooldridge, 2012).

5.3 Random Effects Estimation

In the previous method the aim was to eliminate the unobserved fixed effect a_i because it was assumed to be correlated with the independent variables. In the random effects approach we assume it is uncorrelated with all independent variables in all time periods. Under this assumption we can write the population model as follows:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + v_{it}, \qquad i = 1, \ldots, N, t = 1, \ldots, T,$$

where $v_{it} = a_i + u_{it}$ is the so called composite term. We could include the intercept β_0 into the equation because it does not get eliminated after the transformation anymore and, more importantly, we can assume the expected value of the unobserved effect a_i is 0 as long as the intercept is included. Because a_i is included in the composite error term in each period, we face a serial correlation problem. The correlation appears to be $Corr(v_{it}, v_s) = \sigma_a^2/(\sigma_a^2 + \sigma_u^2)$ for $t \neq s$, where $\sigma_a^2 = Var(a_i)$ and $\sigma_u^2 = Var(u_i)$. We can now use the transformation called quasi-demeaning:

$$y_{it} - \theta\bar{y}_i = \beta_0(1-\theta) + \beta_1(x_{it1} - \theta\bar{x}_{1i}) + \cdots + \beta_k(x_{itk} - \theta\bar{x}_{ki}) + (v_{it} - \theta\bar{v}_i),$$

where $\theta = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2}$. The errors in this equation are serially uncorrelated so we can estimate it by pooled OLS. However, we seldom know the value of parameter θ. But we can estimate it as $\hat{\theta} = 1 - \{1/[1 + T(\hat{\sigma}_a^2/\hat{\sigma}_u^2)]\}^{1/2}$, where $\hat{\sigma}_a^2$ and $\hat{\sigma}_u^2$ are consistent estimators of σ_a^2 and σ_u^2, respectively, resulting in the so called feasible generalized least squares estimation. The quasi-demeaning transformation with the use of $\hat{\theta}$ is called the random effects estimation.

Some of the assumptions needed for the random effects estimation are the same as those needed for the fixed effects estimation. Three of them has to be changed:

$RE.3$: There are no perfect linear relationships among the explanatory variables.

$RE.4$: In addition to $FE.4$, the expected value of a_i given all explanatory variables is constant: $E(a_i|\boldsymbol{X}_i) = \beta_0$.

$RE.5$: In addition to $FE.5$, the variance of a_i given all explanatory variables is constant: $Var(a_i|\boldsymbol{X}_i) = \sigma_a^2$.

Under $FE.1$, $FE.2$, $RE.3$ and $RE.4$ the random effects estimator is consistent and asymptotically normally distributed as $N \to \infty$ for fixed T. Under those first four RE assumptions plus $RE.5$ and $FE.6$, the random effects estimator's standard errors and test statistics are asymptotically valid so we can perform usual statistical inference (Wooldridge, 2012).

5.4 Choosing Between Estimation Methods

As we see, there are more methods to estimate the population parameters of interest. As each of them needs slightly different assumptions to be met, we can distinguished between them to determine the best one in terms of efficiency.

When choosing between FD and FE estimator, we can use the serial correlation in the idiosyncratic errors u_{it} as a criterion (under the homoscedasticity assumption as this is needed for the efficiency comparison). In the case when u_{it} and $u_{i,t-1}$are serially uncorrelated, fixed effects estimation is more efficient than first differencing. On the other hand, when $u_{it} = u_{i,t-1} + \varepsilon_{it}$, where ε_{it} is independent and identically distributed with 0 mean and variance σ_ε^2, i.e. when u_{it} follow a random walk, then first differencing is better than fixed effects because $\Delta u_{it} = \varepsilon_{it}$ are serially uncorrelated (Wooldridge, 2012).

We can test whether Δu_{it} are serially correlated (i.e. whether $\Delta u_{it} = \rho \Delta u_{i,t-1} + \varepsilon_{it}$) by estimating [2] by pooled OLS, get the residuals $\widehat{\Delta u}_{it}$ and regress them on $\widehat{\Delta u}_{i,t-1}$ by using pooled OLS as well. The coefficient on $\widehat{\Delta u}_{i,t-1}$ from this regression, $\hat{\rho}$, is a consistent estimator of ρ and we can also perform standard t test to test its statistical significance. When we find no serial correlation between Δu_{it} and $\Delta u_{i,t-1}$, FD should be used because the original errors probably follow a random walk. On the other hand, when the original errors are uncorrelated, the serial correlation between resulting differenced errors can be showed to be -0.5. Thus if we find significant negative serial correlation in the differenced errors, FE should be used (Wooldridge, 2012, p. 470).

We can also distinguish between the FE and RE estimations. The main difference is that FE does allow the unobserved effects, a_i, to be correlated with the explanatory variables whereas RE estimation rules this possibility out and is in this case more efficient. We can apply the Hausman test to see whether the key RE assumption, $Cov(x_{itj}, a_i) = 0, t = 1, \dots, T, j = 1, \dots, k$ is fulfilled. Moreover, RE estimation is generally more efficient than pooled OLS so it is preferred to it (Wooldridge, 2012).

5.5 Serial Correlation and Heteroscedasticity Robust Inference

Homoscedasticity, i.e. the situation when the error term of the model, u, has a constant variance, σ_u^2, as well as no serial correlation between the error term across time (both conditionally on all explanatory variables), are two important assumptions that need to be met in order for the usual statistical inference, mainly the standard errors and t and F statistics, to be valid. We address the serial correlation problem by testing for it in the differenced errors and then choosing between fixed effects estimation and first differencing. Moreover, we will use robust standard errors to deal with some potentially remaining serial correlation. They are robust also to arbitrary form of heteroscedasticity (Wooldridge, 2012).

Important is that the number of cross sectional units is large enough, meaning it should be *"substantially larger than the number of periods"* (Wooldridge, 2012, p. 511) to justify the use of clustered standard errors. This is fulfilled in our case as we have up to 18 time periods but around 100 individuals (depending on the particular model).

6 Models

6.1 Setting Models

In this section we set up the main model for our analysis. As the dependent variable we use either the post- or pre-tax Gini coefficients, Theil index or the Ratio 90:10 resulting in 4 variations of our model (as a result, we can say whether our results are sensitive or not). Independent variables consist of our 6 measures of globalization and a set of control variables as was discussed in corresponding sections. Moreover, we include a full set of year dummy variables. That is a dummy variable for all time periods except for the first one serving as the base period. This is a normal component for a panel data analysis as it accounts for a linear time trend, allowing different intercepts in each period. So our model is as follows:

$$\begin{aligned} Inequality_{it} &= \beta_0 + \beta_1 Trade_{it} + \beta_2 FDI_{it} + \beta_3 Tourists_{it} + \beta_4 Migrants_{it} \\ &\quad + \beta_5 Phone_{it} + \beta_6 Internet_{it} + \beta_7 lGDP_{it} + \beta_8 lGDP\_sq_{it} + \beta_9 Agric_{it} \\ &\quad + \beta_{10} Depend_{it} + \beta_{11} Educ_{it} + \beta_{12} Educ\_sq_{it} + \beta_{13} Tech_{it} \\ &\quad + \beta_{14} y1996_t + \cdots + \beta_{30} y2012_t + a_i + u_{it}, t = 1995, \dots, 2012 \end{aligned}$$

First of all, we decide which method of estimation to use. We perform the Hausman test to decide between FE and RE and the test for serial correlation proposed by Wooldridge (2012) to test whether FE or FD is better in terms of serial correlation. We perform both tests for all models.

For post-tax Gini and Ratio 90:10 as dependent variables, the Hausman test rejected the null hypothesis of zero correlation between the unobserved effect and regressors at the 10% level, so the fixed effect estimation will be used. For pre-tax Gini and lTheil as dependent variables, there is a strong evidence in favour of the null hypothesis. We are not able to reject it so the random effect estimation will be used with those dependent variables. An overview of exact test's statistics and p-values are enclosed in Appendix 2.

These results have actually some very interesting interpretation. We can see that for inequality variables based on post-tax measures (post-tax Gini and Ratio 90:10, that measures the actual wealth – that means after taxation) the fixed effect estimation was chosen. On the other hand, for variables based on pre-tax measures (pre-tax Gini and lTheil, we here remind that Theil is a measure of wage inequality in the manufacturing sector) the random effects model is a better alternative. An explanation for this can be that taxation schemes, something that can be viewed as constant and thus captured by the unobserved effects, are important for explaining the post-tax inequality, so we need to remove them by fixed effect estimation. On the other hand, they are irrelevant when it comes to the pre-tax inequality, so the unobserved effects are not important and can be actually viewed as random (and also other fixed unobserved characteristics are hereby shown to be random but they are unimportant relative to the taxation in post-tax inequality models).

However, we cannot say anything about the effect of taxation on the post-tax inequality as we are not explicitly including it in our model. We can only say it is important and next research could possibly try to find out its actual effect.

In the next step we test for serial correlation in our models. We proceed as described in the previous section. All results are listed in Appendix 3. In the case of lTheil and Ratio 90:10 as dependent variables, we have found significant negative serial correlation in the differenced errors. However, when we tested the null hypothesis that the correlation estimator, $\hat{\rho}$, actually equals -0.5 (which is equivalent to testing the serial correlation of underlying undifferentiated errors equals 0), we rejected the null hypothesis in both cases. That means some serial correlation is present in the original model. More importantly, we also rejected the null hypothesis that the underlying serial correlation of undifferentiated errors equal 1 in both cases. That means original errors do not follow random walk so the estimation method of first differencing is not so efficient. In the case of Ratio 90:10 as the dependent variable, original undifferentiated errors follow probably significant negative serial correlation (based on that estimated serial correlation of differentiated errors is in (-0.5, -1)) so first differencing does not help at all.

In the case of lTheil as the dependent variable, we should compare fixed effects with first differences to see what is better as the serial correlation of original undifferentiated errors is probably somewhere in (0,1) (because the estimated serial correlation of differentiated errors is in (-0.5, 0)). Complete regression results with first differences can be seen in Appendix 5. The outcome is that model with lTheil as the dependent variable performs worse with first differences than with fixed effects estimation in terms of joint significance of all variables, significance of individual variables and the R^2. Thus we will stick to methods resulting from the Hausman test and will deal with some remaining serial correlation with clustered standard errors which should be robust to any form of both serial correlation and heteroscedasticity.

The case of Gini coefficients is more complicated. The $\hat{\rho}$ coefficient resulting from the described procedure equals 0.2731 and 0.2678 for the post-tax and pre-tax Gini, respectively. This means the original undifferentiated errors neither follow random walk, nor are uncorrelated, nor anything in between in the case of AR(1) serial correlation. In turned out the differentiated errors are serially correlated up to the 5. order (meaning 5 lags are significant at the 5% level in explaining the current differentiated error by pooled OLS as can be viewed in Appendix 4). In this case we cannot say much about the serial correlation of the underlying undifferentiated errors except for that they suffer from significant serial correlation of higher order. We have also no tools to model the correlation in such case and first differencing also does not help much. However, the detection of autocorrelation of such a high order is not that accurate with our number of time periods (from 14 to 18 for Gini models depending on the globalization variable). Its detection in this case may potentially mean also some other kinds of problems appear.

The result is that we will also stick to models proposed by the Hausman test, having in mind we cannot interfere here because the needed assumptions are most probably not met (probably the serial correlation assumption). However, resulting coefficients are still unbiased (because we based our control variables on both the theory and an extensive past empirical work on this topic) which turns out to be the key factor in our models as we will need to discuss the sign of resulting coefficients. We interfere in the other models that better satisfies the assumptions (models with lTheil or Ratio 90:10 as the dependent variables). Moreover, available empirical literature is still not quite sure about the sign of the studied effect so a discussion about its sign, without determination of statistical significant in some cases, is still very contributive.

As the very last assessment of models' performance we will do the RESET test for functional misspecification. This test adds powers of fitted values to independent variables and tests for their joint significance. We add second and third power. This accounts for powers of individual variables as well as their interaction terms up to the third order. A good functionally specified test should not be any more explained by additional powers of its fitted values so they should be jointly insignificant. We include the test's results in the last row of the reporting table. Each of our models proved to be functionally good specified. This is another feature that indicates our models are probably unbiased.

6.2 Underlying Variables‘ Effects

6.2.1 Base Model

We now estimate all main models, each with the determined estimation method.

Table 2: Underlying variables’ effects – Base models estimation.

Dependent variable	Post-tax (FE)	Pre-tax (RE)	lTheil (RE)	Ratio 90:10 (FE)
Trade	0.018 (0.029)	0.032 (0.029)	0.009** (0.004)	1.106 (0.88)
FDI	-0.00003 (0.011)	-0.01 (0.011)	-0.004 (0.004)	-0.075 (0.127)
Tourists	-0.071 (0.055)	-0.069 (0.049)	-0.006 (0.006)	-1.171 (0.732)
Migrants	0.047 (0.083)	0.059 (0.066)	0.016** (0.007)	-1.28 (0.828)
Phone	-0.041*** (0.01)	-0.027*** (0.01)	-0.003** (0.001)	-0.231** (0.104)
Internet	0.007 (0.014)	0.004 (0.013)	0.001 (0.002)	0.01 (0.079)
lGDP	23.155** (9.505)	21.444*** (6.204)	0.94 (0.731)	54.287 (47.282)
lGDP_sq	-1.087** (0.515)	-1.143*** (0.359)	-0.065 (0.043)	-4.067 (2.866)
Agric	-0.038 (0.034)	-0.046 (0.033)	0.003 (0.005)	-0.127 (0.382)
Depend	0.182*** (0.047)	0.27*** (0.054)	0.012 (0.009)	0.127 (0.401)
Educ	-2.346 (3.891)	-5.513 (4.212)	-0.235 (0.572)	-36.59 (32.882)
Educ_sq	0.949 (1.725)	4.022* (2.321)	-0.135 (0.283)	13.976 (14.363)
Tech	0.043** (0.019)	0.038* (0.021)	-0.003 (0.004)	0.385 (0.329)
const.	-88.601* (44.914)	-67.015** (27.892)	-7.324** (3.109)	-136.169 (241.745)
Observations	1542	1541	939	1212
Countries	115	115	96	107
R^2 (within)	0.2038	0.2211	0.0597	0.0544
R^2 (between)	0.0650	0.1666	0.4217	0.0739
R^2 (overall)	0.0665	0.2043	0.3734	0.0487
F (χ^2) test of all variables	0.0000	0.0000	0.0000	0.1660
F (χ^2) test of globalization var.	0.0042	0.0875	0.0039	0.3916

F (χ^2) test of period dummies	0.0107	0.0002	0.0015	0.0926
Periods	1995 - 2012	1995 - 2012	1995 - 2008	1995 - 2012
RESET (Prob>F)	0.6096	0.8269	0.9657	0.1693

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. All estimations include a full set of period dummies.

Source: Author

First of all, we can see signs of independent variables are not sensitive. For most variables the sign is the same in every model. Only in 4 cases the signs differ in one of the models (variable Migrants for Ratio 90:10 and variables Aric, Educ_sq and Tech for lTheil) but they are not significant. So our results are insensitive which is very important and indicates that our models are reliable.

Second, all control variables have expected signs. Except for the sectorial employment and education where more workers in the agriculture and higher average years of education actually leads to lower income inequality in our dataset. Moreover, we can see that we found some evidence in favour of existence of the Kuznets curve.

Next, most of the regressors have a little economic effect on the dependent variable. On the other hand, except for the Ratio 90:10, all dependent variables have limited and relatively small range of values ((19, 68) for Post-tax, (17, 71) for Pre-tax and even (-7.2, -0.7) for lTheil as we can see in summary statistics enclosed in Appendix 6). So also small effects count in this cases. Contrary to this, the most economically significant variables are lGDP, lGDP squared, Education, Education squared and Dependency.

Finally, we discuss our globalization variables. We can see two interesting results. First of all, they are almost all individually insignificant. Except for Trade and Migrants that are significant at the 5% level for lTheil, the only variable significant in all cases is Phone. This does not make any economic sense as we cannot think of any channel through which the number of mobile subscriptions per 100 people could affect the income inequality. On the other hand, their joint significance is unambiguous as they are jointly significant for lTheil and insignificant for the Ratio 90:10 (we do not interfere for Gini coefficients models as they do not meet required assumptions). Actually, the performance of the whole model for Ratio 90:10 is suspicious as it has relatively small R^2 and all independent variables are jointly insignificant even at 15% significance level. It turns out it will be better after allowing for different levels of development in the next section. The bottom line here is, however, that effects of composite indices can be actually driven by variables that do not make any economic sense. This is one of our most important results of estimating effects of individual variables.

The second very important result is that pairs of our globalization variables representing economic, social and technological/informational globalization, respectively, each has a variable with positive sign and a variable with negative sign. That means, if they are put together in one globalization index, and even if they are split according to the globalization dimensions, they disturb each other making the joint effect less significant. So we can say it is better to estimate the effect of every single underlying variable individually in order to uncover the structure in the data. This is the second of our most important results.

6.2.2 Different Levels of Development

One of our questions of interest is whether the effect of globalization on the income inequality is different for countries based on their level of development. For this purpose we added interaction terms of our 6 globalization variables with the dummy variable Dependency to every model. They are marked with the sign Variable_D in the following results table. Coefficients on normal variables correspond to developing variables whereas effects for developed countries are sums of normal variables plus the interaction terms. We add this sums in our table for clarity.

Table 3: Underlying variables' effects – Development models estimation.

Dependent variable	Post-tax (FE)	Pre-tax (RE)	lTheil (RE)	Ratio 90:10 (FE)
Trade	-0.089* (0.053)	-0.076 (0.054)	-0.02 (0.019)	0.166 (0.482)
FDI	0.005 (0.019)	-0.009 (0.02)	-0.012** (0.005)	-0.196 (0.35)
Tourists	0.109 (0.083)	0.119 (0.089)	0.024 (0.022)	-0.119 (0.432)
Migrants	0.14 (0.124)	0.075 (0.133)	0.168** (0.072)	-1.215 (1.194)
Phone	-0.019 (0.038)	-0.026 (0.043)	-0.009** (0.004)	-0.272* (0.157)
Internet	-0.044 (0.057)	-0.04 (0.065)	-0.003 (0.018)	0.053 (0.251)

Trade_D	0.122** (0.051)	0.122** (0.053)	0.031* (0.019)	1.164 (0.894)
FDI_D	-0.009 (0.017)	-0.004 (0.018)	0.008** (0.004)	0.076 (0.281)
Tourists_D	-0.218** (0.096)	-0.223** (0.098)	-0.031 (0.023)	-1.374 (0.941)
Migrants_D	-0.113 (0.143)	-0.024 (0.153)	-0.155** (0.072)	0.08 (1.556)
Phone_D	-0.025 (0.037)	-0.002 (0.043)	0.006 (0.004)	0.03 (0.189)
Internet_D	0.054 (0.057)	0.046 (0.064)	0.004 (0.018)	-0.045 (0.282)
Σ Trade	0.033	0.046	0.011	1.33
Σ FDI	-0.004	-0.013	-0.004	-0.12
Σ Tourists	-0.109	-0.104	-0.007	-1.493
Σ Migrants	0.027	0.051	0.013	-1.135
Σ Phone	-0.044	-0.028	-0.003	-0.242
Σ Internet	0.01	0.006	0.001	0.008
lGDP	23.839** (9.705)	22.721*** (5.672)	1.302* (0.682)	38.848 (49.034)
lGDP_sq	-1.143** (0.522)	-1.23*** (0.32)	-0.085** (0.038)	-3.228 (3.501)
Agric	-0.044 (0.032)	-0.049 (0.033)	0.003 (0.005)	-0.179 (0.409)
Depend	0.17*** (0.049)	0.259*** (0.056)	0.012* (0.006)	0.014 (0.463)
Educ	-2.071 (3.794)	-5.688 (4.215)	-0.528 (0.601)	-38.464 (32.17)
Educ_sq	0.814 (1.684)	4.049* (2.283)	-0.021 (0.291)	15.116 (14.986)
Tech	0.041** (0.019)	0.036* (0.02)	-0.004 (0.003)	0.398 (0.32)
const.	-89.056* (45.795)	-70.219*** (26.531)	-8.826*** (3.256)	-56.952 (228.178)
Observations	1542	1541	939	1212
Countries	115	115	96	107
R^2 (within)	0.2267	0.2363	0.1521	0.0586
R^2 (between)	0.0604	0.1743	0.3301	0.0735
R^2 (overall)	0.0590	0.2090	0.2875	0.0497
F (χ^2) test of all var.	0.0000	0.0000	0.0000	0.0507
F (χ^2) test of global. variables (single)	0.2434	0.3195	0.0001	0.6427
F (χ^2) test of interaction terms	0.0241	0.0668	0.1106	0.3961
F (χ^2) test of period dummies	0.0065	0.0002	0.0018	0.1972

Periods	1995 - 2012	1995 – 2012	1995 - 2008	1995 - 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. All estimations include a full set of period dummies.

Source: Author

We can see the interaction terms are jointly insignificant in each model where we can interfere which can signal we should conclude the effects do not differ on different levels of development. However, two things lead us to the conclusion that they actually do differ. First of all, joint significance of single globalization variables (that is, those for developing countries) are statistically very significant for lTheil. They are not for the Ratio 90:10, but this model has many imperfections so we should probably not to focus on it much. At least, its F-test for joint significance of all independent variables is now much more satisfying than in the base model.

The second important remark here is that in many cases (11 out of 24) the effect of each variable has the opposite sign for each group of countries. On the other hand, this could lead to an increase of individually significant variables but it does not. We believe the reason for this is that after separation of countries into two groups, too few are left in each one to precisely enough estimate the coefficients. So though we have uncovered another level of the structure of underlying data the estimation is now not that accurate leading to large standard errors. However, the bottom line is that we conclude that the effect of globalization on the income inequality does differ for different levels of development for majority of underlying variables, though we cannot now simply say that the overall effect is higher or lower for some group. It can be also the case that the effects cancels out and the overall effect is insignificant for both groups.

We examine the overall effects of globalization on the income inequality in the following section.

6.3 Effects of Composite Globalization Indices

One of the main contributions of this thesis wants to be the discussion about alternative normalization of globalization variables that satisfies one more assumption. In this section we compare for each model our proposed G Index's performance with well-established and widely-used ones – KOF and DHL indices.

6.3.1 Base Model

We now estimate the base model but as the measure of globalization we include different composite indices – G Index, KOF Index and DHL Index. This results in 12 different estimations. Complete results are reported in Appendix 7. The only significant index is the G Index, in the case of both post- and pre-tax Gini and the Ratio 90:10 as the dependent variables (we do not trust the inference with Gini coefficients but still it is good to mention it). Other indices are statistically insignificant in all cases. This is probably because the indices consist of variables that have opposite effects, cancelling themselves out. But still all estimations are unbiased so we can have a look whether there is some consistency of results, at least in terms of signs of coefficients.

For a better overview, let us bring the important results together into a more transparent table. The sign in each cell denotes the major sign that appeared among models the most (out of 4 as we have 4 dependent variables and thus 4 variations of a model). The number in brackets is the number of models the sign appeared in.

Table 4: Number of signs in composite globalization indices base models.

# Of Signs	G Index	KOF Index	DHL Index
Index	- (4)	? (2)	? (2)
lGDP	+ (3)	+ (4)	+ (4)
lGDP_sq	- (3)	- (4)	- (4)

Source: Author

We can see the G Index has the best consistency in its results as it reports negative coefficient for all dependent variables. On the other hand, both the KOF and the DHL indices are ambiguous as they report negative coefficients in a half of estimations and positive effect in the second half. This can be attributed either to the different normalization we have used or the different variables it consists of. However, the variables that build up the G Index are the key variables in the other indices so it is probable that stable results of the G Index are driven by the different, year-to-year normalization.

The economic effect of globalization on the income inequality is also small. More precisely, the highest effect (in absolute terms) has the G Index in the Ratio 90:10 model with the effect of -0.528. This means a unit increase of the index results in a decrease of the Ratio 90:10 by a half of unit.

It should be mentioned that the overall effect of a composite index is driven by its underlying variables. We have seen variables that have no economic sense can be the most significant as was the case of Phone. So it is possible this variable is responsible for most of the negative effect we have found in the composite indices. This however has no interpretation as the number of mobile cellular subscriptions per 100 people cannot have any significant effect on the income inequality. The bottom line is that it is better to estimate effects of underlying variables to better see the structure in the data. This avoids conclusions such as there is some negative effect of globalization on the income inequality if actually the effect is driven by strange variables without interpretation.

Finally, we have found evidence of the Kuznets curve in our data. Many of the corresponding coefficients (on lGDP and lGDP_sq) are significant as can be seen the Appendix 7. Altogether they report positive effect of level of the GDP on the income inequality and negative effects of its square which is exactly what the Kuznets curve predicts.

6.3.2 Different Levels of Development

In this section we allow globalization indices to vary between countries depending on their level of development. Complete results can be found in Appendix 8. We also include their sums for clearly show the effect for developed countries.

All indices either normal or in interaction terms are insignificant in all cases (except for DHL in Ratio 90:10 model being positively significant at 10% - but it is based on half the observations of other models so it is not so trustworthy). This is probably because they consist of variables that have opposite effects, making the overall effect small as they cancel out. But it can also be we don’t have enough data. This is particularly true in the case of lTheil model with DHL Index as the covered periods overlap for 4 years 2005 to 2008 only. This can be also the reason why this models perform rather badly in terms of R^2. The effects are statistically insignificant but, nevertheless, all our estimations are unbiased so we can tell if there is some consistency in the results, at least in terms of directions of effects.

Table 5: Number of signs in composite globalization indices development models.

# Of Signs	G Index	KOF Index	DHL Index
Index (developing countries)	- (4)	? (2)	- (3)
Index_D	+ (4)	+ (3)	+ (3)
Σ (developed countries)	- (3)	? (2)	+ (3)
lGDP	+ (3)	+ (4)	+ (4)
lGDP_sq	- (4)	- (4)	- (4)

Source: Author

Firstly, we can see in the table that the effect of globalization on the income inequality in developing countries is most probably negative as 9 out of 12 cases predict. Then the effect is most probably higher (closer to 0 or positive) for developed countries (10 out of 12 cases). But the total effect in developed countries is ambiguous as in half of the cases it is positive and in the other half negative. So we have found some negative effect for developing countries but an unclear effect for development ones with overall indices of globalization. This is actually the opposite result as in the research of Dreher and Gaston (2008) who found robustly positive effect of globalization on the income inequality, mainly in OECD countries, and unclear effect for developing countries.

The economic effect of the globalization indices is small in every cases, as the highest effect (in absolute terms) is -0.579 in the case of G Index in the Ratio 90:10 model. That means a unit increase of the G Index leads to a drop in the Ratio 90:10 by a half of a unit.

Secondly, we assess the performance of globalization indices as comparison of our newly proposed index is one of the major topics in this thesis. We can see the most consistent one is our G Index as it predicts the same sign almost in every cases. The worst performance has the KOF Index because it predicts ambiguous effect for both developed and developing countries. The DHL Index is somewhere in between. Of course this can be driven by underlying variables that differ in al indices but it could also be that the year-to-year normalization yields more stable results. At least, this result means our thoughts were no nonsense and they could be considered in the next research.

Finally, the evidence of Kuznets curve was found in all cases and individual variables are even often significant.

Conclusion

In this thesis we firstly proposed a new, year-to-year method of normalization for a globalization index that yields more stable data in a regression analysis. We created a simple index with this new method based on 6 most used variables. Then we compared its performance with the standard indices – KOF and DHL – and together we tried to found out the effect of globalization on the income inequality, assessing its sensitivity as we used 3 different globalization measures and 4 different measures of within-country income inequality. Moreover, we estimated effects of our 6 variables individually to see how effect of a composite globalization index is build up. We also allowed for different levels of development to see if the effects differ for developed and developing countries. Finally, we included variables that searched for the Kuznets curve.

Our results are that underlying variables that build up globalization indices can have different signs of effects and they can cancel out in the composite index. This was the most probable reason why in most of the cases globalization indices were both statistically and economically insignificant.

The signs of overall globalization indices were consistent – negative in all cases – only in the case of our G Index. KOF and DHL indices had ambiguous effects. This is probably the result of the different normalization method because variables that build up the G Index are key variables in the other ones and should cover lots of their variance. In any case, next research should probably consider both estimations methods with more precise data. At least, the considerations of a year-to-year normalization proved to be consistent with the standard ones (and probably better).

Then, we found the most significant effect in a composite index (and thus the one that drives its overall effect the most) can have a variable that has no economic explanation as was the case with the effect of mobile cellular subscriptions per 100 people on the income inequality. Such results have to be interpreted carefully. So we conclude a better approach in this sense is to estimate effects of underlying variables individually to better uncover the structure of the data and better see the true relationships.

After allowing for different levels of development, the results were still statistically insignificant, possibly because of the lack of data. But because all estimation were unbiased, from the pattern of their signs we concluded the effect of globalization on the income inequality in developing countries is most probably negative. Then this effect in developed countries is higher (closer to zero or positive) but its sign is ambiguous.

We also found that in our post-tax inequality measures individual unobserved effects were important so the Hausman test suggested they should be dealt with. One explanation is that this unobserved effects captures the taxation schemes which are important in the determination of post-tax inequality measures. Next research could possibly add variables that would control for the different taxation schemes to uncover their effect on the post-tax income inequality.

Finally, we found significant evidence of the Kuznets curve in the data and thus confirmed the theory proposed by Simon Kuznets (1955).

To sum up, in this thesis we examined some features of methodology used in the majority of research on this topic that can have significant impact on results but they are not addressed in the publicly available research and we proposed their solutions.

Bibliography

ATIF, Syed Muhammad; SRIVASTAV, Mudit; SAUYTBEKOVA, Moldir and ARACHCHIGE, Udeni Kathri, 2012. *Globalization and Income Inequality: A Panel Data Analysis of 68 Countries.* [online]. [cited 21.4.2016]. Available at: https://www.econstor.eu/handle/10419/65664

BARRO, Robert and LEE, Jong-Wha, 2016. *Educational Attainment Dataset*. [online.] [cited 7.7.2016]. Available at: http://barrolee.com/

BERGGREN, Niclas, 1999. Economic freedom and equality: Friends or foes? *Public Choice.* [online]. Vol. 100. pp. 203-223. [cited 30.4.2016]. DOI: 10.1023/A:1018343912743. Available at: http://link.springer.com/article/10.1023/A:1018343912743

BERGH, Andreas and NILSSON, Therese, 2010. Do liberalization and globalization increase income inequality? *European Journal of Political Economy.* [online]. Vol. 26. pp. 488-505. [cited 21.4.2016]. DOI: 10.1016/j.ejpoleco.2010.03.002. Available at: http://www.sciencedirect.com/science/article/pii/S0176268010000091

DAMGAARD, Christian and WEINER, Jacob, 2000. Describing Inequality in Plant Size or Fecundity. *Ecology.* [online]. Vol. 81, Issue 4. pp. 1139-1142. [cited 8.7.2016]. DOI: 10.1890/0012-9658(2000)081[1139:DIIPSO]2.0.CO;2. Available at: http://www.jacobweiner.dk/site/Publications_files/Damgaard_Weiner_2000.pdf

DREHER, Alex, 2006. Does globalization affect growth? Evidence from a new index of globalization. *Applied Economics.* [online]. Vol. 38, Issue 20. pp. 1091–1110. [cited 30.4.2016]. DOI: 10.1080/00036840500392078. Available at: http://www.tandfonline.com/doi/abs/10.1080/00036840500392078

DREHER, Alex and GASTON, Noel, 2008. Has Globalization Increased Inequality? *Review of International Economics.* [online]. Vol. 16, Issue 3. pp. 516-536. [cited 30.4.2016]. DOI:10.1111/j.1467-9396.2008.00743.x. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9396.2008.00743.x/abstract

GALBRAITH, James K. and KUM, Hyunsub, 2002. *Inequality and Economic Growth: Data Comparisons and Econometric Tests.* [online]. University of Texas Inequality Project Working Paper Number 21. [cited 1.5.2016]. Available at: http://utip.lbj.utexas.edu/papers.html

GHEMAWAT, Pankaj and ALTMAN, Steven A., 2014. *DHL Global Connectedness Index.* [online]. [cited 6.5.2016]. Available at: http://www.dhl.com/en/about_us/logistics_insights/studies_research/global_connectedness_index/global_connectedness_index.html#.V4rQohV97IV

KUZNETS, Simon, 1955. Economic Growth and Income Inequality. *The American Economic Review.* [online]. Vol. 45, Issue 1. [cited 13.7.2016]. Available at: https://assets.aeaweb.org/assets/production/journals/aer/top20/45.1.1-28.pdf

LALL, Subir; JAUMOTTE, Florence; PAPAGEORGIOU, Chris; TOPALOVA, Petia et al., 2007. *Globalization and Inequality.* [online]. Chapter 4. World Economic Outlook. [cited 13.6.2016]. Available at: https://www.imf.org/external/pubs/ft/weo/2007/02/

POTRAFKE, Niklas, 2014. *The Evidence on Globalization.* [online]. CESIFO Working Paper No. 4708. Center for Economic Studies and Ifo Institute. [cited 18.4.2016]. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425513

SOLT, Frederic, 2014. *The Standardized World Income Inequality Database.* [online]. Working paper. SWIID Version 5.0. [cited 21.4.2016]. Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/11992

STURM, Jan-Egbert, 2016. *KOF Index of Globalization.* [online]. ETH, Swiss Federal Institute of Technology Zurich. [cited 12.4.2016]. Available at: http://globalization.kof.ethz.ch/

THE WORLD BANK DATABANK, 2016. *The World Databank.* [online]. [cited 7.5.2016]. Available at: http://databank.worldbank.org/data/home.aspx

UNITED NATIONS, 2015. *Sustainable Development Goals.* [online]. [cited 13.7.2016]. Available at: http://www.un.org/sustainabledevelopment/sustainable-development-goals/

VUJAKOVIC, Petra, 2010. *How to Measure Globalisation? A New Globalisation Index (NGI).* [online]. FIW Working Paper Nr. 46. [cited 13.7.2016]. Available at:

http://www.fiw.ac.at/fileadmin/Documents/Publikationen/Working_Paper/N_046-Vujakovic.pdf

WOOLDRIDGE, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach.* 5th edition. South-Western, Mason, OH: Cengage Learning. 910 p. ISBN 1-111-53104-8.

List of Tables and Figures

List of appendices

Appendices

Appendix 1: Correlations between globalization indices (table).

Year: 2005	G Index	KOF Index	DHL Index
G Index	1.0000		
KOF Index	0.8020	1.0000	
DHL Index	0.8341	0.8514	1.0000

Year: 2010	G Index	KOF Index	DHL Index
G Index	1.0000		
KOF Index	0.7859	1.0000	
DHL Index	0.7816	0.8310	1.0000

Source: Author

Appendix 2: Hausman test's results (table).

Dependent Variable	Test statistic	P-Value of H_0
Post-tax Gini	81.39	0.0000
Pre-tax Gini	16.43	0.9442
lTheil	12.80	0.9789
Ratio 90:10	40.24	0.0800

Source: Author

Appendix 3: Results of testing for serial correlation (table).

Dependent Variable	$\hat{\rho}$ (estimator of SC in differentiated errors)	Standard Error	P-Value of H_0: $\hat{\rho} = 0$	P-Value of H_0: $\hat{\rho} = -0.5$
Post-tax Gini	0.2731	0.0252	0.000	-
Pre-tax Gini	0.2678	0.2529	0.000	-
lTheil	−0.2759	0.0350	0.000	0.000
Ratio 90:10	−0.7360	0.0341	0.000	0.000

Source: Author

Appendix 4: Serial correlation of differenced errors for the Gini post-tax model (table).

Dependent variable: Residuals of the differenced post-tax Gini coefficient model after pooled OLS.

Independent var.	Coefficient (St. Err.)
res_lag	0.379*** (0.032)
res_lag2	0.299*** (0.031)
res_lag3	-0.271*** (0.029)
res_lag4	-0.049* (0.029)
res_lag5	0.11*** (0.027)
res_lag6	-0.008 (0.025)
const.	-0.0002 (0.021)
Observations	813
R^2	0.3172
Adjusted R^2	0.3121
F test of all variables	0.0000

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Pooled OLS estimation.

Source: Author

Appendix 5: First differences estimation of lTheil model (table).

Dependent variable: dlTheil

Independent var.	Coefficient (St. Err)
dTrade	0.002 (0.003)
dFDI	-0.006* (0.003)
dTourists	0.005 (0.005)
dMigrants	-0.006 (0.016)
dPhone	0.0001 (0.001)
dInternet	-0.0001 (0.002)
dlGDP	-2.536 (1.64)
dlGDP_sq	0.117 (0.097)
dAgric	0.0005 (0.003)
dDepend	-0.014 (0.013)
dEduc	1.035 (1.018)
dEduc_sq	-0.583 (0.557)
dTech	-0.002 (0.002)
_cons	0.174* (0.096)
Observations	843
Countries	92
R^2	0.0417
R^2 (adjusted)	0.0124
F (χ^2) test of all var.	0.0830
F (χ^2) test of glob. var.	0.6320
F (χ^2) test of period dummies	0.0820
Periods	1996 - 2008

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Pooled OLS estimation.

Source: Author

Appendix 6: Summary statistics (table).

Variable		Mean	Std. Dev.	Min	Max	Observations	
Id	overall	107.5	61.784	1	214	N =	3852
	betw.		61.921	1	214	n =	214
	within		0	107.5	107.5	T =	18
Year	overall	2003.5	5.189	1995	2012	N =	3852
	betw.		0	2003.5	2003.5	n =	214
	within		5.189	1995	2012	T =	18
Post-tax	overall	38.692	8.833	19.233	68.155	N =	2374
	betw.		8.639	23.049	65.063	n =	167
	within		2.402	27.063	57.917	Tbar =	14.216
Pre-tax	overall	45.336	7.294	17.989	71.13	N =	2373
	betw.		7.105	27.794	68.671	n =	167
	within		2.684	31.971	65.776	Tbar =	14.21
lTheil	overall	-3.222	0.869	-7.27	-0.706	N =	1252
	betw.		0.862	-5.655	-0.942	n =	121
	within		0.307	-5.545	-1.547	T =	10.347
Ratio 90:10	overall	21.073	51.281	2.824	1846.333	N =	1905
	betw.		25.042	5.828	202.929	n =	155
	within		45.037	-171.098	1664.477	Tbar =	12.29
G Index	overall	17.325	11.288	1.507	66.529	N =	2659
	betw.		10.872	3.766	62.169	n =	172
	within		4.057	0.506	32.125	Tbar =	15.459
KOF	overall	52.311	17.678	15.427	92.629	N =	3446
	betw.		17.087	22.779	91.389	n =	192
	within		4.651	28.87	67.582	Tbar =	17.948
DHL	overall	43.828	17.688	3	90	N =	1112
	betw.		17.52	6.5	88.5	n =	139
	within		2.8	26.203	55.453	T =	8
Trade	overall	19.566	13.453	0	100	N =	3328
	betw.		12.145	0.114	90.616	n =	193
	within		5.66	-19.589	59.446	T-bar =	17.244
FDI	overall	9.19	12.314	0	100	N =	2874
	betw.		8.964	4.492	99.041	n =	177
	within		9.795	-27.314	90.047	Tbar =	16.237
Tourists	overall	12.016	14.871	0	100	N =	3102
	betw.		14.317	0.204	95.524	n =	189
	within		5.013	-16.625	44.715	Tbar =	16.413
Migrants	overall	13.314	19.396	0	100	N =	3761
	betw.		19.477	0	99.545	n =	212
	within		2.14	-10.696	27.199	Tbar =	17.741
Phone	overall	25.865	24.886	0	100	N =	3659
	betw.		19.768	0.091	84.805	n =	208
	within		15.355	-23.204	86.116	Tbar =	17.591
Internet	overall	22.072	27.232	0	100	N =	3648
	betw.		22.85	0	92.438	n =	205
	within		14.773	-41.372	81.422	Tbar =	17.795

lGDP	overall	8.132	1.641	4.242	11.974	N =	3471
	betw.		1.648	5.006	11.734	n =	199
	within		0.189	5.819	8.989	Tbar =	17.442
lGDP_sq	overall	68.819	27.162	17.999	143.38	N =	3471
	betw.		27.363	25.061	137.702	n =	199
	within		2.992	32.512	84.166	Tbar =	17.442
Agric	overall	21.654	22.036	0.1	92.2	N =	2550
	betw.		23.54	0.194	92.2	n =	181
	within		5.147	-9.761	70.197	Tbar =	14.088
Depend	overall	64.494	19.433	16.329	114.304	N =	3495
	betw.		18.692	27.1	107.59	n =	195
	within		5.403	39.937	94.378	Tbar =	17.923
Educ	overall	0.409	0.35	-0.048	1.836	N =	2592
	betw.		0.337	0.013	1.616	n =	144
	within		0.098	-0.086	1.234	T =	18
Educ_sq	overall	0.29	0.466	0	3.371	N =	2592
	betw.		0.434	0.0002	2.625	n =	144
	within		0.173	-0.82	2.155	T =	18
Tech	overall	9.283	12.413	0	87.404	N =	2869
	betw.		11.804	0	68.71	n =	182
	within		5.53	-23.631	79.4	Tbar =	15.764
Develop	overall	0.612	0.487	0	1	N =	3852
	betw.		0.488	0	1	n =	214
	within		0	0.612	0.612	T =	18

Source: Author

Appendix 7: Indices models – Base models (table).

Dependent: Post-tax	G Index	KOF	DHL
Index	-0.128*** (0.043)	0.017 (0.047)	0.009 (0.044)
lGDP	14.092 (8.926)	14.19** (6.937)	25.266*** (9.504)
lGDP_sq	-0.572 (0.471)	-0.645* (0.379)	-1.396** (0.541)
Agric	-0.045 (0.038)	-0.048 (0.04)	0.05 (0.065)
Depend	0.22*** (0.047)	0.236*** (0.042)	0.323*** (0.083)
Educ	-0.094 (3.634)	0.114 (3.614)	4.017 (4.658)
Educ_sq	0.223 (1.572)	0.625 (1.667)	-1.164 (2.244)
Tech	0.044** (0.021)	0.041** (0.018)	-0.002 (0.025)
const.	-51.801 (43.785)	-49.775 (32.551)	-94.483** (43.827)
Observations	1542	1648	683
Countries	115	119	106
R^2 (within)	0.1699	0.1641	0.1728
R^2 (between)	0.0548	0.0677	0.1600
R^2 (overall)	0.0585	0.0741	0.1793
F (χ^2) test of all var.	0.0000	0.0000	0.0003
F (χ^2) test of period dummies	0.0077	0.1742	0.1572
Periods	1995 - 2012	1995 - 2012	2005 - 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Fixed effects estimation. All estimations include a full set of period dummies.

Source: Author

Dependent: Pre-tax	G Index	KOF	DHL
Index	-0.084* (0.043)	0.066 (0.047)	-0.011 (0.042)
lGDP	15.794** (6.977)	15.124*** (5.8)	22.841*** (5.137)
lGDP_sq	-0.798** (0.4)	-0.822** (0.332)	-1.208*** (0.296)
Agric	-0.051 (0.035)	-0.05 (0.035)	0.023 (0.057)
Depend	0.296*** (0.052)	0.311*** (0.049)	0.318*** (0.056)
Educ	-3.174 (3.907)	-2.517 (3.713)	-0.761 (3.769)
Educ_sq	3.261 (2.264)	3.445* (2.089)	2.23 (1.69)
Tech	0.039* (0.023)	0.036** (0.018)	-0.03* (0.017)
const.	-45.962 (31.393)	-44.224* (25.746)	-76.419*** (23.827)
Observations	1541	1647	682
Countries	115	119	106
R^2 (within)	0.2024	0.2018	0.1986
R^2 (between)	0.1654	0.1550	0.0972
R^2 (overall)	0.1992	0.1940	0.1535
F (χ^2) test of all var.	0.0000	0.0000	0.0000
F (χ^2) test of period dummies	0.0006	0.0066	0.0354
Periods	1995 - 2012	1995 - 2012	2005 - 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Random effects estimation. All estimations include a full set of period dummies.

Source: Author

Dependent: lTheil	G Index	KOF	DHL
Index	-0.001 (0.005)	-0.008 (0.008)	-0.009 (0.01)
lGDP	-0.177 (0.998)	0.444 (0.793)	0.334 (1.518)
lGDP_sq	0.001 (0.06)	-0.033 (0.049)	-0.025 (0.086)
Agric	0.002 (0.005)	0.001 (0.005)	0.002 (0.007)
Depend	0.011 (0.009)	0.007 (0.008)	0.012 (0.016)
Educ	0.172 (0.597)	-0.007 (0.617)	-0.111 (0.745)
Educ_sq	-0.357 (0.292)	-0.26 (0.302)	0.073 (0.433)
Tech	-0.002 (0.004)	-0.002 (0.004)	-0.001 (0.005)
const.	-2.48 (4.146)	-4.49 (3.441)	-4.331 (7.676)
Observations	936	975	203
Countries	96	97	68
R^2 (within)	0.0456	0.0330	0.0444
R^2 (between)	0.2686	0.3423	0.2776
R^2 (overall)	0.2701	0.3269	0.2920
F (χ^2) test of all var.	0.0000	0.0000	0.0000
F (χ^2) test of period dummies	0.0064	0.0095	0.7054
Periods	1995 - 2008	1995 - 2008	2005 - 2008

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Random effects estimation. All estimations include a full set of period dummies.

Source: Author

Dependent: Ratio 90:10	G Index	KOF	DHL
Index	-0.528* (0.273)	-0.043 (0.714)	0.413 (0.276)
lGDP	9.049 (32.914)	34.536 (28.915)	47.563* (25.034)
lGDP_sq	-1.266 (1.792)	-3.153* (1.677)	-3.877** (1.938)
Agric	-0.226 (0.383)	-0.236 (0.356)	0.024 (0.156)
Depend	0.355 (0.282)	0.389 (0.237)	0.774** (0.299)
Educ	-16.157 (26.404)	-14.742 (29.433)	2.266 (13.298)
Educ_sq	7.439 (10.72)	8.54 (12.601)	-1.702 (4.721)
Tech	0.418 (0.316)	0.327 (0.232)	0.081 (0.077)
const.	25.695 (214.227)	-60.332 (189.666)	-154.603 (99.865)
Observations	1212	1306	630
Countries	107	111	97
R^2 (within)	0.0313	0.0277	0.0199
R^2 (between)	0.0545	0.0495	0.0073
R^2 (overall)	0.0414	0.0302	0.0211
F (χ^2) test of all var.	0.1326	0.0971	0.0027
F (χ^2) test of period dummies	0.0986	0.1615	0.5330
Periods	1995 - 2012	1995 - 2012	2005 – 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Fixed effects estimation. All estimations include a full set of period dummies.

Source: Author

Appendix 8: Indices models – Different levels of development (table).

Dependent: Post-tax	GIndex	KOF	DHL
Index	-0.143 (0.118)	-0.03 (0.087)	-0.026 (0.084)
Index_D	0.017 (0.117)	0.057 (0.088)	0.054 (0.091)
Σ (developed countries)	-0.126	0.027	0.028
lGDP	14.487* (8.598)	16.98** (7.226)	26.993*** (9.318)
lGDP_sq	-0.597 (0.444)	-0.823** (0.402)	-1.512*** (0.526)
Agric	-0.045 (0.038)	-0.044 (0.038)	0.05 (0.066)
Depend	0.219*** (0.048)	0.233*** (0.042)	0.32*** (0.084)
Educ	-0.159 (3.751)	-0.181 (3.724)	4.052 (4.654)
Educ_sq	0.255 (1.586)	0.771 (1.686)	-1.266 (2.224)
Tech	0.044** (0.021)	0.04** (0.018)	-0.002 (0.025)
const.	-53.242 (42.756)	-59.974* (32.839)	-100.697** (43.148)
Observations	1542	1648	683
Countries	115	119	106
R^2 (within)	0.1700	0.1662	0.1747
R^2 (between)	0.0536	0.0615	0.1297
R^2 (overall)	0.0567	0.0660	0.1390
F (χ^2) test of all var.	0.0000	0.0000	0.0001
F (χ^2) test of glob. indices	0.0147	0.7441	0.7634
F (χ^2) test of period dummies	0.0093	0.1681	0.1708
Periods	1995 - 2012	1995 - 2012	2005 - 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Fixed effects estimation. All estimations include a full set of period dummies.

Source: Author

Dependent: Pre-tax	GIndex	KOF	DHL
Index	-0.155 (0.12)	0.006 (0.06)	-0.052 (0.063)
Index_D	0.079 (0.121)	0.08 (0.056)	0.066 (0.062)
Σ (developed countries)	-0.076	0.086	0.014
lGDP	16.892*** (6.02)	17.009*** (5.097)	23.044*** (5.067)
lGDP_sq	-0.873*** (0.337)	-0.98*** (0.28)	-1.253*** (0.287)
Agric	-0.049 (0.036)	-0.045 (0.036)	0.028 (0.057)
Depend	0.289*** (0.054)	0.308*** (0.048)	0.316*** (0.055)
Educ	-3.555 (3.964)	-3.245 (3.785)	-0.716 (3.771)
Educ_sq	3.42 (2.22)	3.712* (2.043)	2.067 (1.688)
Tech	0.038 (0.023)	0.035** (0.017)	-0.031* (0.017)
const.	-49.251* (28.301)	-48.513** (23.994)	-75.239*** (23.909)
Observations	1541	1647	682
Countries	115	119	106
R^2 (within)	0.2040	0.2078	0.2019
R^2 (between)	0.1702	0.1772	0.1068
R^2 (overall)	0.2086	0.2131	0.1627
F (χ^2) test of all var.	0.0000	0.0000	0.0000
F (χ^2) test of glob. indices	0.1284	0.1542	0.5702
F (χ^2) test of period dummies	0.0008	0.0050	0.0851
Periods	1995 - 2012	1995 - 2012	2005 - 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Random effects estimation. All estimations include a full set of period dummies.

Source: Author

Dependent: lTheil	GIndex	KOF	DHL
Index	-0.029 (0.029)	-0.012 (0.015)	-0.003 (0.012)
Index_D	0.029 (0.029)	0.005 (0.012)	-0.008 (0.01)
Σ (developed countries)	0.0005	-0.007	-0.011
lGDP	-0.023 (0.897)	0.394 (0.851)	0.784 (1.307)
lGDP_sq	-0.01 (0.053)	-0.034 (0.048)	-0.046 (0.076)
Agric	0.004 (0.005)	0.001 (0.005)	0.002 (0.007)
Depend	0.01 (0.009)	0.007 (0.008)	0.012 (0.016)
Educ	0.117 (0.608)	-0.031 (0.622)	-0.119 (0.763)
Educ_sq	-0.317 (0.302)	-0.248 (0.305)	0.066 (0.441)
Tech	-0.002 (0.004)	-0.002 (0.004)	-0.002 (0.005)
const.	-2.947 (3.88)	-4.07 (3.882)	-6.572 (6.556)
Observations	939	975	203
Countries	96	97	68
R^2 (within)	0.0553	0.0388	0.0431
R^2 (between)	0.2739	0.3231	0.2916
R^2 (overall)	0.2717	0.3072	0.2935
F (χ^2) test of all var.	0.0000	0.0000	0.0000
F (χ^2) test of glob. indices	0.6008	0.5864	0.4585
F (χ^2) test of period dummies	0.0068	0.0188	0.6105
Periods	1995 - 2008	1995 - 2008	2005 - 2008

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Random effects estimation. All estimations include a full set of period dummies.

Source: Author

Dependent: Ratio 90:10	GIndex	KOF	DHL
Index	-0.579 (0.678)	0.325 (1.036)	0.312* (0.159)
Index_D	0.065 (0.703)	-0.503 (0.608)	0.152 (0.287)
Σ (developed countries)	-0.514	-0.178	0.464
lGDP	11.382 (46.719)	3.741 (46.075)	53.848* (32.157)
lGDP_sq	-1.416 (2.539)	-1.184 (2.881)	-4.287* (2.516)
Agric	-0.223 (0.402)	-0.27 (0.349)	0.029 (0.156)
Depend	0.351 (0.287)	0.388 (0.245)	0.762** (0.294)
Educ	-16.573 (25.434)	-11.29 (26.403)	1.918 (13.446)
Educ_sq	7.641 (10.443)	7.312 (11.624)	-1.755 (4.743)
Tech	0.417 (0.309)	0.342 (0.239)	0.082 (0.078)
const.	17.21 (260.392)	56.674 (222.503)	-177.738 (112.711)
Observations	1212	1306	630
Countries	107	111	97
R^2 (within)	0.0313	0.0288	0.0200
R^2 (between)	0.0561	0.0216	0.0090
R^2 (overall)	0.0422	0.0142	0.0236
F (χ^2) test of all var.	0.1606	0.1114	0.0015
F (χ^2) test of glob. indices	0.1469	0.5750	0.1510
F (χ^2) test of period dummies	0.1209	0.1564	0.5278
Periods	1995 - 2012	1995 - 2012	2005 - 2012

*** Denotes significant at 1% level, ** significant at 5% level, * significant at 10% level. Robust standard errors in brackets. Fixed effects estimation. All estimations include a full set of period dummies.

Source: Author

Appendix 9: KOF variables and weights (table).

Indices and Variables		Weights
A.	Economic Globalization	[36%]
	i) Actual Flows	(50%)
	Trade (percent of GDP)	(22%)
	Foreign Direct Investment, stocks (percent of GDP)	(27%)
	Portfolio Investment (percent of GDP)	(24%)
	Income Payments to Foreign Nationals (percent of GDP)	(27%)
	ii) Restrictions	(50%)
	Hidden Import Barriers	(23%)
	Mean Tariff Rate	(28%)
	Taxes on International Trade (percent of current revenue)	(26%)
	Capital Account Restrictions	(23%)
B.	Social Globalization	[37%]
	i) Data on Personal Contact	(33%)
	Telephone Traffic	(26%)
	Transfers (percent of GDP)	(2%)
	International Tourism	(26%)
	Foreign Population (percent of total population)	(21%)
	International letters (per capita)	(25%)
	ii) Data on Information Flows	(35%)
	Internet Users (per 1000 people)	(36%)
	Television (per 1000 people)	(38%)
	Trade in Newspapers (percent of GDP)	(26%)
	iii) Data on Cultural Proximity	(32%)
	Number of McDonald's Restaurants (per capita)	(46%)
	Number of Ikea (per capita)	(46%)
	Trade in books (percent of GDP	(7%)
C.	Political Globalization	[27%]
	Embassies in Country	(25%)
	Membership in International Organizations	(27%)
	Participation in U.N. Security Council Missions	(22%)
	International Treaties	(26%)

Source: STURM (2016)

Appendix 10: DHL variables and weights (table).

Pillar (Weight % of total)	Depth Component (Weight % of Pillar)	Breadth Component (Weight % of Pillar)
1. Trade (35%)	1.1 Merchandise Trade (75%)	1.1 Merchandise Trade (100%)
	1.2 Services Trade (25%)	-
2. Capital (35%)	2.1 FDI Stocks (25%)	2.1 FDI Stocks (25%)
	2.2 FDI Flows (25%)	2.2 FDI Flows (25%)
	2.3 Portfolio Equity Stocks (25%)	2.3 Portfolio Equity Stocks (50%)
	2.4 Portfolio Equity Flows (25%)	-
3. Information (15%)	3.1 International Internet Bandwidth (40%)	-
	3.2 Telephone Call Minutes (40%)	3.2 Telephone Call Minutes (67%)
	3.3 Trade in Printed Publications (20%)	3.3 Trade in Printed Publications (33%)
4. People (15%)	4.1 Migrants (33%)	4.1 Migrants (33%)
	4.2 Tourisms (33%)	4.2 Tourisms (33%)
	4.3 Students (33%)	4.3 Students (33%)

Source: Ghemawat and Altman (2014)

Appendix 11: Number of original and interpolated values in the dataset (table).

Variable	Original	Interpolated	Difference	%
Post-tax	2337	2374	37	0.96
Pre-tax	2336	2373	37	0.96
Theil	1110	1252	142	4.74
Ratio 90:10	1015	1905	890	23.10
Trade	3296	3328	32	0.83
FDI	2785	2874	89	2.31
Tourists	3102	3151	49	1.27
Migrants	837	3761	2924	75.91
Phone	3617	3659	42	1.09
Internet	3449	3648	199	5.17
GDP	3471	3471	0	0.00
Agric	1869	2550	681	17.68
Depend	3495	3495	0	0.00
Educ	576	2592	2016	52.34
Tech	2662	2869	207	5.37
Develop	3852	3852	0	0.00
In total interpolated including Migrants and Educ (%)				11.92
In total interpolated excluding Migrants and Educ (%)				3.90

Note: Variables Migrants and Educ have more interpolated values because they are only available with 5-year frequency. Number of observation for the full dataset 1995-2012 is 214×18=3852. Number of observations for Theil is 214×14=2996.

Source: Author