DOI: 10.20472/IAC.2019.046.014

ANA CECILIA PARADA ROJAS
Instituto Politécnico Nacional, Mexico

HUMBERTO RÍOS BOLÍVAR
Instituto Politécnico Nacional, Mexico

JORGE OMAR RAZO DE ANDA
Instituto Politécnico Nacional, Mexico

MINING OF CLASSIFICATION TREES TO ANALYZE A MULTIDIMENSIONAL PHENOMENON

Abstract:

During periods of remarkable trade openness, increase income inequality in many countries. This paper analyzes how factors that influence inequality due to commercial globalization interact each other. For which a reliable Classifier Tree -selected through a modeling process of bootstrapping- is built, it has 14 knowledge rules and classifies 84% of the observations correctly. This model indicates that inequality‘s changes into a country, due greater economic integration, depend principally on the labor market’ structure –in agricultural countries and urbanization processes (industrialization) it reduces depending in turn on the rule of law; on the other hand, in countries with a strong service sector and good trade terms it increases in periods of stagnation or with low levels of high technology exports.

Keywords:

Income Inequality, Globalization, International Trade, Data Mining, Classification and Regression Tree (CART)

JEL Classification: C44, D33, F00

Introduction

Increases in house income inequality within countries has been observed since he decade of the eighties and considering that this phenomenon is accompanied by a series of social, cultural and economic problems (IMF, 2007, Atkinson, Piketty, and Saez, 2011), the theme has being gain importance in the research. Although the dynamics of inequality is a complex process that arises from different social phenomena and occurs through different mechanisms, the recent increases are attributed to processes of globalization (Harrison & Hanson, 1999) that were potentiated by the change technology (Bourguignon, 2017), mainly in communications (Dollar, 2004).

The most common argument that relates the processes of economic integration with wage changes, indicate that free trade between countries, especially when it comes to rich or developed countries with poor or developing countries (Dollar, 2004), leads to increases in the wage gap between unskilled and highly skilled workers. Based on the theorem of Stolper & Samuelson (1941), inequality decreases due to the increase in the wage of unskilled work in countries with abundant labor, considered a comparative advantage that balances the wages. However, empirically Han et al.(2012) find no evidence of this, and Goldberg & Pavcnik (2007) argue that regardless of the type of economy inequality increases between groups of workers and according to Helpman et al. (2010) also individually. That without neglecting the combined effect of technological progress that affects the supply of work (Tinbergen, 1970) and the type of production. By another hand the efficiency of government institutions can explain and even reverse the adverse effects of trade liberalization, through a strong welfare state, measured by governance indices (Atkinson et al., 2001; Kaufmann et al., 2009). This suggests that the effect of trade globalization has different effects on income inequality depending on the period of study, the characteristics of the data and the circumstances of each country, specifically the structure of the labor market, the education of the labor force, the kind of exports, in addition to the institutions and social policies. Therefore, the objective of this paper is to analyze how the different factors that influence the effect on inequality interact when commercial globalization increases.

In order to capture only the effects derived from a commercial interaction raise, an index of commercial globalization is constructed using the Mahalonobis distance, based on which the records are filtered by this index increases. Then the Classification trees are built, due this data mining technique is flexible and non-parametric, which allows analyzing the relationships between the variables and how they are combined so that a situation is presented, in this case the income distribution changes**.** In addition this tool allows to understand the phenomenon of study, since it provides knowledge through the analysis and the extraction of knowledge rules (Faraway, 2016). However, the decision criteria may change based on the parameters and input data set. So, a bootstrapping process of random sampling of 75% of the data is carried out to build a little more than one hundred

trees, among which the most reliable one is selected considering the efficiency and stability**.**

In the results analysis basing on the knowledge rules provided by the best model, the structure of the labor market is identified as a determining characteristic. Discriminating agricultural countries but with potential for urbanization as the beneficiaries of trade, while countries with the highest employment in the services sector increase their inequality in periods of stagnation or if they do not export high technology.

In the next section, a literature review of the determinants of inequality and mechanisms associated to increases in trade globalization is made, in which the debate about wage effects, the technological progress rol and the welfare state is described. Next the factors included in the analysis -which differentiate the countries in certain periods of time- are described, and an index of commercial globalization is proposed that allows filtering situations with increasing degree of commercial openness. Next, we describe the Classification Tree technique -including the training algorithm and its evaluation- and the modeling process to select the most reliable and stable tree. Later, each one of the 14 knowledge rules of the best model is analyzed, and finally the general paper conclusions are presented.

Literature Review

Inequality Owed Trade Globalization.

Income inequality changes within a country are mainly attributed to urbanization phenomena (Kuznets, 1955), economic growth (Dollar & Kraay, 2001), technological progress (Lawrence et al., 1993) and recently to globalization (Aghion, 1999, Atkinson, 2015, Jaumotte, 2013, Bourguignon, 2017).

To measure the inequality between households within a country, the Gini coefficient is taken after taxes and transfers, due it is an aggregate index of the income distribution besides being more sensitive to changes in the extremes (upper and lower class) than in the media (middle class).

In order to analyze the inequality changes over the time and between countries, the values of the Standardized World Income Inequality Database (SWIID 5.0) developed by Frederick Solt, are used. This database contains the standardized values of the Gini coefficients estimated by the main international institutions, through multiple imputation algorithms of lost data (Solt, 2016); besides being a resource used in several research about this subject (Palma & Stiglitz, 2016; Heathcote et. al., 2017; Jeon & Kabukcuoglu, 2018).

Determinants of inequality

In 1950 Kuznets identifies a relationship between progress and inequality, through two forces that increase income distribution inequality due to the countries development: the

concentration of the richest savings and the urbanization, the latter considers the income gap between the rural and industrial population, on the other hand the concentration of savings is affected mainly by fiscal policies, demographic phenomena, entrepreneurship of new industries and changes in the portion of income from services.

Bluestone & Harrison (1982) study the low productivity and structural change of employment in the services sector, industrial and agricultural. When the cities are industrialized the income of the people who worked in this sector is higher than in the agricultural sector, so the inequality increases, but as it is given the mobility decreases. Then the structural change in income was characterized by economic growth due to the process of industrialization but followed by a development process that combined mechanisms of population reduction, fiscal and social policies, consistent with the inverted U of the Kuznets hypothesis (1955). The conventional theory of social welfare maintains that increasing the per capita product of a country will improve the welfare of the entire population, including that of the poorest (Deininger & Squire, 1996) "International trade is good for growth and growth is good for poverty "(Dollar and Kraay, 2001).

Empirically the Kuznets hypothesis is satisfied in 1970 for most of the OECD countries, but not for developing or underdeveloped countries, since 1980 the Kuznets curve is no longer so clear, so the need arises of new theories to understand this relationship (Aghion, 1999), context in which it is argued that globalization (mainly commercial opening) in combination with technological progress are the forces responsible for contemporary inequalities (Bourguignon, 2017). The report of the International Monetary Fund (2007), Atkinson, Piketty, and Saez (2011), among others, indicate that in the last decades, income inequality increased in most countries; Harrison & Hanson, (1999) attribute this phenomenon to the processes of globalization that deepened in the eighties.

Trade Globalization

Globalization is an economic integration process between the economies around the world that reconfigure the structure and interaction of markets at an international level. Economically we can separate globalization in two stages, the first from 1870 to 1914 focused mainly on financial integration and mobility, and the second considering contemporary globalization from 1914 (Milanovic, 2016). Despite the fact that during World War II there was a setback in globalization, with the end of the war the course of economic integration was re-established, under the umbrella of the General Agreements on Tariffs and Trade (GATT) which encourages free trade. However, Dollar (2004) identifies an Ito detonated by the commercial opening of China in 1978, which contributed to the development of the crises caused by foreign debt in Latin American countries, which in turn led these economies to change their import strategies substitution (to strengthen the domestic industry) to an externally oriented strategy, that is, to the increase of exports to developed economies. The entry of China into the world market had such an impact, that the terms of trade of many economies increased drastically (Bourguignon, 2017).

So this work focuses on contemporary globalization from 1980, period in which this phenomenon was potentiated by technological advances and communication, coupled with the signing of free trade agreements between developing and developed countries (Dollar, 2004).

Effect through wages.

Wage inequality is different from income inequality, they can even move in different directions. By definition, the wage doesn't include income from capital returns; although there is no simple direct relationship, part of the income inequality comes from this wage gap, in addition to the fact that the proportion of income received as a salary corresponds mostly to people who do not own capital.

The classic argument to justify the trade liberalization and the free exchange of goods, falls on the comparative advantage of David Ricardo associated with the optimal reallocation of productive factors that increases economic welfare. However, information asymmetries and incomplete markets distort this efficiency, it also depends on the initial endowments and does not have to do with equity (Stiglitz, 2010).

The debate about the effect of trade globalization, understood as the opening of markets and economic interaction through trade relations for the exchange of goods and services, on the economy and welfare, is not new; the belief that the protection of the internal market is necessary in the face of imminent competition from international markets, was already a common argument when Stolper & Samuelson presented their famous theorem in 1941, ensuring that it was possible to "unequivocally infer" the effects of trade on the remuneration of work. This theorem argues that trade liberalization decreases the inequality between the wages of workers in a developing country endowed with unskilled labor force, due to a greater demand for this type of labor, however, the opposite happens in developed countries - where the demand for unskilled workers decreases and inequality increases (Stolper & Samuelson, 1941; FitzGerald, 1996). Hence, the prices of the factors of production are balanced in the countries that trade with each other, because production increases in countries with abundant labor. This process makes less productive companies leave the market, and encourages companies to select the most capable workers, increasing the wage inequality between skilled and unskilled labor known as salary premium (Helpman et al., 2010), although in the long term the working class wins due to the increases in its productivity (Shahbaz, 2012).

Empirically Han et al, (2012) use the Heckscher-Ohlin model including the salary advantage of workers with higher levels of education, and indicate that the theorem is not fulfilled. Goldberg & Pavcnik (2007) find that after increases in trade globalization wage inequality between groups of workers by economic sector increases both in developing and developed countries and within groups (Helpman et al., 2010).

Role of technological progress.

From technological advances in 1980, not only did economic integration increase, but it affected the way in which markets are distributed and the type of exports from developing countries to developed economies; "CD players from China, refrigerators from Mexico and software from Thailand" (Dollar, 2004: 150p). According to Asteriou et al. (2014) the inequality decreased in European countries due to the fact that they export high technology (machinery for industrial processes produced with a high degree of research and development).

Lawrence et al. (1993) argues that inequality comes mostly from technological change. The rise of technological markets affects the demand for work, since companies dedicated to innovations require employees with a higher degree of education. Technological changes are born in developed countries and are usually focused on saving labor and replacing capital with unskilled labor. What generates inequalities in wages in both developed and developed countries (Agénor, 2002), in addition to generating a technological dependence. Jaumotte (2013) finds that as many technological advances as financial integration benefit 20% of the richest population in a country.

Institutional approach

The effect of globalization on income inequality also varies depending on the institutions and public and fiscal policies, that is, the welfare state of each country, since it can avoid increases in inequality, thus allowing sustained growth and successful economic integration. (Atkinson, 2003, 2015).

Akerman et al. (2013) show that countries without strong institutions that protect the labor market has a pronounced inequality wage, in addition to important variations between sectors and occupations. According to Kaufmann et al. (2009) policies often have adverse effects due to the ineffectiveness of the state to implement them. Atkinson (2001) emphasizes the role of governance indices, protectionist measures, as a key factor in the determination of inequality. Countries with high levels of corruption can hardly redistribute resources, also face higher export costs, making it difficult to obtain benefits from globalization (Dollar, 2004), however, North (1990) argues that globalization can also induce changes in the institutional environment, although these can be slow.

Why the theories and results diverge

Despite being a widely studied topic, there is no concession about the effect of globalization on inequality. The contributions of empirical research depend on a part of the countries included in the study, the period, the measurement of inequality both in form and rigor (Ravallion, 2003), also depend on the theoretical perspective influenced by the type of inequality analyzed (Nissanke & Thorbecke, 2006), that is, the approach, for example, from the point of view of developed or industrialized countries, in which inequality increased after

signing free trade agreements (Alderson, 2001) combined to technological progress, and from the point of view of developing countries or emerging economies, which during the 1990s reduced their inequality by benefiting from trade Dollar (2004) through wages.

In general, the developing and developed countries perceived different effects due to the structure of employment, education or training of labor and type of exports, but also influence, their institutions and social policies, since they allow a country to benefit or not from the processes of economic integration.

Data and Empirical Methodology

Data Specification

Due to the complexity of the interaction between commercial globalization and income inequality between households within a country, there is no general theory or rule that unequivocally describes the relationship between these phenomena, on the contrary the effect on inequality before Trade openness increases depend on the circumstances of each country, as described in the previous section. For this reason, a multidimensional analysis is carried out that includes factors of population, employment, spending and investment, in addition to the governance indexes, commercial, financial and theological variables, which can be seen in table 1, structured in a data panel by country and year.

Table 1. Definition and sources variables.

Cod.	Definition	Cod.	Definition
Population and Employment*		**Trade ***	
GDP_R	Growth rate		
POB_URB	Population Urban (Growth rate)	MT	Merchandise trade (% GDP)
TFERA	Fertility rate in adolescents	TERM_TRADE	Exchange terms index
ESP_VIDA	Life expectancy at birth (years)	IM_GS_P	Import of goods and services (% GDP)
EM_IND	Emp. in the industrial sec. (% total)	EXP_GS_P	Export of goods and services (% GDP)
EM_SERV	Emp. in the services sec. (% total)	HTE_T	Export of high technology (% GDP)
EM_AGR	Emp. in the agricultural sec. (% total)	HTE_PM	Export of high technology (% Manufac.)
Expenditure and Investment * **		**Financial**	
IPRIV	Private investment (% GDP)	DEB_GDP	Historical public debt (% GDP)
KGOV	Stock of capital public (% GDP)	RERV	Reserves and others (% GDP)***
KPRIV	Stock of capital private (% GDP)	LDI	Total direct investment liabilities (% GDP) ***
GMILTAR	Military expenditure (% GDP)*		
Governance indices **		**Technological***	
COR	Corruption control		
GEF	Government efficiency	CRED_TICS	Credit to TICs (% GDP) ***
ESP	Political stability	US_INT	Internet users (per 100 people)
CREG	Quality of regulations		

RL	Rule of law	

*WB-World Bank Data, World Development Indicators.
**WGI-Worldwide Governance Indicators
***IMF-International Monetary Found

It is worth mentioning that the factor developing or developed country was not included, because there are different classifications that include different characteristics, we also consider a dynamic approach where an economy changes over time.

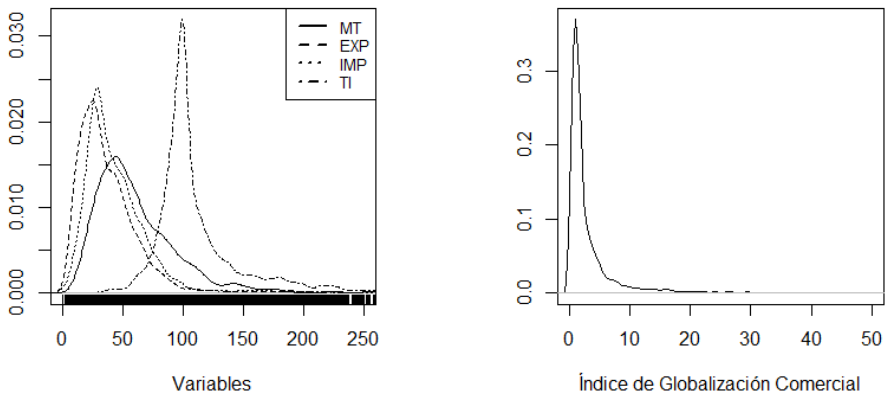
Globalization index

To measure how globalized a country is, the calculation of an index using the Mahalonobis distance between the main commercial variables is proposed. This measure is widely used in various fields for classification and recognition of statistical patterns in multivariate relationships; this distance considers the deviations and covariance between the variables as shown in the following equation.

$$D_M(X) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Where $(x-\mu)^T$ is a transposed matrix that contains the differences of the values with respect to its media μ, y Σ^{-1} is the matrix of variances-covariance inverse of the variables involved (McLachlan, 1999), and X is the matrix of attributes, from which the index is estimated. The Trade globalization index (I_{GC}) is computed with 4,234 observations and includes the volume of imports and exports with respect to GDP, the terms exchange and exchange of final goods $\boldsymbol{I_{GC}} = D_M\ (EXP_{GS_P}, IMP_{GS_P}, MT\ , TI)$.

Figure 1. Distribution of the Trade Globalization Index



Source: Own elaboration in R

As can be seen in Figure 1, both the index and the variables that comprise it have some extreme values, corresponding to countries with high trade volume -with respect to their GDP- such as Aruba, Guinea and Luxembourg. Countries with exports and imports greater than 150% of their gross domestic product are Singapore, Hong Kong and Luxembourg.

With the objective of analyzing the changes of inequality when commercial globalization increases, the observations with positive changes in $\boldsymbol{I_{GC}}$ are filtered from one year to the next, and those records where no inequality changes C_{GINI} were discarded were discarded.

$$C\_I_{GC} = 'C + ' \ \& \ C_{GINI} <> 'SC'$$

The model is constructed based on a data set that considers the 27 factors described in table 1; after filtering the data -according to the previous condition- there are 1,241 records but when eliminating those with at least a null value these are considerably reduced. Then, a DATA database is used with 356 records structured in a panel by country and year unbalanced, despite the loss of information we have data from 72 countries and different regions as shown in table 2.

Table 2. Countries by region and without null values

	Region	Countries by region	Percentage of countries covered by region
1	South of Asia	4	50%
2	Europe and Central Asia	29	49%
3	Middle East and North Africa	4	19%
4	East Asia and the Pacific	11	29%
5	Africa	7	15%
6	Latin America and the Caribbean	15	37%
7	North America	2	67%
	Total	72	33%

Source: Own elaboration

Table 2 shows the number and percentage of countries by region with information in the database, according to that of the World Bank.

Classification decision trees

Given that, the objective of this paper is to find relationships that provide knowledge about the behavior of inequality when commercial globalization increases, Classification and Regression Decision Trees (CART) are used, which allow us to understand this phenomenon through extraction and analysis of knowledge rules.

Econometric and statistical models presuppose the behavior of the data or estimate certain parameters based on the exploratory analysis of the information, however, this becomes less clear when the models are complex or include a large number of variables with non-normal behavior. Decision Trees, on the other hand, are flexible models, between linear and non-parametric, that capture the interaction between the variables from which knowledge rules are extracted (Faraway, 2016)

Formally a CART is a combination of attributes $B(A \cup Y)$, where A is the set of n factors $= \{a_1, a_1, \ldots., a_i, \ldots, a_n\}$, Y the target variable with a domain $dom(Y) = \{c_1, c_{2,}, ..\}$ which contains the possible classes c of Y. In this case, the target variable is the direction of the change in the Gini coefficient (C_{GINI}), given a certain combination of the some of the 27 proposed factors, as described below.

$$B(A_{t-1} \cup C\_GINI_t) \quad , \quad |dom(C\_GINI)| = \{C+, C-\}$$

$$A = \{ \quad GDP_R, \quad POB_{URB}, \ldots \quad , \mathrm{RL}, \quad \mathrm{MT}, \ldots \quad , US_{INT} \quad \}^1$$

The training algorithm is an iterative method to find the decision criteria that classify the examples with the lowest possible margin of error, find $(a^* \forall a_i \in A)$ the best classifies attribute according to the division condition, from which a criterion that divides the data in two subsets is defined, then the best attribute for each subset is searched again (Rokach & Maimon, 2008), in this case until all the observations are classified. To manage the noise, the tree is pruned, eliminating the final nodes that classify data in very specific conditions, which could cause overfitting problems.

The intuitive interpretation of the CARTs, allows to visualize the rules as a combination of criteria (branch) that ends in a classification or result (sheet) associated with a probability. As a condition for dividing a branch of the tree, impurity level D is calculated based on the number of observations correctly classified n_k^C and incorrectly n_k^{NC} accumulated in each k final node is taken.

$$D = \sum_k -2n_k[n_k^C \ln(n_k^C) + n_k^{NC} \ln(n_k^{NC})]$$

To evaluate a CART model, the following adjustment measures are used: the rate of misclassification (MC) -which is the percentage proportion of the number of examples classified incorrectly (nd^{NC}) with respect to the total records in a dataset of observations in specific $MC\ (DATA_{ALL}) = \ (nd^{NC}/nd) * 100$ where nd is the number of records in the data set to be evaluated; the rate of precision when classifying (TP) -important for problems with an unbalanced class distribution because it penalizes the examples classified as positive that are in fact negative; and the number of final nodes (NF) -which represents the degree of complexity of the tree.

Modeling process

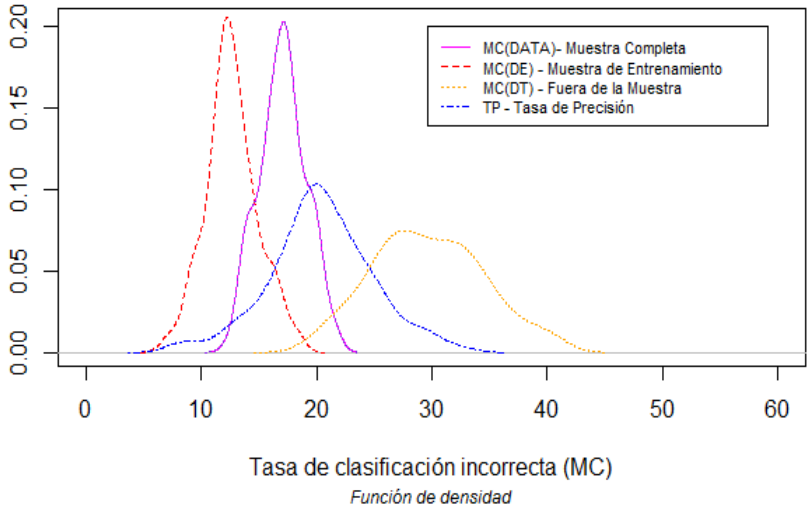
The algorithm for building a tree can generate models with different criteria, which could disregard the results when analyzing the rules of knowledge. To overcome this problem, a model mining process is carried out, which consists of selecting the most stable average tree from a series of models that are constructed by taking different random training samples.

[1] Correspondiente a los 27 factores descritos en el cuadro 1

Around 118 models were build, based on $nS = n/3$ models where n is the number of observations in $DATA$, taking for each one a random training sample (DE) with 267 records corresponding to 75% of the data, to identify each model an initial seed equal to one $SEM\_INI = 1$ is set, which is increased by five in five for each iteration.

For each model, the TP accuracy rate is calculated, which is the percentage of examples that the model identifies as situations where the inequality increases and which coincides with the data in the $DATA$, as well as the incorrect classification rates MC as a function of the observations to consider: out of the sample $MC\ (DT)$, in the training sample $MC\ (DE)$, in the complete sample MC (DATA) and considering the 1,241 MC data ($DATA\_ALL$). Since the construction of the trees is based on training data, the following condition $MC\ (DT) > MC\ (DATA) > MC\ (DE)$ for each model is met. The tree with the lowest MC rate (DE) is also the model with the lowest probability of being generated, that is, it contains the most difficult and rare combination of factors to find, so we focus on the average (stable) trees.

Figure 2. Distribution of measures the model goodness fit

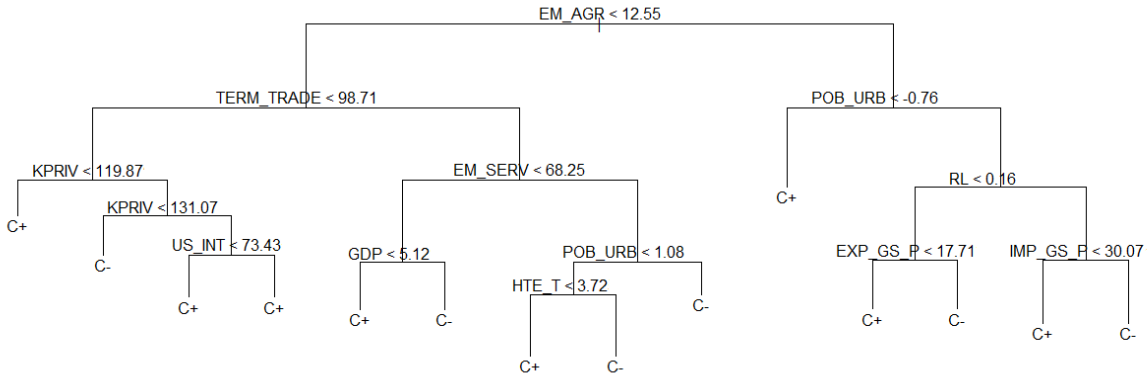


In Figure 2, the behavior of the goodness fit measures of the 118 models is shown, where a model has MC (DE) less than 5%, so it would be correctly classifying 95% of the data but within the certain training sample, so that, despite its efficiency, it has problems of overfiting; It is also appreciated that the sample's raw efficiency of some models is greater than 40%, so regardless of the model to be chosen, it will correctly classify a little less than 60% of the observations.

To choose the best tree among the nS found, the stability of the criteria is considered without leaving aside the efficiency of the model, filtering the models that meet the following condition $(|\ TP - \mu_{TP}| < \varepsilon\ ) \cap \left(\left|\ MC(DT) - \mu_{MC(DT)}\right| < \varepsilon\ \right)$, with an epsilon $\varepsilon = .5$ that increases by 0.1 if there is more than one model within this neighborhood and where μ_TP is the mean of the TP and $\mu_{MC(DT)}$ the average of the MC rates with respect to the test data.

From among models that meet the previous condition, the one with the greatest number of commercial variables is selected, since, we assume that the possible changes in inequality are generated in some way by factors of commercial exchange. In terms of stability, employment in the agricultural sector is the most important criterion since it is presented in 77 of the 118 models built; its interaction with the terms of exchange is presented in 35% of the cases; the first decision nodes of the tree appear in more than 10% of the models; In addition, the factors included in the model are also considered in the other models, that is, each variable appears between 30 and 70 percent of the trees.

Figure 3. Graphical Representation of the Classification Tree



Fuente: Elaboración propia en R con la librería "tree"

The tree model has 14 final rules or nodes, built from 27 criteria and 11 factors, of which 4 are commercial, 3 are population and employment, 1 are technology 1, economic growth, 1 are investment and 1 are governance. The model is constructed with a training sample setting the seed in 3751 within which correctly classifies 89% of the observations, 84% in the complete sample with a precision of 79.61%, out of the sample 70% and an MC considering the 1241 data of 36.74.

Analysis of the Model Results

The results of the tree constructed in the previous section are described and interpreted below, in such a way that we analyze the knowledge extracted from the model. In general, table 3 shows the conditions under which increases in trade openness, measured with the previously estimated index (C_I_GC), are related to changes in income inequality given by changes in the Gini coefficient (C_GINI) .

The most important discriminating factor is the proportion of workers in the agricultural sector with respect to total employment (EM_AGR). This criterion separates countries with a value greater than 12.55 from the right side of the tree, among which are most of the countries in South Asia and more than half of the African and Latin American economies (including Mexico), in total of 397 observations from 74 countries. On the other side of the

tree are the economies with low employment in the agricultural sector that make up a group of 63 countries and 377 data and include more than 65% of European countries[2].

Although, the countries seem to be divided according to the region they belong to, the $idRegion$ variable is not a determining factor, because in both groups there are countries of all regions (except for the Americans that are only Canada and the EU), In addition, the structure of employment changes over time. It should be mentioned that the rules do not classify economies in a certain group of countries with certain characteristics because we assume that these characteristics are dynamic, so that in a country it could be associated with a rule in a certain year and at the same time be associated with another rule in another year.

Table 3. Knowledge rules that lead to changes in inequality when trade globalization increases

RULE	CONDITION	C_GINI	PROB.
R-1	EM_AGR<12.55 & TERM_TRADE<98.7197 & KPRIV< 119.876	C+	1.00
R-2	EM_AGR <12.55 & TERM_TRADE<98.7197 & KPRIV>119.876 & KPRIV<131.077	C-	0.70
R-3	EM_AGR<12.55 & TERM_TRADE<98.7197 & KPRIV>=131.077 & US_INT<73.435	C+	0.90
R-4	EM_AGR<12.55 & TERM_TRADE<98.7197 & KPRIV>=131.077 & US_INT>=73.435	C-	0.60
R-5	EM_AGR<12.55 & TERM_TRADE>=98.7197 & EM_SERV < 68.25 & GDP<5.128	C+	0.90
R-6	EM_AGR<12.55 & TERM_TRADE>=98.7197 & EM_SERV < 68.25 & GDP>=5.128	C-	0.80
R-7	EM_AGR<12.55 & TERM_TRADE>=98.7197 & EM_SERV>=68.25 & POB_URB<1.08183 & HTE_T<3.72539	C+	1.00
R-8	EM_AGR<12.55 & TERM_TRADE>=98.7197 & EM_SERV>=68.25 & POB_URB<1.08183 & TE_T>=3.72539	C-	1.00
R-9	EM_AGR<12.55 & TERM_TRADE>=98.7197 & EM_SERV>=68.25 & POB_URB>=1.08183	C-	0.90
R-10	EM_AGR>=12.55 & POB_URB< -0.76	C+	1.00
R-11	EM_AGR>=12.55 & POB_URB> -0.76 & RL < 0.1631 & EXP_GS_P<17.7135	C+	0.60
R-12	EM_AGR>=12.55 & POB_URB> -0.76 & RL < 0.1631 & EXP_GS_P>=17.7135	C-	0.90

[2] The analysis of the results in each criterion is done by filtering the panel with null data using SQL Server, which although 1241 observations actually depends on the records where the variables don't have null values. For example the variable EM_AGR has 467 observations with some null value, then, the first criterion the remaining 774 records as described in the text

R-13	EM_AGR>=12.55 & POB_URB> -0.76 & RL>=0.1631 & IMP_GS_P< 30.0741	C+	0.90
R-14	EM_AGR>=12.55 & POB_URB> -0.76 & RL>=0.1631 & IMP_GS_P>= 30.0741	C-	0.80

Source: Own elaboration based on the CART model.

In Table 3, the 14 rules of the model are presented, a rule is a combination of conditions that subdivide the observations, each rule corresponds to a route that ends in a final node in the tree of Figure 3 and each has a probability associated to the result factor, that is, {C +, C-}.

Starting from rule 10, in countries with a share of employment in the agricultural sector greater than 12% (EM_AGR> 12.55%) and where the population in urban areas is decreasing to a rate less than 0.76%, inequality increases with a high probability (R-10). As happened in Georgia in 2001 and 2007, where more than 50% of their employment in the agricultural sector their income inequality increased, as in Lithuania (2001-2006) that despite the fact that most employment It focuses on the services sector has significant reductions in growth in the urban population.

On the contrary, if the population growth rate is positive, the rule of law, measured by the globalization index RL, turns out to be a determining factor. The index is shown in standardized values between -2.5 and 2.5, so positive values reveal a good rating regarding the concentration of power. So, if this indicator is practically negative (RL <0.16) and the percentage of its exports with respect to national income is smaller (EXP <17.7), the inequality increases (R-11). This is the least reliable rule of the model, since it indicates increases in inequality with a probability of 60%, classifying Colombia, Egypt and India correctly some years and incorrectly in others, in such a way that there is another phenomenon not contemplated (or that disappeared in the tree pruning process) that prevents inequality in these countries from changing when globalization increases. However, this classification criterion is important for rule 12, which has a probability of correctly classifying 0.9, in addition to 156 observations corresponding to 38 countries having such characteristics in some years. Then the inequality tends to decrease with the commercial globalization, in countries that have a proportion of employment in the agricultural sector greater than 12.5, a constant growth in urban areas, and although a weak rule of law they export more than 17.7% of their product gross domestic product (R-12). It is not that the deterioration of the rule of law promotes economic well-being; on the contrary, this condition could indicate that it is globalization and not the government that has allowed these improvements through wages in these economies with growth potential.

The 24% of the countries have benefited from trade through exports at some point, especially Latin Americans (almost 40%), as was the case of Chile, Ecuador, El Salvador, Guatemala, Nicaragua, Panama, Colombia - which barely achive to export more than 17% for 2008, and Mexico as of 1998 (with the exception of the period between 2004-2007) with

an employment structure by sector averaging 18% in agriculture, 25% in industry and 57% in the services sector. Among the Asian countries with these characteristics and consistent with rule 12 Armenia, Thailand and Cambodia (exports more than 65 of its product), plus Moldova, Turkey, and Egypt, with more than 50% of its production in exports from of 2006.

Like Vietnam (2001-2006), China is an exception to the rule, this country has had increases in inequality since 1983 until 2010 despite its exports exceeding 17.7% of GDP in 1993, with more than 40% of the concentrated employment in the agricultural sector and a high - albeit declining - population growth in urban areas, their exports were not enough to reduce inequality.

On the other hand, countries with good qualifications in the rule of law are usually countries with reliable institutions, good public policies that in turn have high levels of education and, therefore, higher salary premiums, leading to the formation of circles virtuosos (Acemoglu & Robinson, 2013) that by definition tend to reduce inequality. In our model, changes in inequality in countries with an efficient rule of law that have more than 12% of employment in the agricultural sector in addition to urban growth also depend on trade globalization, specifically on the proportion of imports with respect to national income. The model indicates that inequality decreases in countries with a strong state of law, coupled with increases in globalization through imports with levels greater than 30% (R-14), as is the case of Bolivia before 2009, Costa Rica , Honduras, the Philippines, among others, most Asian; but it increases in countries with little commercial iteration IMP <30% (R-13), as is the case of Bangladesh 2000-2003, Pakistan (which also have low percentages of exports) and Indonesia (which has also been reducing its exports in a important). As an exception to this rule (R-13), Brazil has been reducing its inequality since 1988 despite the fact that its imports do not reach even 15 percent of its product, which could be more related to the efficiency of government and public policies focused on education that increase the supply of work in the services sector.

With respect to the rules in which the employment in the agricultural sector is less than 12.55%, corresponding to the left side of the CART model (figure 3), the deterioration of the terms of trade is a determining characteristic for the increase in inequality . This criterion divides the observations, according to whether they have deteriorated (TERM_TRADE <98.8) or not the terms of exchange (TERM_TRADE> = 98.8).

In the first part of the tree (on the right), we find that in countries with little employment in the agricultural sector (EMP_AGR <12.5%) and in which private capital is lower than the gross domestic product (KPRIV <120%), the deterioration of the terms of trade is accompanied by increases in income inequality (R-1), this was the case of Bolivia in 1992 and 2001, Lithuania 2008-2013 and Switzerland 2005-2013; but if investment in the private sector exceeds GDP by more than 30% (KPRIV∈ (120,130)), the Gini coefficient could decrease (R-2). So, one way to prevent inequality from increasing when terms of trade deteriorate is to increase capital but only to a certain extent (119.9 <KPRIV <131.0), this

for economies that have left behind agricultural production, situation in the Netherlands 2007, Estonia 2005-2006, Greece 2004,2007, Ireland, Republic of Mauritius-Africa and Slovakia 2010, although not for all years.

If private capital is the presently large (KPRIV> 131), then the effect of trade globalization on the Gini coefficient -added to the deterioration of the terms of trade- depends on technological inclusion specifically in terms of communication, measured in this case with the number of internet users per a hundred. In such a way that, if more than 27% of the population lacks Internet access, inequality increases (R-3) and, on the contrary, it decreases in countries with more than 73% (R-4).

Austria is an excellent example of the effect of digital inclusion on the inequality determined in rules 3 and 4, with a rate of 39 in 2001 -when has inequality up- then it was increasing Internet access until in 2007 the limit passed of the condition (US_INT> 73) and the measure of inequality began to decrease. Another interesting aspect is that most of the years in which the R-3 is completed coincide with periods of crisis -the technological crisis (2002-2005) or the financial crisis (2007-2010)[3]; while on the other side of the node, in the R-4 it is shown that Switzerland, Austria, Finland and E.U. They presented reductions in inequality after they increased Internet access, including the period between 2007 and 2010. This suggests that digital inclusion will help prevent inequality from increasing in times of crisis. In addition to that, there could be a relationship between crises and inequality, which would be affected by technological inclusion, if so, we would be omitting an important crisis factor.

Returning to the importance of the terms of trade, in those economies in which the value of their exports are worth more than their imports, employment in the agricultural sector is less than 12%, employment in the service sector is less than 68% and , therefore, employment in the industrial sector greater than 20%, the effect of greater economic integration on income inequality is determined by the growth rate of its production. As described in rules 5 and 6, unless that rate exceeds five percent (R-6), inequality increases with a high probability (R-5), which is why countries such as Australia (2004-2006)), Bulgaria (2010-2013), Croatia (2008), Germany (2001-2004) Portugal and Russia, among others, presented increases in inequality. Therefore, in this model, the economic growth rate acts as a discriminant. The model indicates that it is possible to reduce inequality without reaching 68% of employment in the services sector, if the industrial sector is efficient enough to generate growth of more than 5%, with acceptable terms of trade. This suggests a relationship between economic growth and income inequality when commercial globalization increases coupled with the conditions established in rule 6.

[3] Alemania (2005-2006), Luxemburgo (2005), Portugal (2003-2005, 2012), Hong Kong (2005), Japón (2005, 2011-2013), Corea (2001-2004), Singapur (2002-2006), Suiza (2005), EU (2005), Francia (2008-2009), Italia (2008-2011), Irlanda (2008-2010), España (2008-2012), Israel (2001-2008)

For economies that have good terms of trade (TERM_TRADE> 98.7) and a labor market structure in which workers in the agricultural sector constitute less than 12%, those in the industrial sector less than 32%, but above all in which workers in the service sector represent more than 68% of the labor force, income inequality increases due to the slow growth of the urban population (POB_URB <1.08%) coupled with the wasted education of its workforce to export more of 3% in high technology (R-7). The economies that suffered inequality increases due to this type of stagnation were: Australia in 2007, Denmark since 2006, Hong Kon at the turn of the century, New Zealand (2011, 2013), Uruguay 1995-2002 and even the United States and the United Kingdom in 2003.

If the population in urban areas continues to grow at a rate of at least 1%, it is common that the inequality decreases (R-9), otherwise it will decrease only if it exports high technology in a percentage higher than 3.72% of the total exports (R-8). The countries that have benefited from the globalization of trade, due to their urban growth coupled with the characteristics of their labor market are: Argentina, Peru, Canada, The Netherlands, New Zealand, Norway, South Africa, Switzerland and the United Kingdom (R- 9); and for exporting more than 3.7% of goods and services considered high technology are: Sweden in 2001-2002, UK in 2002 and Belgium in 2004-2007 (R-8). The United States is an exception to rule 9, since, despite the rapid growth of its urban population, its inequality stopped growing until 1993 when its exports in high technology exceeded 1.89%, as did Argentina until 2001 when it began to export more .05%.

Conclusions

Although the increase in inequality since 1980 is clear, it is not possible to generalize this phenomenon due to the heterogeneity of the countries and temporary changes (Ravallion, 2003). However, the intuitive interpretation of the presented CART model allows to identify circumstances and key factors to analyze the effect of commercial globalization on social welfare, specifically on the inequality in the income of households within a country.

The structure of the labor market reflects the effects of industrialization and technological progress on the distribution of income of the population, consistent with Kuznets (1955), Milanovic (2016), however, it is necessary to specify under what conditions the result is positive or not for inequality.

In non-agricultural countries, in which private capital is lower than its production level, the deterioration of the terms of trade is accompanied by increases in inequality (R-1), in this case private investment and digital inclusion play an important role, it can reduce inequality, if more than 27% of the population lacks access to the Internet the effect on inequality is positive.

The tree indicates that in countries with high levels of employment in the services sector and without deterioration in the terms of trade, mostly developed economies, income inequality may increase during the period of stagnation (R-7). On the contrary, this type of

countries reduce their inequality with rates of growth above 5% (R-6). So, for the economic well-being of the population of these countries to improve, it is important that they do not stagnate and exploit the potential of their high concentration of skilled labor by exporting high technology. This is consistent with a period of adjustment in the structure of the labor market, in which the society pays a cost of learning, but once the labor force in the services sector is strengthened, sufficient returns are generated to subsequently reduce their inequality (R-8).

The change in the structure of the labor market in which a greater concentration of workers in the service sector is allowed is crucial for the distribution of income, due to the salary premium and the wage differences between sectors. As emphasized in the analysis, agricultural economies (with more than 12% of employment in this sector) but in industrialization processes (reflected in the positive urban population growth), characteristics of developing countries, present reductions in inequality before increases in commercial globalization, consistent with the differentiation highlighted by Dollar (2004). Although the CART model also identifies the efficiency of the rule of law as a factor to promote a reduction via imports (R14), for countries with deficient rule of law, the reduction is via exports (R-10).

References

Acemoglu, D., & Robinson, J. A. (2013). Why nations fail: The origins of power, prosperity, and poverty. *Broadway Business*.

Aghion, P., Caroli, E., & Garcia-Penalosa, C. (1999). Inequality and economic growth: The perspective of the new growth theories. *Journal of Economic literature*, 37(4), 1615-1660.

Asteriou, D., Dimelis, S., & Moudatsou, A. (2014). Globalization and income inequality: A panel data econometric approach for the EU27 countries. *Economic modelling*, *36*, 592-599.

Akerman, A., Helpman, E., Itskhoki, O., Muendler, M. A., & Redding, S. (2013). Sources of wage inequality. *American Economic Review*, *103*(3), 214-19.

Alderson, A. S., & Nielsen, F. (2002). Globalization and the great U-turn: Income inequality trends in 16 OECD countries. *American Journal of Sociology*, *107*(5), 1244-1299.

Atkinson, A. B., Smeeding, T. M., & Brandolini, A. (2001). Producing Time Series Data for Income Distribution: Sources, Methods, and Techniques. *Maxwell School of Citizenship and Public Affairs*, Syracuse University.

Atkinson, Anthony. B. (2003). Income inequality in OECD countries: Data and explanations. *CESifo Economic Studies*, *49*(4), 479-513.

Atkinson, A. B., Piketty, T., y Saez, E. (2011). Top incomes in the long run of history. *Journal of economic literature*, *49*(1), 3-71.

Atkinson, Anthony. B. (2015). *Inequality: what can be done?* Harvard University Press, ISBN 9780-674-28-7037, 398 p.

Bluestone, B., & Harrison, B. (1982). The deindustrialization of America: Plant closings, community abandonment, and the dismantling of basic industry (Vol. 312). New York: Basic Books.

Bourguignon, F. (2017). The globalization of inequality. Princeton University Press.

Deininger, K., & Squire, L. (1996). A new data set measuring income inequality. *The World Bank Economic Review*, *10*(3), 565-591.

Dollar, D., & Kraay, A. (2001). Trade, growth, and poverty. *World Bank, Development Research Group, Macroeconomics and Growth.*

Dollar, D. (2004). Globalization, poverty, and inequality since 1980. *The World Bank*.

Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. (Vol. 124). CRC press.

Goldberg, P. K., & Pavcnik, N. (2007). Distributional effects of globalization in developing countries. *Journal of economic Literature*, *45*(1), 39-82.

Han, J., Liu, R., & Zhang, J. (2012). Globalization and wage inequality: Evidence from urban China. *Journal of international Economics*, *87*(2), 288-297.

Harrison, A., & Hanson, G. (1999). Who gains from trade reform? Some remaining puzzles1. *Journal of development Economics*, *59*(1), 125-154.

Heathcote, J., Storesletten, K., & Violante, G. L. (2017). Optimal tax progressivity: An analytical framework. *The Quarterly Journal of Economics*, *132*(4), 1693-1754.

Helpman, E., Itskhoki, O., & Redding, S. (2010). Inequality and unemployment in a global economy. *Econometrica*, *78*(4), 1239-1283.

International Monetary Fund (2007). Globalization and Inequality. En *World Economic Outlook*, Washington DC. 31–65.

Jaumotte, F., Lall, S., & Papageorgiou, C. (2013). Rising income inequality: technology, or trade and financial globalization? *IMF Economic Review*, 61(2), 271-309.

Jeon, K., & Kabukcuoglu, Z. (2018). Income inequality and sovereign default. *Journal of Economic Dynamics and Control*, *95*, 211-232.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters VIII: aggregate and individual governance indicators, 1996-2008.

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

Lawrence, R. Z., Slaughter, M. J., Hall, R. E., Davis, S. J., & Topel, R. H. (1993). International trade and American wages in the 1980s: giant sucking sound or small hiccup? *Brookings papers on economic activity. Microeconomics*, *1993*(2), 161-226.

Milanovic, B. (2016). Global Inequality: A New Approach for the Age of Globalization. *Harvard University Press*.

Nissanke, M., & Thorbecke, E. (2006). Channels and policy debate in the globalization–inequality–poverty nexus. *World development*, 34(8), 1338-1360.

North, D. C. (1990). A transaction cost theory of politics. *Journal of theoretical politics*, *2*(4), 355-367.

Palma, J. G., & Stiglitz, J. E. (2016). Do nations just get the inequality they deserve? The “Palma ratio” re-examined. In *Inequality and Growth: Patterns and Policy* (pp. 35-97). Palgrave Macmillan, London.

Paskov, M., & Dewilde, C. (2012). Income inequality and solidarity in Europe. *Research in Social Stratification and Mobility*, *30*(4), 415-432.

Ravallion, M. (2003). The debate on globalization, poverty and inequality: why measurement matters. *International Affairs*, *79*(4), 739-753.

Rokach, L., & Maimon, O. Z. (2008). Data mining with decision trees: theory and applications (Vol. 69). *World scientific*.

Shahbaz, M. (2010). Income inequality-economic growth and non-linearity: a case of Pakistan. *International Journal of Social Economics*, *37*(8), 613-636.

Solt, F. (2016). The standardized world income inequality database. *Social science quarterly*, *97*(5), 1267-1281.

Stiglitz, J. E. (2010). *El malestar en la globalización*. Taurus.

Stolper, W. F., & Samuelson, P. A. (1941). Protection and real wages. *The Review of Economic Studies*, *9*(1), 58-73.

Tinbergen, J. (1970). A positive and a normative theory of income distribution. *Review of Income and Wealth*, *16*(3), 221-234.