Journal of Research, Innovation and Technologies https://doi.org/10.57017/jorit.v4.1(7).01

A Machine Learning Approach to Synthetic Gini Coefficient Estimation in Colombian Municipalities

John Michael RIVEROS-GAVILANES ✉ ✉
Faculty of Administration and Economics
University College of Cundinamarca[1], Colombia
https://orcid.org/0000-0003-4939-0268

Abstract

This paper presents two synthetic estimations of the Gini coefficient at a municipality level for Colombia in the years 2000-2020. The methodology relies on several machine learning models to select the best model for imputation of the data. This derives in two Random Forest models where the first is characterized by containing Dominant Fixed Effects, while the second contains a set of Dominant Varying Factors. Upon these estimations, the Synthetic Gini Coefficients for both models are inspected, and public links are generated to access them. The Dominant Fixed Effects models is rather "stiff" in contrast to the Varying Factor model. Hence, for researchers it is recommended to use the Synthetic Gini Coefficient with Varying Factors because it contains greater variability across time than the Dominant Fixed Effects models.

Keywords: Gini, machine learning, random forest, estimation, synthetic, economics.

JEL Classification: C80; H70; O10; P19.

Introduction

The objective of this paper is to present a machine learning estimation for the scarce data of the Gini coefficient at a municipality level for Colombia between the years 2000-2020. Using the power of the CEDE data of the University of Los Andes (CEDE, 2023), and the only existing information for the municipality Gini coefficient in 2005, this empirical exercise extends "synthetically" the measures of the Gini coefficient. By using the best model across several estimations, the imputation of the Gini coefficient is executed, allowing to synthesize the data in the panel data format including the identification keys with the DIVIPOLA municipality and the years. During the estimations, it was noted that the best model had as the majority of important variables (or features), a set of "Dominant Fixed Effects" which drove the Random Forest results. The features of this model are strongly related to municipality clusters, distances, and geographical locations. By interpreting these features related mostly to fixed effects -as they do not fluctuate significantly over time and are mostly related to geographical ID clusters- an alternative model is generated characterized by containing a set of Varying Factors. The latter model relies on a set of socioeconomically time-varying factors mainly concentrated in human capital accumulation and population dynamics.

Both the Dominant Fixed Effects model (R^2 = 98.0%), and the Dominant Varying Factor model (R^2 = 94.3%) estimated via Random Forests, exhibits good properties in terms of the metrics, but they also exhibit differential variability over time and within individuals. Hence, while the first model (Dominant in Fixed Effects) remarks the importance of time-invariant factors such as distances, geographical regions, and potentially static phenomena such as time-invariant infrastructure, institutions and climate conditions.

[1] Address Cl. 28 #5B-02, Faculty of Administration and Economics, Bogotá, Colombia

The second model remarks on the importance of human capital and socioeconomic factors to describe the Synthetic Gini Coefficient. In particular, the Dominant Fixed Effects model also contains as important variables the socioeconomic variables of population and human capital, but they're not as important as the fixed or "semi-fixed" variables in the Random Forests estimations.

Consistently, in both estimations, Chocó is the Department most affected by income inequality given the results of the Synthetic Gini Coefficient. Both estimations also depict that Colombia has not improved significantly the levels of income inequality during this period.

The details of this estimation go as follows; first, a filtering process is implemented to reduce the number of features with missing values in the available data. Next, a model selection between the machine learning approaches of linear regression, random forest, regression trees, and gradient boosting is implemented. The random forest outperforms the rest of the models based on the observed available data through cross-validation techniques. By observing the composition of the most important variables in the initial model, a strong composition of "fixed" and "semi-fixed" variables exists and are characterized by not changing over time. Upon this interpretation, a second model is generated by excluding this fixed and semi-fixed variables. And thus, the Varying Factor model is estimated.

The public available version of the Synthetic Gini Coefficient for the Dominant Fixed Effects model is: https://docs.google.com/spreadsheets/d/1jc1c-X1aum8GkfrsZHOec1gz_1Qo9w_S/edit?usp=sharing&ouid=118158209086311183140&rtpof=true&sd=true. The public available version of the Synthetic Gini Coefficient under the Varying Factor model is: https://docs.google.com/spreadsheets/d/1JUt93lBzp3S_kgWnU4msuDJ2LQBGsbVq/edit?usp=sharing&ouid=118158209086311183140&rtpof=true&sd=true

Considering that the Varying Factor model contains more variability across the municipalities for the sample years, it is recommended to be used for researchers. In essence, because the fixed effects can be isolated by using standard econometrical specifications. The Varying Factor model is in my perspective also more realistic, because the core important predictors are time-varying. Finally, while this article contributes with synthetic measures of the Gini coefficient at a municipality level for Colombia during the 2000-2020, it is important to remember the weaknesses of the machine learning methods for data imputation. This implies that the motivation for solving the missing data problem is specific (Lakshminarayan et al., 1996).

Hence, the Dominant Fixed Effects imputation is not necessarily bad in contrast to the Varying Factor imputation. Both reflect synthetic estimates of the Gini coefficient but based on different natures. Hence, no single machine learning algorithm is the ultimate answer for the missing data problem (Hong & Lynn, 2020). Also is important to highlight that all machine learning methods are subject to bias and errors, implying a risk of biased estimates in the parameters and error in the imputations (Sullivan et al., 2017).

The justification to select this time frame (2000-2020) is mainly related to data limitations, as at a municipality level, there is a significant number of missing values for the Colombian municipalities, which neglects the use of a greater selection of variables. The estimates provided for this period, however, do possess some significant available information which is both time-varying and invariant, allowing to proceed with sensibility analysis. The methodology is based on available real-world administrative information to execute the estimation process, and the reference year is the 2005 since it is the only period with the Gini coefficient at the municipality level according to the CEDE. While it is not possible to discard the estimation backwards (before 2000), the estimates are not as accurate due to missing values, administrative changes in the municipalities, and potential inconsistencies across the measure of the explanatory variables.

The limitation of this study is that the synthetic Gini index distribution estimated in this article, might not describe or align with the real-world inequality distribution of Colombia as synthetic data is only generated from observed available variables in the essence of Rubin (1987). Hence, the imputed dataset is a synthetic estimation, which might not reflect the true variance of the population across units or time. Leading to potential inaccuracies that can exists (Schafer & Graham, 2002). For these reasons and the potential biases that exist within the machine learning imputation, the dataset will never produce the pure and correct estimate for the income inequality (Gelman

and Hill, 2007) and it is impossible to compare with the real data of income inequality as the latter does not exist. Hence, the datasets generated should be used as informative observations.

Nevertheless, this article contributes to the understanding of income inequality through the generation of the Synthetic Gini Index and its potential evolution based on observed variables at the municipality levels. Thus, it provides two approaches based on either the dominance of fixed characteristics, or the time-varying features. It also presents a potential result for the estimation of income inequality over time across the territories of Colombia.

Finally, the additional findings present the well-known correlations between income inequality and human capital accumulation (inverse in nature), the population effects on income inequality (where massification becomes evident deriving in a nonlinear U-shape relationship), and the effect of rurality on income inequality (where the rural population is in a disadvantage since it is likely to encounter higher income inequality).

This study also contributes to the missing data problem that the Colombian territory suffers in terms of the analysis of income inequality, considering that the Colombian territory is one of most unequal economies of the world. This study was also influenced by the relevant works of Xue (2023), Lin et al. (2022), Seu et al. (2022), Alwateer et al. (2024), Sun et al. (2023), Wang et al. (2019), Gond et al. (2021), and Lin and Tsai (2020).

The document continues with Section 1, providing a brief literature review, section 2 describes the estimation process, Section 2.1 describes the data transformation used to preserve the number of variables with the less count of missing values. Section 2.2 provides the generalities of the machine learning estimation, where the model performance through several machine learning models is compared. It presents the importance of the variables of the data, where in Section 2.3 an interpretation of the variables within the random forest is analysed. In this interpretation, the nonlinear correlations of some important variables are analysed relative to the Synthetic Gini Coefficient, to further explore the data at the Department level and over time. Section 2.4 presents the estimation with time-varying factors and the comparison across models, the importance of time-varying factors which are mainly concentrated in human capital accumulation and population dynamics. Section 2.5 presents the nonlinear dynamics between the Synthetic Gini Coefficient and the regressors previously analysed where virtually the same patterns emerge. Finally, Section 2.6 briefly reviews the geographical distribution of the Synthetic Income Inequality for both the Dominant Fixed Effects model and the Varying Factor model.

1. Research Background

The techniques to estimate missing values has greatly increased during the last two decades (Rácz & Gere, 2025) depicting potential alternatives to satisfy the world's data problems. The concept of strategic imputation takes place when we consider the available data, the limitations and methodologies to create the most accurate and potential result for decision making. The literature of Gama et al. (2025), Sharma et al. (2025), Shin et al. (2025) provide some great examples of applied data imputation methods to multiple topics not only in social sciences but also in other fields. In particular, the economic science is not isolated to the use of these techniques. The study of Teng et al. (2024) reviews under machine learning methods how income inequality can be understood from differential effects of oil and gas rent, also investigates if a potential Dutch disease might be occurring in the countries of Bahrain, Kuwait, Oman, and Saudi Arabia, this study constitutes a corner stone in the implementation of machine learning methods in the economic analysis related to income inequality.

In the same line for economic sciences, the study of Silva et al. (2024) performs a forecast analysis of trade patterns for specific commodities, the machine learning models they estimated depend on several available macroeconomic variables, including topological variables. The estimations involved around 200 countries and provide a rich understanding of the behaviour of gross domestic production.

Related to these estimation procedures, the work of Ma et al. (2023) imputes an innovation index based on machine learning methods (k-means and random forest models) considering several economic variables which are available. These variables include income, unemployment, social characteristics, and education. The results of the application of these machine learning methods generated the Global Intelligence Innovation Index (GIII) which helps the evaluation of intelligence innovations across countries.

Another influential study for economics and machine learning methods is the study performed by Zhan et al. (2023) where the goal is to estimate the house prices. For this study, a hybrid machine learning approach is developed with a mixture of Bayesian statistics and machine learning approaches. The output of this study is a multi-source dataset containing information on real estate markets with the imputed prices.

In this line, the study of Caravaggio et al. (2025) applies several machine learning methods to predict the European Union allocation of resources in Italy. These methods involve Random Forest, XGBoost, and Support Vector Machine, it is found that such transfers depend on territorial traits (fixed), and economic factors including income, unemployment, and debt.

In the case of variables such as welfare, machine learning methods have been applied to inspect how debt might interact with social welfare, the study of Zhu & Huang (2024) reveals that through machine learning methods, the relationship between debt and welfare is affected mostly by macroeconomic variables including gross domestic production, unemployment and fiscal status.

The process of recovering economic data, has been executed also with machine learning methods as the article of Combes et al. (2022) demonstrates. In this study, historical information about urban economics is used, where the power of machine learning methods helps to understand and estimate the spatial distribution of several economic outcomes.

Related to recovering income data, the study of Gao et al. (2024) uses human mobility patterns to investigate and estimate the income distribution through machine learning. Human patterns derived in a series of indicators which are used to estimate the income distribution. In general, the deep learning models outperform other approaches, and an important result is that residential differentiation of income is significantly greater than mobility disparities. Some of the models executed in this research involve XGBoost, convolutional neuronal network, and time series graph neuronal networks.

These studies depict how differential machine learning methods are used for imputation, and they provide a framework to operate in the context of this study, considering that the goal is to recover and estimate the Gini coefficient as a measure of income inequality for the municipality level in Colombia. The research gap is the lack of data and studies about this measure of income inequality at the granular level of municipalities, imposing a significant challenge for decision making processes and the overall study of the evolution of income inequality in Colombia.

2. Methodology and General Process

Since the CEDE (2023) is one of the sources of information that consolidates the municipality data of Colombia for a significant number of years, I selected this data to be core of the empirical estimations. That said, CEDE data composes for this version of the study (1.02.2025) seven panels of information. These are related to: 1) Agriculture and land. 2) Good government. 3) General characteristics (2022). 4) General characteristics (2023). 5) Conflict and violence. 6) Education. 7) Health and services.

Unfortunately, for all these datasets not all the data is available. In particular, in the panel of General Characteristics (both years), the Gini coefficient only presents the information for the year 2005 at the municipality level. The Gini then is missing for all the remaining years, consolidating a gap in our knowledge of the behaviour of income inequality across the country.

2.1. Data Transformations

The first step involved identifying which of the seven information panels contained the highest number of available observations. The analysis determined that Panel 2: Good Government had the most extensive dataset (N = 43,541). Consequently, this panel was retained as the primary dataset for the merging process.

The second step was to select the join keys for the grand merger between panels. As it is well known, the municipality codes (called DIVIPOLA) and years are the proper identifications for the subsequent task to estimate variables at the municipality level. The result of this joint created a panel of 43541 observations with 2718 variables.

The third step was to retain in this aggregated panel the information between the years 2000 and 2020. After the suppression of observations by the previous condition, the resulting panel contained 23,563 observations. As some of the panels had the same variables (thus repeated), an algorithm to delete the duplicates was executed. This let a panel of 2,654 variables and 23,563 observations.

The fourth step was to count the missing values for all the variables and then calculate the percentage of missing values per variable. A strong condition was later applied to the data at this point, and it was to retain only the variables with 1% or lower relative to the count of missing values. This to ensure as possible the existence of real information (without any imputation on the features). Further some empty (but not missing) character variables were dropped [2]. The resulting panel at this point contained 48 variables and 23, 563 observations.

The fifth step was to convert all character variables to numerical variables, this left some empty but not missing variables again[3]. In this point, a final condition is applied to retain only complete cases for the panel. The resulting panel then contained 44 variables and 23 082 observations.

The sixth step was to recover the unique information of the Gini available only in the year 2005 which was inside the panel of General Characteristics 2023. Then it was merged back into the cleaned panel without missing values for the panel to proceed with the estimation techniques. The panel for these processes contains 45 variables and 23,082 observations.

2.2. Machine Learning Estimation

The software and programs used to estimate the different machine learning models, and their graphics are developed by Ridgeway (2007), Kuhn (2008), Wickham (2011), Therneau et al. (2015), Wickham et al. (2019), Yarberry (2021). Where the structure was checked first to ensure that all variables/predictors or "features" were numerical (including the conversion from categorical ones). The second step for the machine learning imputation was to identify the data where the Gini exists and where it is missing.

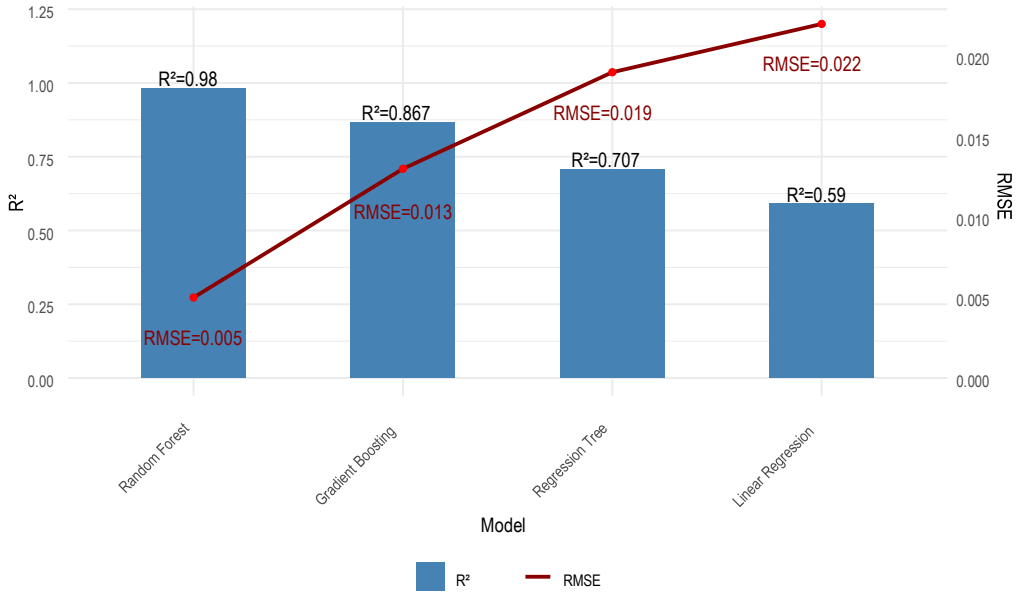
The third step was to define the predictors (44 variables) and the target variable (the Gini). Then a seed was placed to train models upon cross validation. The number of folds selected for these processes was established in five to train the algorithm with the cross-validation approach. Given that some variables are duplicated in essence, and that some ID variables are also included in the panels, a cleaning process is done to let one type of cluster ID variable for geographical/administrative locations.

The fourth step was to estimate the linear regression, regression tree, random forest and gradient boosting models to investigate which was the most suitable on the training data. The results of step 4 are represented in Figure 1 which contains the classic measures of R^2 and RMSE for continuous outcomes. These estimations suggest that random forest approach outperforms the rest of the models. It exhibits an $R^2 = 98.67\%$ with an RMSE = 0.005 suggesting a precise result even after cross-validation with five folds. With this best model (the random forest), the Gini data is imputed for the whole panel without missing values (K = 45, N = 23,082) including the target variable. The panel hence contains information of about 1091 municipalities (this vary depending on the number of non-missing values on the predictors) for the years 2000 and 2020.

[2] This applied to the variables containing a "" in the panel, and it was related to the variables "DF2 categorica, DF2 doinicial, DF2 rango, categoria" which belongs to the panel of good government but were considered not empty. Hence these were eliminated

[3] The variables of "depto, provincia, municipio" and "act adm" which were incomplete identifications of the municipalities, not important as the code for each of them are in other variables

Figure 1. Model estimation performance



Source: Own elaboration

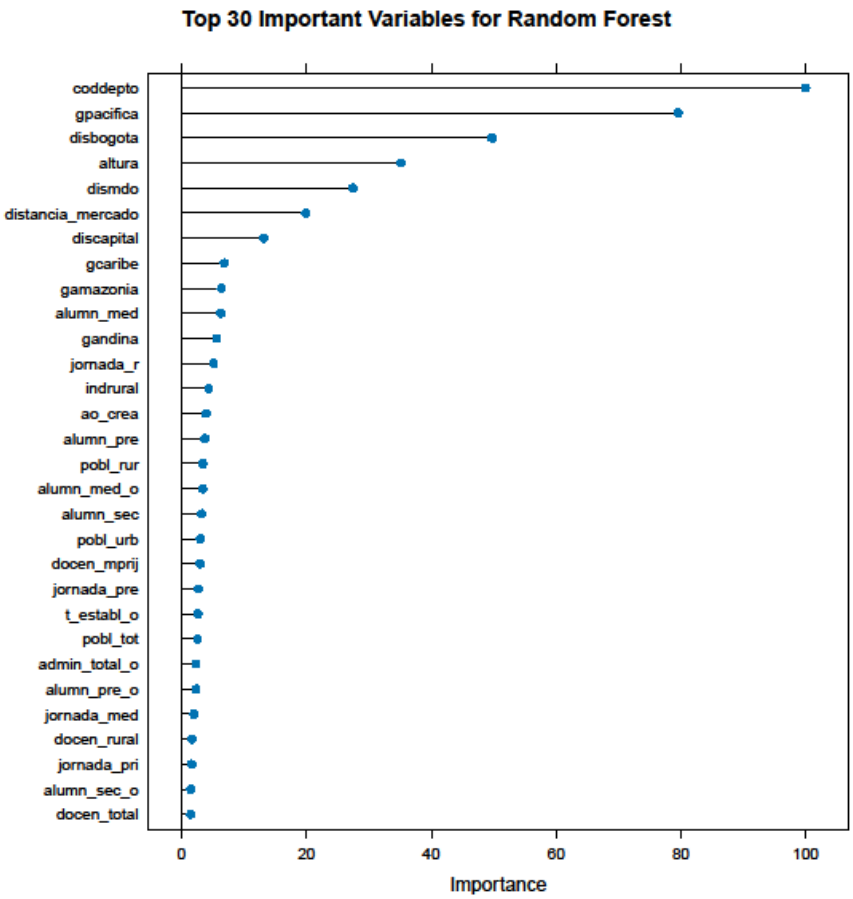
2.3. Interpretation of Variables within the Random Forest

An additional important result is the importance of the variables for the random forest estimation. Which are displayed in Figure 2. By looking at Figure 2 it is possible to identify a pattern in scale of importance for the random forest estimation. The starting variable "coddepto" which refers to the geographical cluster of the administrative divisions of Departments of Colombia. This is the next administrative division after the Nation. The variable "gpacifica" relates to the location of the "pacific region" of Colombia. Then is followed by the distance towards Bogota (the capital of Colombia), the altitude, the linear distance to the most import food market nearby (in *Km*), the linear distance to the nearby municipality where the highest food market is located, the linear distance to the capital of the Department followed by other fixed location variables such as the Andina, Amazonia, and Caribbean regions. This first set of variables has one thing in common. They are variables related to the geographical physical locations of the municipality.

On the second set of important variables, it is found that socioeconomic characteristics involve population dynamics, school, educational and human capital topics. This set the variables includes the rural density of the municipalities, the total number of school schedule (classes), total number of students in middle education, the total number of students in kindergartens, the total number of students in secondary education, the urban and total population of the municipality, the number of teachers, the number of educational facilities, and the number of school schedules for primary education.

The selection of fixed or "semi-fixed" characteristics, such as geographical clusters and distances, is interpreted as a form of fixed effects from an econometric perspective. In general, there are some existing time-invariant characteristics which can explain a significant portion of individuals' heterogeneity; this implies that such constant fixed effects are present. This fits into the categories of distance and geographical static clusters. Beyond this interpretation, they could include institutional time invariant characteristics like local authorities' performance, constant physical infrastructure and roads of communication or access to other municipalities, in particular, the capital of Colombia, Bogota, and capital of the departments. Market interconnections play a key role here as well.

Figure 2. Important features of the estimation

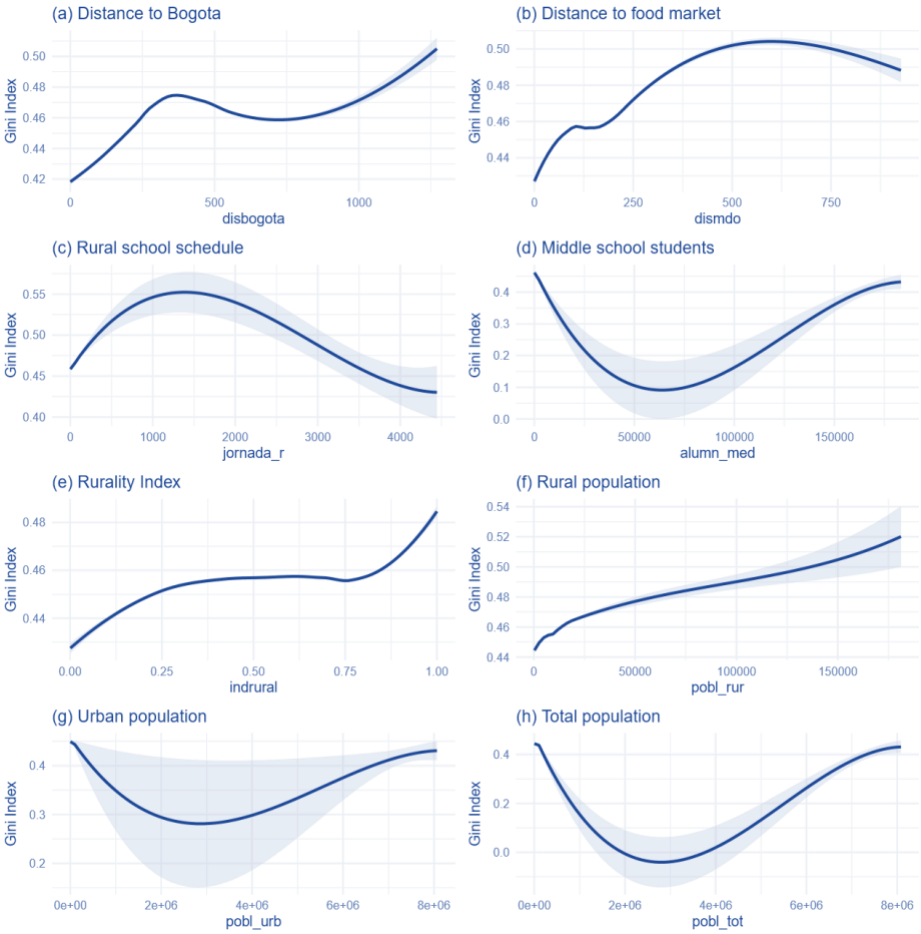


Source: Own elaboration

According to Figure 2 these fixed effects are the main drivers of the inequality across the Colombian municipalities. The question of what and how much will change the estimates will be addressed in the next section with varying factors. For now, using this data it is interesting to see how income inequality might relate to the continuous variables used in the random forest estimations. This is to understand a bit better the black box inside the synthetic data and what patterns can be empirically identified.

Figure 3 depicts some of the non-linear interrelations between the Synthetic Gini coefficient and some of the primary continuous regressors (or features) used in the Random Forest model. Some clear patterns emerge from these dispersions: a) The larger the distance of the municipality relative to the capital of the country (Bogota) implies a larger income inequality. b) The larger the distance of the municipality relative to the largest food market nearby implies a larger income inequality. c) An increasing size of the class schedules in rural areas tends to drive a higher inequality until a turning point where class schedules are large enough to invert the pattern, resulting in a decrease in inequality (inverted U-shape relation). d) There is a decreasing relationship between inequality and the number of middle school students, however, at some point, when students conglomerate, the inequality starts to grow again. e) When the ratio of rural population surpasses the urban population implies a growing inequality. f) As a consequence, there's a positive and almost linear relationship between the rural population and income inequality. g) Urban population has a decreasing effect on the estimations of income inequality, however, when urban population massifies, the inequality starts to grow again (U-shape relation). h) The total population mimics the behaviour of the urban population, hence when massification arises, inequality grows again.

Figure 3. Dispersion of Synthetic Gini Index and features



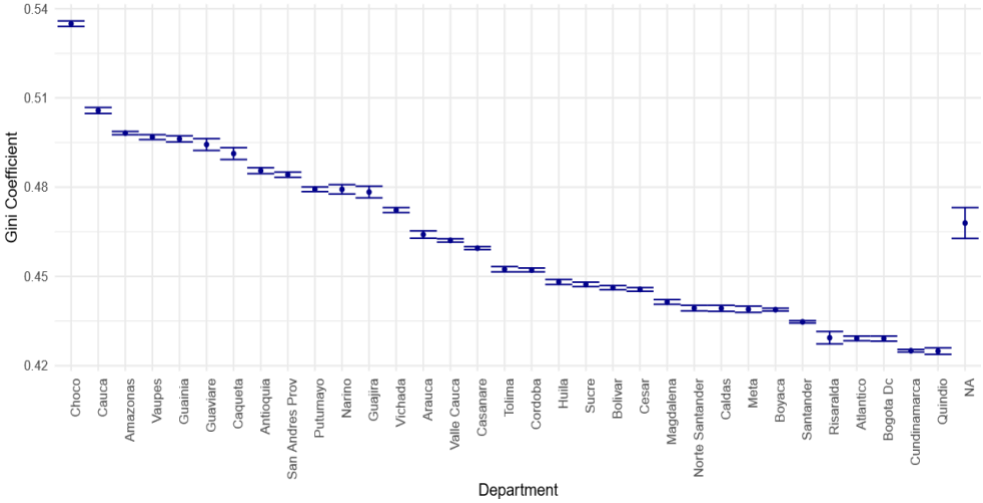
Source: Own elaboration

These patterns are consistent with the economic literature[4], where for example, the peripheral regions or zones are prone to suffer from income inequality given the lack of market interconnexion, public services and the provision of goods. This is why patterns of Figure 3 relative to panels (a) and (b) are strong. These capture the relative distance to the capital of the country, and the linear distance to the central food markets. According to this argument the population dynamics such as rural concentration and rural population, seem to be positive correlated with the synthetic inequality, as witnessed in panels (e) and (f). Finally, the nonlinear dynamics of human capital accumulation are also very interesting to analyse. The fact that rural school schedules have an inverted U-shape reflects some of the internal dynamics related to opportunities and school attendance which will tend to decrease inequality if they are high enough. When school schedules satisfy the demand of schooling in rural territories, it is most likely to encounter a reduction in the synthetic inequality. On the other hand, from the side of the demand as seen in panel (d), when the number of middle school students increases enough, students will face a constraint in the access of education, delivering an increase in synthetic inequality. Finally, population dynamics relative to the urban population (g) and total population (h) have a U-shape relationship, implying that when massification of the municipalities occurs, synthetic inequality will tend to increase.

[4] This aligns with the studies of Paas and Schlitte (2008), Rey (2004), Salvati (2016), Kühn (2015), Oppido et al. (2023), Riveros-Gavilanes (2023), Lee and Lee (2018), Castelló-Climent & Doménech (2021), Lee & Vu (2020) relative to these observe variables and income inequality.

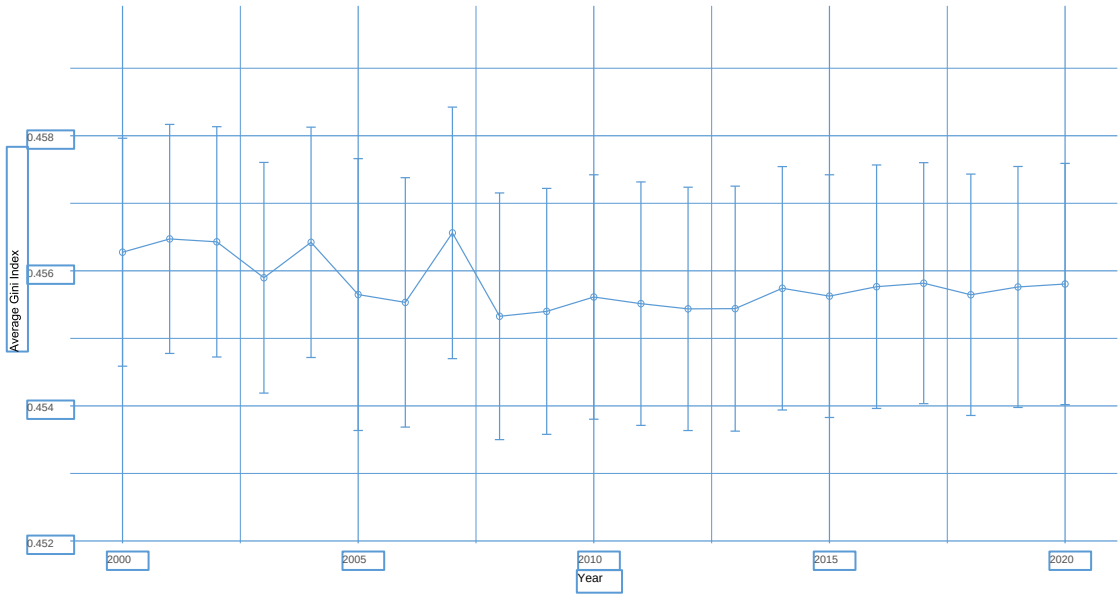
Next, the behaviour of the synthetic Gini index produced by the Random Forest at the Department level and its evolution over time is shown in Figure 4 and Figure 5. The error bars of the plot imply that there is no significant intra-Department level variation. Which is consistent from the most important variables used in the Random Forest which are in essence "fix". The Department with the highest synthetic income inequality is Chocó with almost a Gini of 0.54 meanwhile the lower income inequality is reflected in Bogota with a Gini closer to 0.43. This is no surprise as Chocó has been historically characterized for the lack of opportunities, massive income concentration, and distance to the capital. On the other hand, Bogotá becomes the most equal place in the estimations with the fixed factors. The peripherical areas including Cauca, Amazonas, Vaupes, Guainía, Guaviare and Caquetá are also according to the pattern of high-income inequality. The yearly evolution of the synthetic income inequality produced by the Random Forest does not have significant changes. In fact, the synthetic inequality seems to round a consistent 0.457 of the Gini and a 0.454 reflecting a stagnating income inequality average for each year. At the best case, it has decreased just 0.002 of the Gini in the lapsus of 20 years.

Figure 4. Synthetic Gini Index at the Department level (Dominant Fixed Effects)



Source: Own elaboration

Figure 5. Synthetic Gini Index annual evolution (Dominant Fixed Effects)



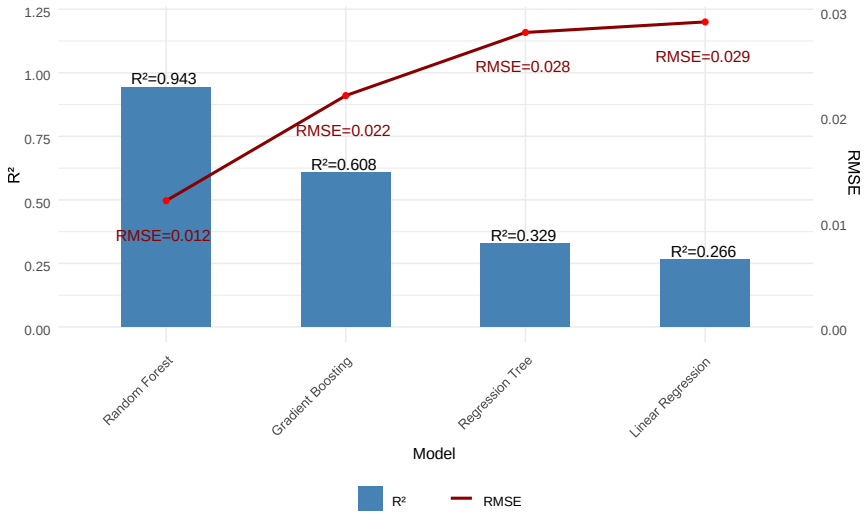
Source: Own elaboration.

2.4. Estimation with Time-Varying Factors

Considering the previous interpretation, and particularly the pattern displayed in Figure 5, one could ask how the synthetic Gini coefficient might change by leaving in the Random Forest a set of pure varying regressors. It is easy to identify the fixed and "semi-fixed"[5] regressors from Figure 2.

The Random Forest model was re-estimated while excluding the fixed and "semi-fixed" regressors[6]. Following this adjustment, the performance of the machine learning models is presented in Figure 6, where the Random Forest once again demonstrates superior performance compared to gradient boosting, regression decision trees, and linear regression. The evaluation metrics for the Random Forest indicate an R^2 of 94.3 and an RMSE of 0.012.

Figure 6. Model estimation performance with varying regressors



Source: Own elaboration

While the first model "Dominant in fixed effects" estimates with a precision of R^2 =98.67%, the model with "Dominant in varying factors" estimates the precision in R^2 = 94.3%. Surprisingly, the loss is about 4.37%, which is interpreted as a non-significant decrease. Now, what kind of regressors dominate the varying factors models? Let's inspect this in the next Figure 7, where the importance of variables is presented.

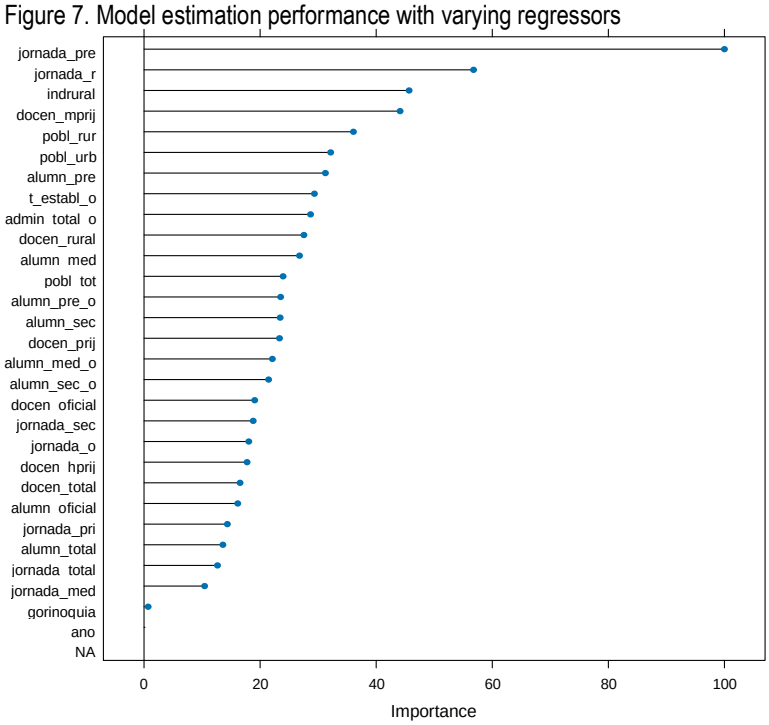
The set of factors which drive the Random Forest are now all related to socioeconomic characteristics, in particular, human capital accumulation is one of the major drivers along with population dynamics. The initial stage of human capital accumulation also displays an important role.

Regarding the human capital accumulation variables, kindergarten class schedules is the most important variable for this Regression Forest, followed by the rural class schedules. The third important variable is related to the rural index, followed by the number of female teachers in kindergarten. The rural and urban population are next in the importance. Then the number of students in kindergarten, the number of educational facilities, the number of administrative personal in the educational facilities, and the number of teachers in rural areas, followed by the number of middle school students. Total population also enters here with more human capital indicators related to total count of teachers, total school students, total number of students in public institutions, and total number of dedicated class schedules to middle education.

[5] The term semi-fixed is used here to refer to those continuous variables that do not vary significantly. For example, the linear distance to the capital, this variable while it is continuous in nature, it is not changing overtime, constituting a fixed effect.

[6] Specifically the next set of fixed affects are deleted: "codmpio", "codprovincia", "codmdo", "coddepto", "gpacifica", "disbogota", "altura", "dismdo", "discapital", "gcaribe", "gamazonia", "gandina","ao cre", "ao crea", "distancia mercado", "mercado cercano" where in particular the last fourth variables are in essence the semi-fixed effects as these refer to the year of creation of the municipality, distance to closest food market (which is repeated with mercado cercano), and hence I let just one measure of distance to food markets.

Figure 7. Model estimation performance with varying regressors

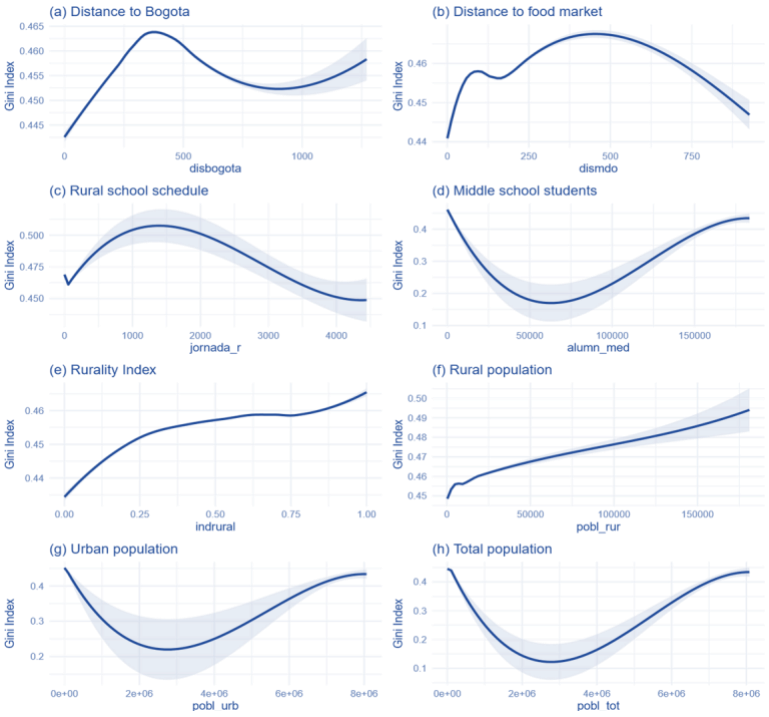


Source: Own elaboration

2.5. Machine Learning Estimation with Time-Varying Factors

Considering that the model with dominant fixed effects might potentially drive differential results in terms of pure "memory", this section will present the behaviour of the Synthetic Gini Coefficient with the model of varying factors. To have some comparisons, the nonlinear relationships are presented with the same variables as the first models. This is to inspect some potential robustness of the results and the patterns. Figure 8 shows the relative patterns of the Synthetic Gini Coefficient with the regressors of the model with dominant fixed effects.

Figure 8. Dispersion of Synthetic Gini Index and features - varying factors

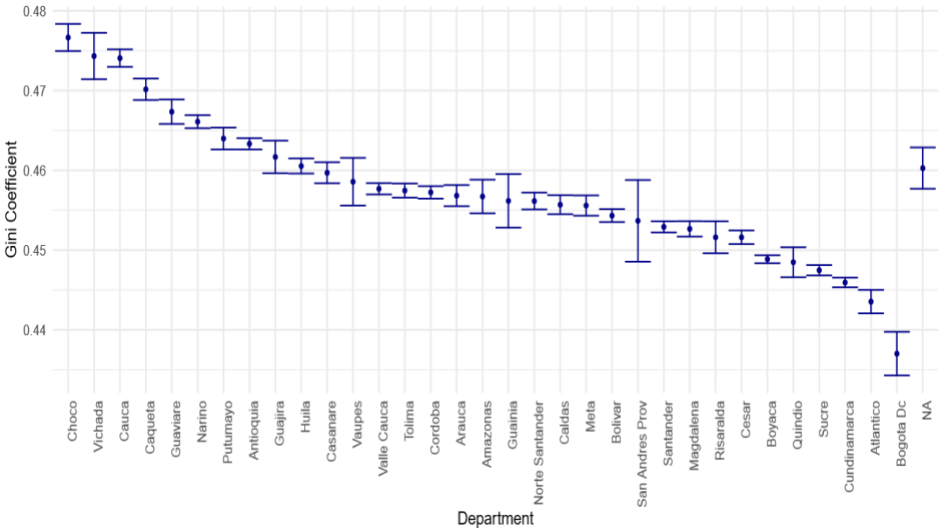


Source: Own elaboration

At first glance, it is surprising that the patterns are virtually the same, however, the magnitude of the estimates differs slightly. The same pattern described in the previous section exists as shown in Figure 8. However, the magnitude is what might significantly change according to the values on the y-axis. This behaviour cannot be seen easily in the previous plots.

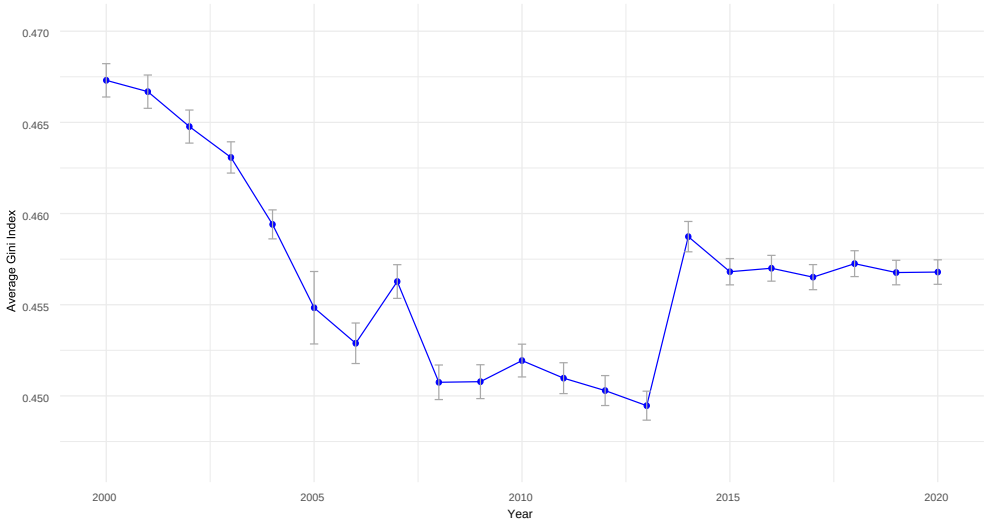
The next Figure 9 and Figure 10 present the variation across Departments and the evolution over time of the Synthetic Gini Index under the dominant factor varying model. As expected, in contrast to the dominant fixed effects model, the factor varying model presents a higher volatility across Departments and across time. This is because the drivers are "stiffer" in the first model when it comes to synthetize the Gini coefficient across time. On the other hand, more variability exists when relying on the varying regressors. However, some consistent findings do exist. Chocó is in fact the most unequal Department of Colombia again, but now instead of Bogotá, Quindío seems to be best in terms of synthetic inequality. Across time, differential dynamics are witnessed, first the starting inequality is about 0.469 but it reduces to 0.459 with some significant increases in the year 2006 and 2013. Whereas in the dominant fixed effects model the starting inequality started in 0.4567 and decreased to 0.4557. Hence, confirms the "stiffness" of the dominant fixed effects model.

Figure 9. Synthetic Gini Index at the department level (varying factors)



Source: Own elaboration

Figure 10. Synthetic Gini Index annual evolution



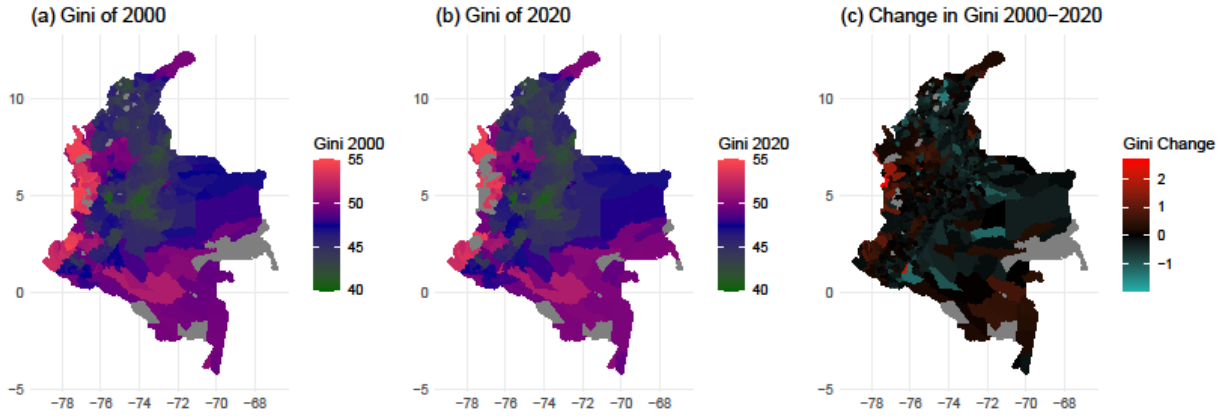
Source: Own elaboration

2.6. Geographical Analysis

Considering both imputations, the Synthetic Gini Coefficient presents a relative asymmetric evolution from the Dominant Fixed Effects model and the Varying Factor model. To present these asymmetries, the Figure 11 and Figure 12 show the geographical distribution of the Synthetic Gini coefficient for both Dominant Fixed Effects model and the Varying Factor model.

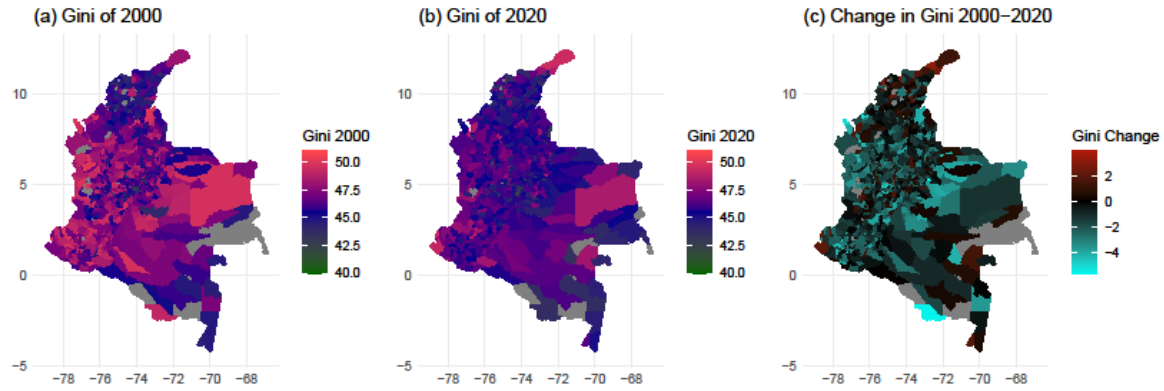
Important aspects to remark on are that the Dominant Fixed Effects model does present a conservative evolution of the income inequality, with a barely noticeable improvement. Meanwhile, the Varying Factor model does depict some improvements in Colombia. The figures contain a set of three panels, panel (a) that shows the Gini Synthetic estimate for the year 2000, panel (b) which do the same for the year 2020, and panel (c) which shows the absolute change. The black areas of the last panel represent a null change in the lapsus of 20 years, the red areas show a worsening situation where inequality increased, and the light blue areas represent a potential improvement in income inequality. This distributional framework concludes that the Varying Factor model does present the most optimistic evolution of income inequality across Colombia, as panel (c) presents lighter zones in contrast to the Dominant Fixed Effects model.

Figure 11. Synthetic Gini Index Dominant Fixed Effects



Source: Own elaboration

Figure 12. Synthetic Gini Index Varying Factor model



Source: Own elaboration.

A diverging result is found for the areas where income inequality increased. For example, in the conservative estimates of the Dominant Fixed effects model, the west of Colombia, close to the pacific region has deteriorated in terms of their income inequality. In contrast, the Varying Factor model shows that the northern part of Colombia has worsened its income inequality condition.

2.7. Descriptive Statistics

This section only presents the tables of descriptive statistics as the final part of the study.

Table 1. Summary statistics for Gini dominant fixed effects model

Year	N	Min	Max	Mean	Median	Std_Dev
2000	1091	40.43	54.84	45.62772	44.870	2.844194
2001	1102	40.44	54.61	45.64716	44.870	2.871935
2002	1107	40.41	54.73	45.64290	44.830	2.894591
2003	1104	40.41	54.46	45.58965	44.810	2.894910
2004	1109	40.38	55.27	45.64208	44.850	2.897762
2005	1095	39.43	56.81	45.56487	44.860	3.397956
2006	1099	40.13	55.96	45.55332	44.730	3.123218
2007	1044	40.22	55.38	45.65627	44.875	3.069900
2008	1104	40.17	55.77	45.53265	44.690	3.095974
2009	1111	40.12	55.63	45.53998	44.690	3.096774
2010	1114	40.18	55.50	45.56136	44.705	3.081708
2011	1115	40.19	55.62	45.55137	44.690	3.070800
2012	1116	40.17	55.43	45.54370	44.710	3.070403
2013	1107	40.15	55.42	45.54397	44.680	3.079029
2014	1089	40.30	55.67	45.57413	44.670	3.034139
2015	1085	40.31	55.66	45.56259	44.670	3.019958
2016	1084	40.26	55.72	45.57653	44.670	3.028247
2017	1104	40.31	55.69	45.58164	44.655	3.024949
2018	1095	40.27	55.34	45.56455	44.670	3.017022
2019	1103	40.25	55.46	45.57606	44.660	3.023271
2020	1104	40.24	55.64	45.58049	44.670	3.027407
Total	23082	39.43	56.81	45.58135	44.750	3.032628

Source: Own elaboration

Table 2. Summary statistics for Gini varying factors model

Year	N	Min	Max	Mean	Median	Std_Dev
2000	1091	42.33	50.21	46.73063	46.670	1.540815
2001	1102	42.19	50.19	46.66831	46.550	1.545259
2002	1107	42.14	50.33	46.47703	46.320	1.539203
2003	1104	42.34	50.19	46.30774	46.305	1.453158
2004	1109	42.21	49.72	45.94070	45.750	1.349514
2005	1095	39.43	56.81	45.48358	44.860	3.353495
2006	1099	41.22	51.82	45.28877	44.990	1.880035
2007	1044	41.16	51.74	45.62739	45.640	1.523703
2008	1104	41.33	51.96	45.07485	44.910	1.606435
2009	1111	41.25	51.04	45.07825	45.000	1.579002
2010	1114	41.29	51.07	45.19388	45.115	1.530773
2011	1115	41.22	50.49	45.09755	44.950	1.440838
2012	1116	41.32	50.63	45.02939	44.930	1.399990
2013	1107	41.38	50.87	44.94650	44.870	1.349895
2014	1089	41.98	50.81	45.87338	45.790	1.401379
2015	1085	41.76	50.23	45.68106	45.680	1.212002
2016	1084	41.96	50.04	45.70012	45.730	1.188298
2017	1104	42.03	50.30	45.65135	45.690	1.165503
2018	1095	41.81	49.49	45.72513	45.740	1.197512
2019	1103	41.96	49.44	45.67634	45.730	1.141141
2020	1104	41.82	49.54	45.67902	45.730	1.137567
Total	23082	39.43	56.81	45.66182	45.590	1.651774

Source: Own elaboration

Conclusion

This study is an effort to contribute to the missing data problem of income inequality in Colombia by synthesizing two datasets estimating the Gini coefficient at the municipality level between the years 2000-2020, this is done through machine learning techniques. This resulted in two models estimated through the technique of Random Forest, which was the best model across the estimations, and outperformed in comparison the gradient boosting, the regression trees, and the linear regression approaches. The two models using the Random Forest are described as the following: 1) A Dominant Fixed Effects model denominated like this given the nature of the importance of the predictors/features that compose it, which is mainly based on a set of fixed or semi-fixed variables. 2) A Dominant Varying Factor model as an alternative to the Dominant Fixed Effects models, estimated mainly with time-varying predictors/features. Both models perform in their metrics quite well, with an R^2 = 98.67% for the Dominant Fixed Effects model and an R^2 = 94.3% for the Dominant Varying Factor model. Hence, the estimation of these models to the rest of panel is denominated as the Synthetic Gini Coefficient/Index.

The Dominant Fixed Effects model presents the Synthetic Gini Coefficient with minimum/conservative/stiff changes, where inequality in average has decreased from 45.63 in 2000 to 45.58 in 2020. In contrast the Dominant in Varying Factors model presents more variability over time, starting with a mean inequality of 46.73 which has reduced to 45.67. All the estimates, reflect that Colombia has not improved their income inequality condition. Also, provides evidence of existing heterogeneities that are present in the municipalities of Colombia. For both estimations, the Department of Chocó is the most unequal of the Colombian territory. The fact that Colombia has not improved greatly the situation of income inequality over this period, raises questions about the level of welfare improvement in the country[7].

In the Dominant in Varying Factors models, the composition of predictors/features used to estimate the Synthetic Gini index consists of two essential categories: human capital accumulation (including both supply and demand variables) and population dynamics. The nonlinear relationships between the Synthetic Gini index and predictor information confirm several patterns observed in the economic literature.

Credit Authorship Contribution Statement:

The author performed all tasks involved in manuscript preparation, research, and writing.

Acknowledgments

Special thanks to the civil organization Veeduria Estudios y Evaluación de la Gestión Pública Colombiana -EEGPC- to raise the interest in this topic. Moreover, special thanks to the Corporación Centro de Interés Público -CIPJUS- to allow the research over open topics in economics and public policy. This research was not funded or financially related. I declare there is no financial or other substantive conflict of interest that could be seen to influence your results or interpretations.

Conflict of Interest Statement

The author declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

[7] This due income inequality has been negative correlated with the levels of welfare among economies, see Dagum (1990), Clark & Kavanagh (1996), Abdel-Rahman & Wang (1997), Coburn (2015), Kim (2017), Riveros-Gavilanes (2021), Wildowicz-Szumarska (2022), Coady et al. (2022), Riveros-Gavilanes et al. (2022), Yang & Tang (2023), and Sologon et al. (2023).

References

Abdel-Rahman, H. M. & Wang, P. (1997). Social welfare and income inequality in a system of cities. *Journal of Urban Economics*, 41(3), 462–483. https://doi.org/10.1006/juec.1996.2013

Alwateer, M., Atlam, E.-S., Abd El-Raouf, M. M., Ghoneim, O. A., & Gad, I. (2024). Missing data imputation: A comprehensive review. *Journal of Computer and Communications*, 12(11), 53–75. https://doi.org/10.4236/jcc.2024.1211004

Caravaggio, N., Resce, G., & Vaquero-Piñeiro, C. (2025). Predicting policy funding allocation with Machine Learning. *Socio-Economic Planning Sciences*, 98, 102175. https://doi.org/10.1016/j.seps.2025.102175

Castelló-Climent, A. & Doménech, R. (2021). Human capital and income inequality revisited. *Education Economics*, 29(2), 194–212. https://doi.org/10.1080/09645292.2020.1870936

CEDE (2023). Panel municipal Centro de Estudios sobre el Desarrollo Económico. https://datoscede.uniandes.edu.co/catalogo-de-datos/

Clark, C. M. & Kavanagh, C. (1996). Basic income, inequality, and unemployment: rethinking the linkage between work and welfare. *Journal of Economic Issues*, 30(2), 399– 406. https://doi.org/10.1080/00213624.1996.11505803

Coady, D., D'Angelo, D., & Evans, B. (2022). Fiscal redistribution, social welfare and income inequality: 'doing more' or 'more to do'? *Applied Economics*, 54(21), 2416–2429. https://doi.org/10.1080/00036846.2021.1990840

Coburn, D. (2015). Income inequality, welfare, class and health: A comment on Pickett and Wilkinson. *Social Science & Medicine*, 146, 228–232. https://doi.org/10.1016/j.socscimed.2015.09.002

Combes, P.-P., Gobillon, L., & Zylberberg, Y. (2022). Urban economics in a historical perspective: Recovering data with machine learning. *Regional Science and Urban Economics*, 94, 103711. https://doi.org/10.1016/j.regsciurbeco.2021.103711

Dagum, C. (1990). On the relationship between income inequality measures and social welfare functions. *Journal of Econometrics*, 43(1-2), 91–102. https://doi.org/10.1016/0304-4076(90)90109-7

Gao, Q.-L., Zhong, C., Yue, Y., Cao, R., & Zhang, B. (2024). Income estimation based on human mobility patterns and machine learning models. *Applied Geography*, 163, 103179. https://doi.org/10.1016/j.apgeog.2023.103179

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gond, V. K., Dubey, A., & Rasool, A. (2021). A survey of machine learning-based approaches for missing value imputation. In *2021 3rd International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1–8. IEEE. https://doi.org/10.1109/ICIRCA51532.2021.9544957

Hong, S. & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, 20:1–12. https://doi.org/10.1186/s12874-020-01080-1

Kim, K.-t. (2017). The relationships between income inequality, welfare regimes and aggregate health: A systematic review. *The European Journal of Public Health*, 27(3), 397–404. https://doi.org/10.1093/eurpub/ckx055

Kuhn, M. (2008). Building predictive models in *r* using the caret package. *Journal of Statistical Software*, 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Kühn, M. (2015). Peripheralization: Theoretical concepts explaining socio-spatial inequalities. *European Planning Studies*, 23(2), 367–378. https://doi.org/10.1080/09654313.2013.862518

Lakshminarayan, K., Harp, S. A., Goldman, R. P., Samad, T., et al. (1996). Imputation of missing data using machine learning techniques. In *KDD*, Volume 96. https://cdn.aaai.org/KDD/1996/KDD96-023.pdf

Lee, J.-W. & Lee, H. (2018). Human capital and income inequality. *Journal of the Asia Pacific Economy*, 23(4), 554–583. https://doi.org/10.1080/13547860.2018.1515002

Lee, K.-K. & Vu, T. V. (2020). Economic complexity, human capital and income inequality: A cross-country analysis. *The Japanese Economic Review*, 71(4), 695–718. https://doi.org/10.1007/s42973-019-00026-7

Lin, W.-C. & Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53, 1487–1509. https://doi.org/10.1007/s10462-019-09709-4

Lin, W.-C., Tsai, C.-F., & Zhong, J. R. (2022). Deep learning for missing value imputation of continuous data and the effect of data discretization. *Knowledge-Based Systems*, 239, 108079. https://doi.org/10.1016/j.knosys.2021.108079

Ma, X., Hao, Y., Li, X., Liu, J., & Qi, J. (2023). Evaluating global intelligence innovation: An index based on machine learning methods. *Technological Forecasting and Social Change,* 194, 122736. https://doi.org/10.1016/j.techfore.2023.122736

Oppido, S., Ragozino, S., & Esposito De Vita, G. (2023). Peripheral, marginal, or noncore areas? setting the context to deal with territorial inequalities through a systematic literature review. *Sustainability*, 15(13), 10401. https://doi.org/10.3390/su151310401

Paas, T. & Schlitte, F. (2008). Regional income inequality and convergence processes in the EU-25. *Scienze regionali: Italian Journal of Regional Science: 7, Supplemento 2, 2008*, 29-49. https://www.francoangeli.it/riviste/articolo/33743

Rácz, A., & Gere, A. (2025). Comparison of missing value imputation tools for machine learning models based on product development cases studies. LWT, 117585. https://doi.org/10.1016/j.lwt.2025.117585

Rey, S. J. (2004). Spatial analysis of regional income inequality. *Spatially Integrated Social Science*, 1, 280–299. https://doi.org/10.1093/oso/9780195152708.003.0014

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the GBM package. *Update*, 1(1). https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf

Riveros-Gavilanes, J. M. (2021). Estimation of Amartya Sen's social welfare function for Latin America. *Ensayos de Economía*, 31(59), 13-40. https://doi.org/10.15446/ede.v31n59.88235

Riveros-Gavilanes, J. M. (2023). On the empirics of violence, inequality, and income. *Journal of Economics and Management*, 45(1), 102–136. https://doi.org/10.22367/jem.2023.45.06

Riveros-Gavilanes, J. M., Al Akayleh, F., Oduniyi, O., Samuel, A. H., & Hassan, S. M. (2022). On the welfare trends: A view from the Sen's social welfare function. Technical Report, M & S Research Hub institute. https://ideas.repec.org/p/ris/msrwps/2022_003.html

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons. ISBN 0-471-08705-X, https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316696.fmatter

Salvati, L. (2016). The dark side of the crisis: disparities in per capita income (2000–12) and the urban-rural gradient in Greece. *Tijdschrift voor economische en sociale geografie*, 107(5), 628–641. https://doi.org/10.1111/tesg.12203

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. https://pubmed.ncbi.nlm.nih.gov/12090408/

Seu, K., Kang, M.-S., and Lee, H. (2022). An intelligent missing data imputation technique: A review. *International Journal on Informatics Visualization*, 6(1-2), 278– 283. http://dx.doi.org/10.30630/joiv.6.1-2.935

Silva, T. C., Wilhelm, P. V. B., & Amancio, D. R. (2024). Machine learning and economic forecasting: The role of international trade networks. *Physica A: Statistical Mechanics and Its Applications*, 649, 129977. https://doi.org/10.1016/j.physa.2024.129977

Sologon, D. M., Doorley, K., & O'Donoghue, C. (2023). Drivers of income inequality: what can we learn using microsimulation? *Handbook of Labor, Human Resources and Population Economics*, 1–37. https://doi.org/10.1007/978-3-319-57365-6_392-1

Sullivan, T. R., Lee, K. J., Ryan, P., & Salter, A. B. (2017). Multiple imputation for handling missing outcome data when estimating the relative risk. *BMC Medical Research Methodology*, 17, 1-10. https://doi.org/10.1186/s12874-017-0414-5

Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, 227, 120201. https://doi.org/10.1016/j.eswa.2023.120201

Teng, W., Mamman, S. O., Xiao, C., & Abbas, S. (2024). Impact of natural resources on income equality in Gulf Cooperation Council: Evidence from machine learning approach. *Resources Policy*, 88, 104427. https://doi.org/10.1016/j.resourpol.2023.104427

Therneau, T., Atkinson, B, & Ripley, B. (2015). Package 'rpart'. https://cran.r-project.org/web/packages/rpart/rpart.pdf

Wang, S., Li, B., Yang, M., & Yan, Z. (2019). Missing data imputation for machine learning. In *IoT as a Service: 4th EAI International Conference, IoTaaS 2018, Xi'an, China, November 17–18, 2018, Proceedings 4*, 67–72. Springer. http://dx.doi.org/10.1007/978-3-030-14657-3_7

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), 180–185. http://dx.doi.org/10.1002/wics.147

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidy verse. *Journal of Open-Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wildowicz-Szumarska, A. (2022). Is redistributive policy of EU welfare state effective in tackling income inequality? A panel data analysis. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, 17(1), 81–101. https://doi.org/10.24136/eq.2022.004

Xue, J. (2023). Review on data imputation methods in machine learning. *Journal of Physics: Conference Series*, Volume 2646, 012034. IOP Publishing. https://doi.org/10.1088/1742-6596/2646/1/012034

Yang, X. & Tang, W. (2023). Additional social welfare of environmental regulation: The effect of environmental taxes on income inequality. *Journal of Environmental Management*, 330, 117095. https://doi.org/10.1016/j.jenvman.2022.117095

Yarberry, W. (2021). *CRAN Recipes: DPLYR, Stringr, Lubridate, and RegEx in R*, pages 1–58. Apress Berkeley, CA, ISBN: 978-1-4842-6875-9, eBook ISBN: 978-1-4842-6876-6. https://doi.org/10.1007/978-1-4842-6876-6

Zhan, C., Liu, Y., Wu, Z., Zhao, M., & Chow, T. W. S. (2023). A hybrid machine learning framework for forecasting house price. *Expert Systems with Applications,* 233, 120981. https://doi.org/10.1016/j.eswa.2023.120981

Zhu, J., & Huang, T. (2024). Public debt and welfare with machine learning. *Finance Research Letters*, 69, 106164. https://doi.org/10.1016/j.frl.2024.106164

Cite this article

Riveros-Gavilanes, J. M. (2025). A machine learning approach to synthetic Gini Coefficient estimation in Colombian municipalities. *Journal of Research, Innovation and Technologies*, Volume IV, 1(7), 7-24. https://doi.org/10.57017/jorit.v4.1(7).01

Article's history:

Received 14th of February, 2025; Revised 1st of March, 2025; Accepted for publication 20th of March, 2025;
Available online: 24th of March, 2025 Published as article in Volume IV, Issue 1(7), 2025