Municipality synthetic Gini index for Colombia: A machine learning approach

John Michael, Riveros-Gavilanes

Veeduria Estudios y Evaluacion de la Gestion Publica Colombiana, Corporacion Centro de Interés Público y Justicia - CIPJUS

1 February 2025

Veeduria Estudios y Evaluación de la Gestión Pública Colombiana - EEGPC
Corporacion Centro de Interés Público y Justicia - CIPJUS
Working Paper No. 001, January 2025, pp. 1-30

Municipality Synthetic Gini Index for Colombia: A Machine Learning Approach

Índice Sintetico Municipal de Gini para Colombia: Un enfoque de Machine Learning

John Michael Riveros-Gavilanes[1]

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 were 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

Abstract

Este documento presenta dos estimaciones sinteticas del coeficiente de Gini a nivel municipal en Colombia entre los años 2000-2020. La metodología utiliza varios modelos de machine learning para seleccionar el mejor modelo para la imputación de datos. Esto deriva en dos modelos de Random Forest, el cual, el primero es characterizado por ser Dominante en Efectos Fijos, mientras el segundo tiene un conjunto de variables Dominantes en Factores Variantes. Con estas estimaciones, el Índice Sintetico de Gini para los modelos es revisado, y links publicos son generados para su acceso. El modelo Dominante de Efectos Fijos es "rigido" en contraste con el modelo de Dominante en Factores Variantes. Se recomienda a los investigadores usar el Índice Sintético con factores variantes por que contienen mayor variabilidad a través del tiempo.

Palabras Clave: Gini, Machine learning, Random forest, estimación, sintetico, economía

1. M. Sc. Economics. Ludwig-Maximilians-Universität - LMU München, Deutschland
E-mail: {riveros}@ms-researchhub.com

Contents

1. 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 identificators 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 strong related to municipality clusters, distances, and geographical locations. By interpreting these features related mostly to fixed effects, -as they do not fluctuate significant 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. The second model remarks the importance of 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.

The details of this estimation goes as follow; 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, this is described by a strong composition of "fixed" and "semi-fixed" variables which 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=drive_link&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=drive_link&ouid=118158209086311183140&rtpof=true&sd=true

Considering that the Varying Factor model contains more variability across the municipalities and the 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 of them 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 and 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 limitations of this study is that in fact, 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 and Graham, 2002). For these reasons and the potential biases that exists 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 exists. 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 presents 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 2 that describes the estimation process, Section 2.1 describes the data transformation used to preserved 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 analyzed. In this interpretation, the nonlinear correlations of some important variables are analyzed 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 analyzed 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.

2. General process

Since the CEDE (2023) is one of the sources of information that consolidates the municipality data of Colombia for a significant amount of years , I selected this data to be core of the empirical estimations. That said, CEDE data compose 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 of these topics 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 behavior of income inequality across the country.

2.1 Data transformations

The first step was to identify within the seven panels of information which contained the highest number of observations available. I identified that panel 2) Good Government contained the most information available (N=43541), so I kept this panel as the central initial data for the merging process.

The second step was to select the join keys for the grand merge between panels. As is well known, the municipality codes (called DIVIPOLA) and years are the proper for the subsequent task to estimate variables at the municipality level. The result from this merge 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 23563 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 2654 variables and 23563 obs.

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 [1]. The resulting panel at this point contained 48 variables and 23563 obs.

The fifth step was to convert all character variables to numerical variables, this let some empty but not missing variables again [2]. In this point, a final condition is applied to retain only complete cases for the panel. The resulting panel then contained 44 variables and 23082 obs.

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 to the cleaned panel without missing values. For where the panel for estimation techniques is going to be executed. The panel for this processes contained 45 variables and 23082 obs.

2.2 Machine learning estimation

The software and programs used to estimate the different machine learning models and their graphics are developed by Ridgeway and Ridgeway (2004), Kuhn (2008), Wickham (2011), Therneau et al. (2015), Wickham et al. (2019), Yarberry and 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 this 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 the panels, a cleaning process is done to let one type of cluster ID variable for geographical 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 in figure 1 which contains the classic measures of R^2 and $RMSE$ for continuous outcomes.

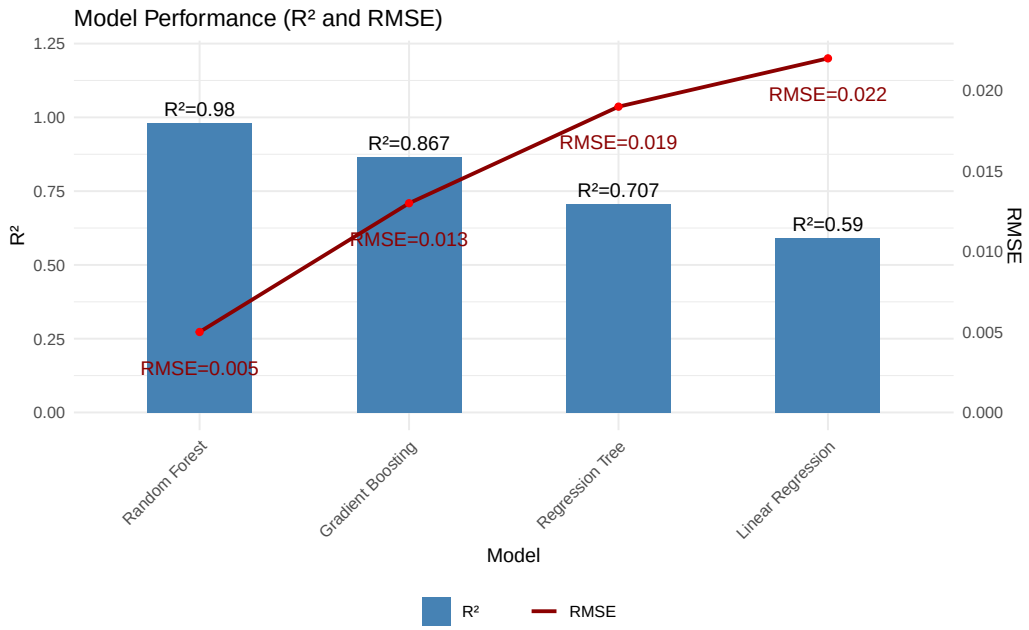
These estimations suggest that random forest approach outperforms the rest of the models. In particular, it exhibits an $R^2 = 98.67\%$ with an $RMSE = 0.005$ suggesting a precise result even after cross-validation with five folds.

With this best model (the random forest), the Gini data is imputed for the whole panel without missing values ($K = 45, N = 23082$) including the target variable. The panel hence contains information of about 1091 municipalities for the years 2000 and 2020.

An additional important result is the importance of the variables for the random forest estimation. Which are displayed in figure 2.

1. 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 elimiated
2. The variables of "depto, provincia, municipio" and "act adm" which were incomplete identificators of the municipalities, not important as the code for each of thems are in other variables

Figure 1: Model estimation performance



Source: Own elaboration

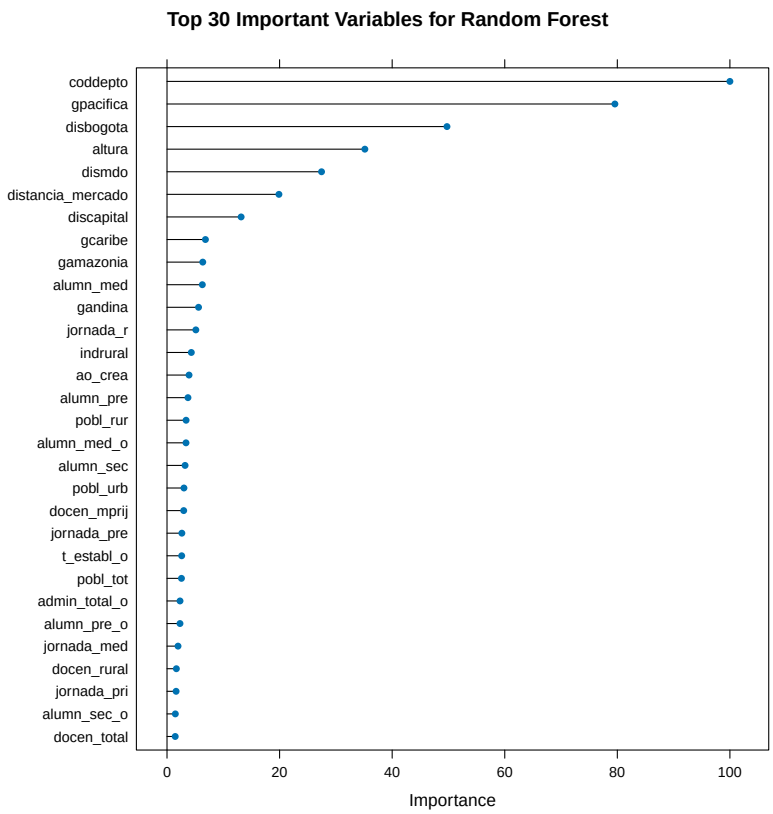
2.3 Interpretation of variables within the random forest

By looking at figure 2 it is possible to identify a pattern in scale of importance for the random forest estimation. In particular the starting variable "coddepto" which refers to the geographical cluster of the administrative divisions of Departaments 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 altitute, 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 Departament followed by other fixed location variables such as the Andina, Amazonia, and Caribbean regions. This first set of variables are have one thing in common. They are variables related to the geographical physical locations of the municipality.

On the second set of important variables it can be found socioeconomic characteristics involving population dynamics, school and educational topics. This set the variables involve 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.

I interpret the selection of the fixed or "semi" fixed characteristics such as the geographical clusters and the distances as a form of "fixed-effects" from the econometric perspective. In general, there are some existing time-invariant characteristics which are

Figure 2: Importance of features



Source: Own elaboration

able to 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 the capital of the departments. Market's interconnections play a significant role here as well.

According to figure 2 these fixed effects are the main drivers of the inequality across the Colombian municipalities. The question of what and how will change the estimates will be adressed in the next section. For now, using this data it is interesting to see how income inequality might related to the continous variables used in the random forest estimations. This to understand a bit better the blackbox 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 continous regressors (or features) used in the Random Forest model. Some clear patters emerge from this dispersions: a) The larger the distance of the municipality relative to the capital of the country (Bogota) implies a larger

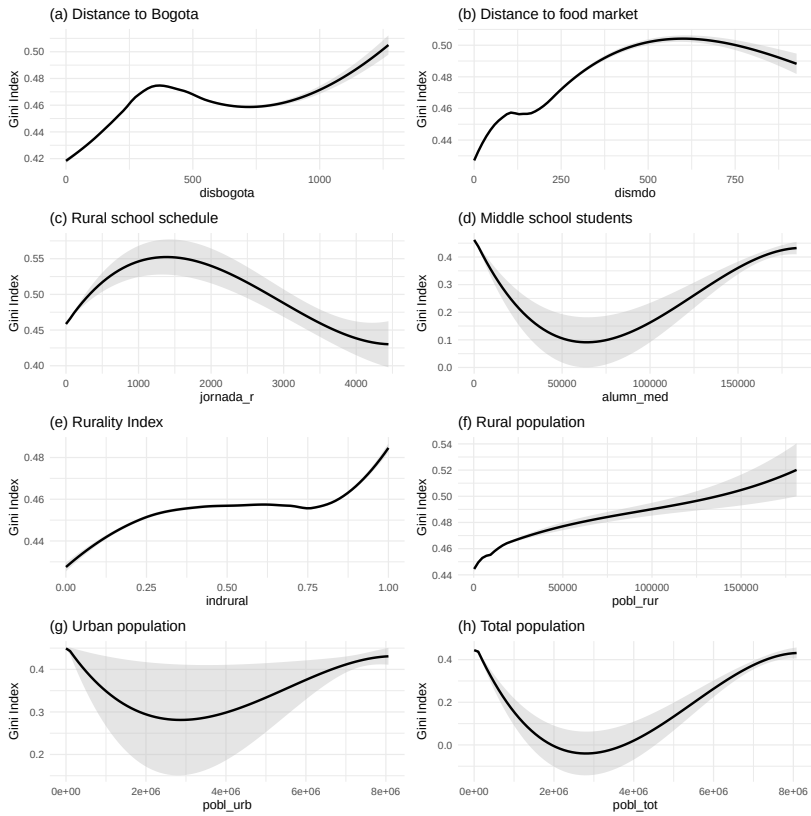
income inequality. b) The larger the distance of the municipality relative to the largest food market nearby implies a larger income inequality. c) An increasing size of the class schedules in rural areas tends to drive a higher inequality until a turning point where class schedules are large enough 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, howevever, at somepoint, 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 rural population and the 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 behavior of the urban population, hence when massification arises, inequality grows again.

These patterns are consistent with the economic literature[3], where for example, the peripheral regions or zones are prone to suffer from income inequality given the lack of market interconexion, 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 also the linear distance to the central food markets. According to this argument the population dynamics such as rural concetration, rural population, seems 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 analyze. 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. In particular, when school schedules satisfy the demand of schooling in the rural territories, it is most likely to encounter a reduction in the synthetic inequality. On the otherhand, from the side of the demand as seen in panel (d). When the number of middle school students increase enough, students will face a constraint in the access of education, delivering an increase in the 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.

Next, the behavior of the synthetic Gini index produced by the Random Forest at the Deparment level and its evolution over the time is shown in figures 4 and 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 suprise as Chocó has been historically characterized for the lack of opportunities, massive income inconcentration, and distance to the capital. On the otherhand, Bogotá becomes the most unequal place in the estimations with the fixed factors. The peripherical areas including Cauca, Amazonas, Vaupes, Guainia, 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

3. In particular, 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 and Doménech (2021), Lee and Vu (2020) relative to these observe variables and income inequality.

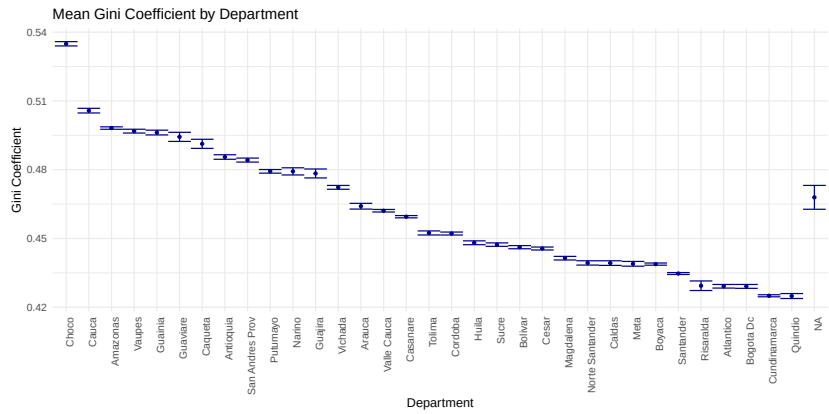
Figure 3: Dispersion of Synthetic Gini Index and features



Source: Own elaboration

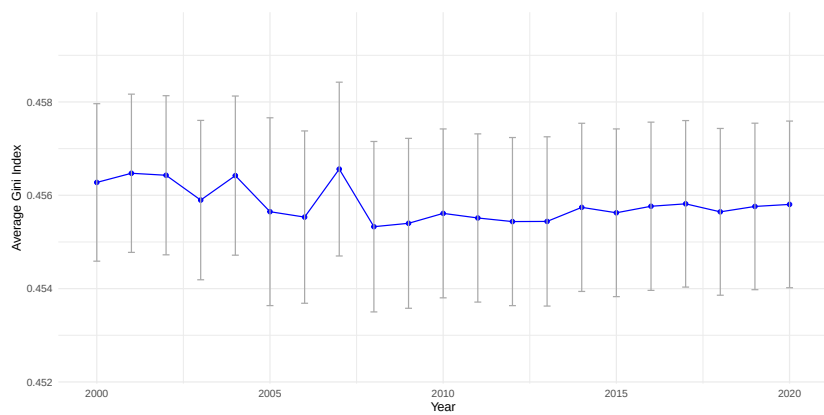
0.457 of the Gini and a 0.454 reflecting a stagnating income inequality average for each year. At the best case, it has decreased just a 0.002 of the Gini in the lapsus of 20 years.

Figure 4: Synthetic Gini Index at the Department level



Source: Own elaboration

Figure 5: Synthetic Gini Index annual evolution



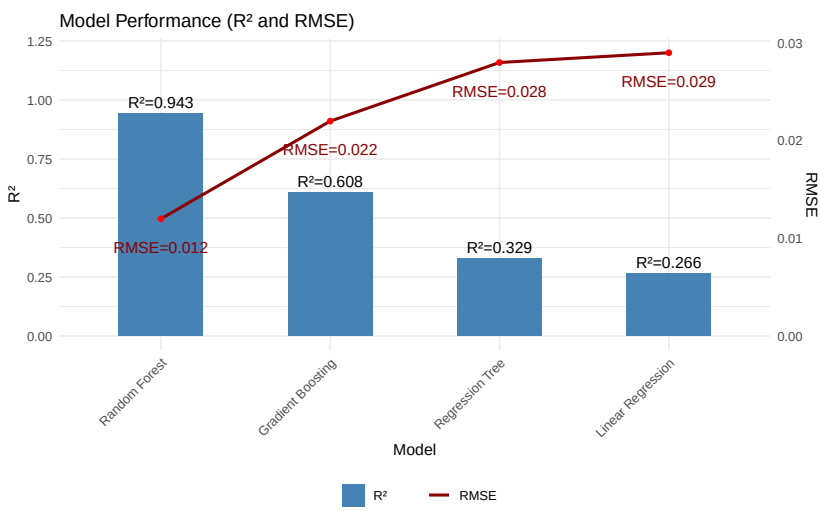
Source: Own elaboration

2.4 Estimation with time-varying factors

Considering the previous interpretation, and in particular 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. In particular, it is easy to identify the fixed and "semi-fixed"[4] regressors from figure 2.

With this idea I re-estimate the Random Forest model but excluding the fixed and "semi-fixed" regressors[5]. Upon this process, the performance of the machine learning models is presented in figure 6. Where it is visible that the Random Forest once again outperforms the gradient boosting, the regression decision trees and the linear regression. The metrics for the Random Forest is an $R^2 = 94.3$ and a $RMSE = 0.012$.

Figure 6: Model performance with varying regressors



Source: Own elaboration

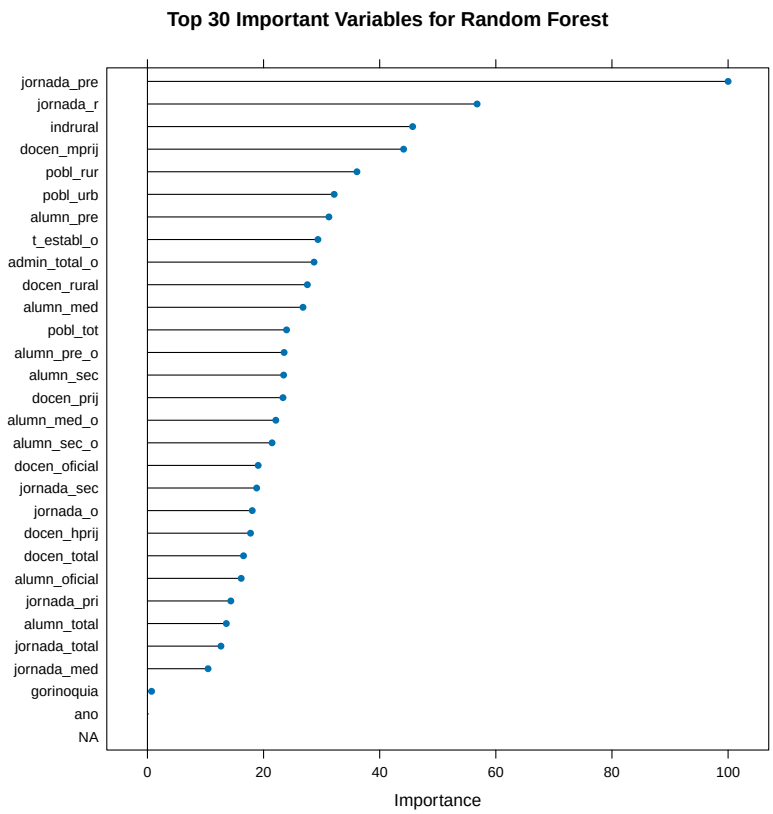
While the first model "Dominant in fixed effects" estimates with a precision of $R^2 = 98.67\%$, the model with "Dominant in varying factors" estimates the precision in $R^2 = 94.3\%$, Suprisingly, the loss is about 4.37% which is interpreted as a non-significant decrease. Now, what kind of regressors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

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

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

5. 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: Importance of features with varying factors



Source: Own elaboration

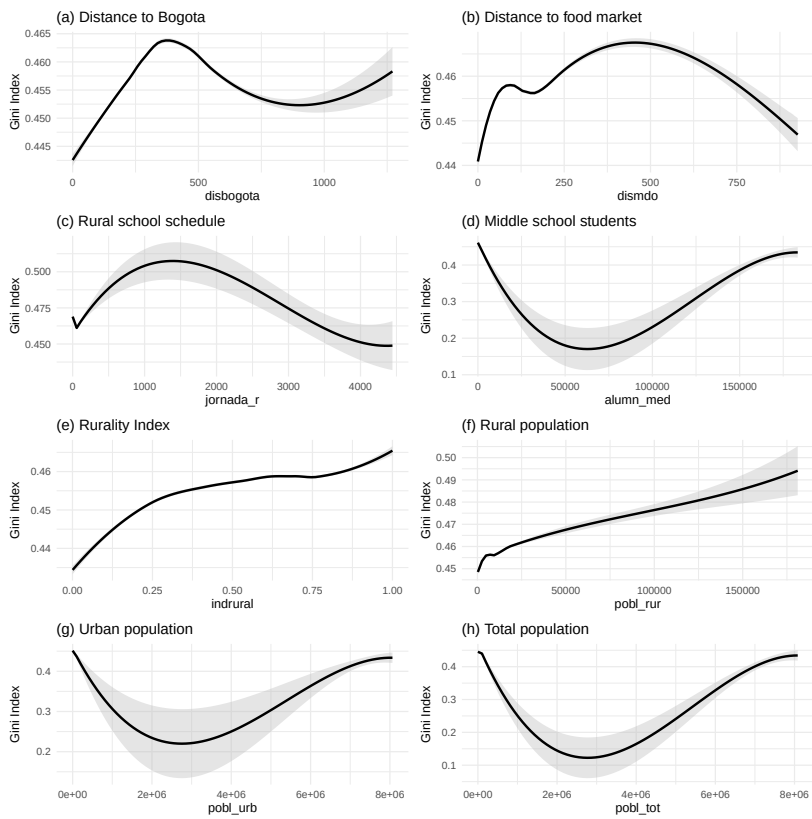
refereed as kindergarten class schedules is the most important variable for this Regression Forest, followed by the rural class schedules. The third important variable is related to the rural index, followed by the number of female teachers in kindergarten. The rural and urban population are next in 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.

2.5 Machine learning estimation with varying factors

Considering that the model with dominant fixed effects might potentially drive differential results, this section will present the behavior of the Synthetic Gini Coefficient with the model of varying factors. In order to have some comparisons, the nonlinear relationships are presented with the same variables of the first models. This 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.

At first glance, it is surprising that the patterns are virtually the same, however, the magnitude of the estimates differs slightly.

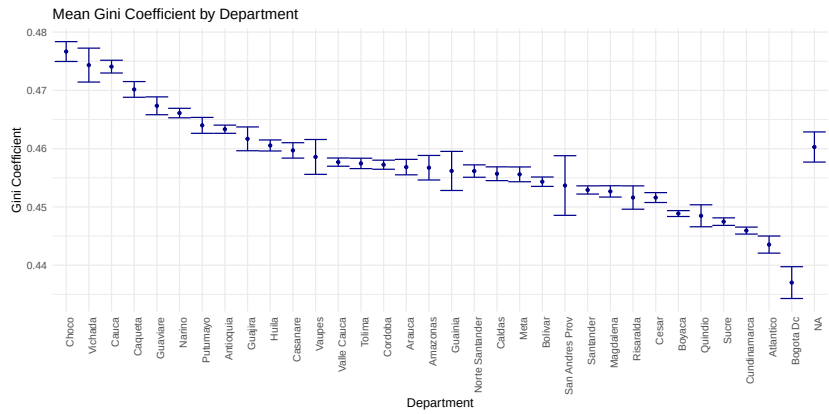
Figure 8: Dispersion of Synthetic Gini Index and features - Varying factors



Source: Own elaboration

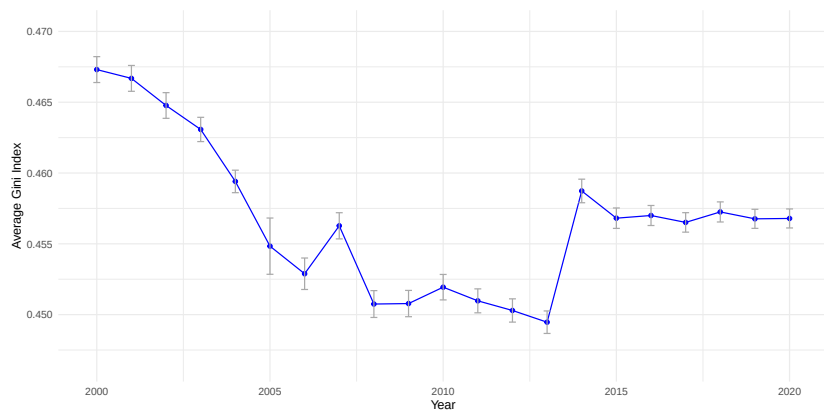
The same pattern exists as shown in figure 8 and described in the previous section. However, the magnitude is what might significantly change according to the values on the y-axis. As this behavior cannot be seen easily in the previous plots. The next figures 9 and 10 presents the variation across Departments 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 more "stiff" 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 does exists. Chocó is in fact the most unequal Department of Colombia again, but now instead of Bogotá, Quindio 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 a 0.4557. Hence, confirms the "stifness" of the dominant fixed effects model.

Figure 9: Synthetic Gini Index at the Department level



Source: Own elaboration

Figure 10: Synthetic Gini Index annual evolution

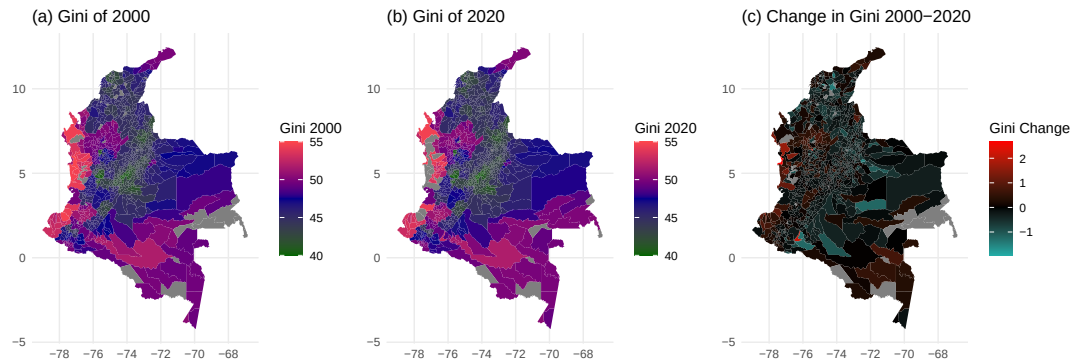


Source: Own elaboration

2.6 Geographical analysis

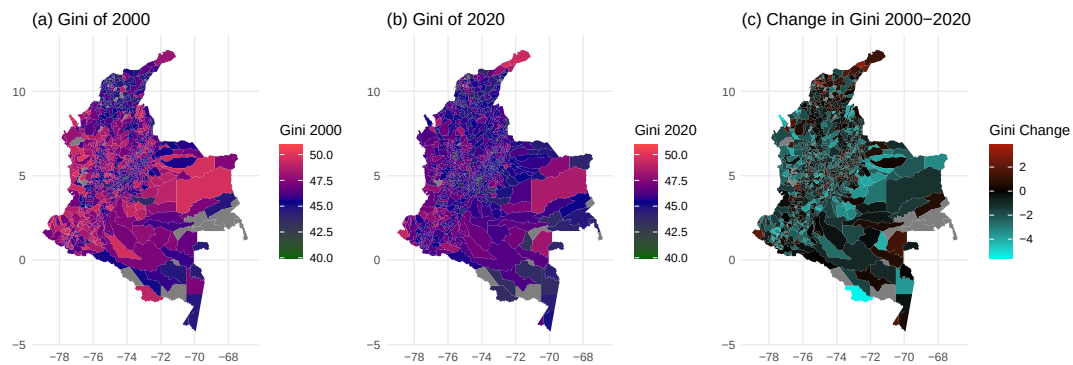
Considering both imputations, the Synthetic Gini Coefficient presents a relative asymmetric evolution from the Dominant Fixed Effects model, and the Varying Factor model. To present these asymmetries, the figures 11 and 12 show the geographical distribution of the Synthetic Gini coefficient for both Dominant Fixed Effects model and the Varying Factor model. Crucial aspects to remark is 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 on Colombia.

Figure 11: Synthetic Gini Index Dominant Fixed Effects



Source: Own elaboration

Figure 12: Synthetic Gini Index Varying Factor model



Source: Own elaboration

The figures contains 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 improvements 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.

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 in this language.

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

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

2.8 Conclusions

This study in an effort to contribute to the missing data problem of income inequality in Colombia has synthesized two datasets estimating the Gini coefficient at the municipality level between the years 2000-2020, through machine learning techniques. In particular, this delivered 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 of the estimates, reflect that Colombia has not improved their income inequality condition. Also, existing heterogeneities are present in the municipalities of Colombia. In particular, for both of the 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 of time, raises questions about the level of welfare improvement in the country [6].

In the Dominant in Varying Factors, I highlight that the composicion of predictors/features to estimate the Synthetic Gini index has two essential categories related to the human capital accumulation (with variables of supply and demand) and population dynamics. The nonlinear relationships between the synthetic Gini, and the information of the predictors, confirms several patters encountered in the economic literature.

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=drive_link&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=drive_link&ouid=118158209086311183140&rtpof=true&sd=true

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

3. Introducción en Español

El objetivo de este artículo es presentar una estimación mediante aprendizaje automático para los datos escasos del coeficiente de Gini a nivel de municipio en Colombia entre los años 2000-2020. Utilizando el poder de los datos de CEDE de la Universidad de los Andes (CEDE, 2023), y la única información existente para el coeficiente de Gini a nivel municipal en 2005, este ejercicio empírico extiende "sintéticamente" las medidas del coeficiente de Gini. Al usar el mejor modelo a través de varias estimaciones, se ejecuta la imputación del coeficiente de Gini, permitiendo sintetizar los datos en formato de datos de panel, incluyendo las claves identificadoras con el municipio DIVIPOLA y los años. Durante las estimaciones, se observó que el mejor modelo tenía como las variables (o características) más importantes un conjunto de "Efectos Fijos Dominantes" que guiaron los resultados del Random Forest. Las características de este modelo están fuertemente relacionadas con los clústeres de municipios, distancias y ubicaciones geográficas. Al interpretar estas características relacionadas principalmente con efectos fijos, -ya que no fluctúan significativamente con el tiempo y están mayormente relacionadas con clústeres geográficos ID- se genera un modelo alternativo caracterizado por contener un conjunto de Factores Variables. Este último modelo se basa en un conjunto de factores socioeconómicos que varían con el tiempo, concentrados principalmente en la acumulación de capital humano y la dinámica poblacional.

Tanto el modelo de Efectos Fijos Dominantes ($R^2 = 98.0\%$), como el modelo de Factores Variables Dominantes ($R^2 = 94.3\%$) estimado mediante Random Forests, muestran buenas propiedades en términos de las métricas, pero también exhiben variabilidad diferencial a lo largo del tiempo y dentro de los individuos. Por lo tanto, mientras que el primer modelo (Dominante en Efectos Fijos) subraya la importancia de factores invariables en el tiempo como distancias, regiones geográficas y fenómenos potencialmente estáticos como infraestructura invariable en el tiempo, instituciones y condiciones climáticas. El segundo modelo subraya la importancia de los factores socioeconómicos para describir el Coeficiente de Gini Sintético. En particular, el modelo de Efectos Fijos Dominantes también contiene como variables importantes las variables socioeconómicas de población y capital humano, pero no son tan importantes como las variables fijas o "semi-fijas" en las estimaciones de Random Forests.

Consistentemente, en ambas estimaciones, Chocó es el Departamento más afectado por la desigualdad de ingresos según los resultados del Coeficiente de Gini Sintético.

Los detalles de esta estimación son los siguientes; primero, se implementa un proceso de filtrado para reducir el número de características con valores faltantes en los datos disponibles. A continuación, se implementa una selección de modelo entre los enfoques de aprendizaje automático de regresión lineal, random forest, árboles de regresión y boosting de gradiente. El random forest supera al resto de los modelos basado en los datos disponibles observados a través de técnicas de validación cruzada. Al observar la composición de las variables más importantes en el modelo inicial, este se describe por una fuerte composición de variables "fijas" y "semi-fijas" que se caracterizan por no cambiar con el tiempo. Sobre esta interpretación, se genera un segundo modelo excluyendo estas variables fijas y semi-fijas. Y así, se estima el modelo de Factores Variables.

La versión pública del Coeficiente de Gini Sintético para el modelo de Efectos Fijos Dominantes es:

https://docs.google.com/spreadsheets/d/1jc1c-X1aum8GkfrsZHOec1gz_1Qo9w_S/edit?usp=drive_link&ouid=118158209086311183140&rtpof=true&sd=true

La versión pública del Coeficiente de Gini Sintético bajo el modelo de Factores Variables es:
https://docs.google.com/spreadsheets/d/1JUt93lBzp3S_kgWnU4msuDJ2LQBGsbVq/edit?usp=drive_link&ouid=118158209086311183140&rtpof=true&sd=true

Considerando que el modelo de Factores Variables contiene más variabilidad a través de los municipios y los años, se recomienda su uso para los investigadores. Esencialmente, porque los efectos fijos pueden ser aislados mediante especificaciones econométricas estándar. El modelo de Factores Variables es, desde mi perspectiva, también más realista, porque los predictores más importantes son variables con el tiempo. Finalmente, aunque este artículo contribuye con medidas sintéticas del coeficiente de Gini a nivel municipal para Colombia durante el periodo 2000-2020, es importante recordar las debilidades de los métodos de aprendizaje automático para la imputación de datos. Esto implica que la motivación para resolver el problema de datos faltantes es específica (Lakshminarayan et al., 1996). Por lo tanto, la imputación de Efectos Fijos Dominantes no es necesariamente mala en comparación con la imputación de Factores Variables. Ambas reflejan estimaciones sintéticas del coeficiente de Gini, pero basadas en diferentes naturalezas. Por lo tanto, ningún algoritmo de aprendizaje automático es la respuesta definitiva para el problema de datos faltantes (Hong and Lynn, 2020). También es importante destacar que todos los métodos de aprendizaje automático están sujetos a sesgos y errores, lo que implica un riesgo de estimaciones sesgadas en los parámetros y errores en las imputaciones (Sullivan et al., 2017).

Las limitaciones de este estudio son que, de hecho, la distribución sintética del índice de Gini estimado en este artículo, podría no describir o alinearse con la distribución real de la desigualdad en Colombia, ya que los datos sintéticos solo se generan a partir de variables observadas disponibles en la esencia de Rubin (1987). Por lo tanto, el conjunto de datos imputado es una estimación sintética, que podría no reflejar la verdadera varianza de la población a través de unidades o tiempo. Lo que podría llevar a inexactitudes potenciales que pueden existir (Schafer and Graham, 2002). Por estas razones y el sesgo potencial que existe dentro de la imputación mediante aprendizaje automático, el conjunto de datos nunca producirá la estimación pura y correcta de la desigualdad de ingresos (Gelman and Hill, 2007) y es imposible compararlo con los datos reales de desigualdad de ingresos, ya que estos no existen. Por lo tanto, los conjuntos de datos generados deben ser utilizados como observaciones informativas.

No obstante, este artículo contribuye a la comprensión de la desigualdad de ingresos mediante la generación del Índice Sintético de Gini y su posible evolución basada en variables observadas a nivel municipal. Así, proporciona dos enfoques basados ya sea en la dominancia de características fijas, o en las características variables a lo largo del tiempo. También presenta un posible resultado para la estimación de la desigualdad de ingresos a través del tiempo en los territorios de Colombia.

Finalmente, los hallazgos adicionales presentan las bien conocidas correlaciones entre la desigualdad de ingresos y la acumulación de capital humano (inversa en naturaleza), los efectos de la población sobre la desigualdad de ingresos (donde la masificación se vuelve

evidente derivando en una relación no lineal en forma de U), y el efecto de la ruralidad sobre la desigualdad de ingresos (donde la población rural está en desventaja ya que es más probable que se enfrente a una mayor desigualdad de ingresos).

Este estudio también contribuye al problema de los datos faltantes que sufre el territorio colombiano en términos del análisis de la desigualdad de ingresos, considerando que el territorio colombiano es una de las economías más desiguales del mundo. Este estudio también fue influenciado por los trabajos relevantes de 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), y Lin and Tsai (2020).

El documento continúa con la sección 4 que describe el proceso de estimación, la Sección 4.1 describe la transformación de los datos utilizada para preservar el número de variables con menos valores faltantes. La Sección 4.2 presenta las generalidades de la estimación mediante aprendizaje automático, donde se compara el desempeño del modelo a través de varios modelos de aprendizaje automático. Presenta la importancia de las variables de los datos, donde en la Sección 4.3 se analiza la interpretación de las variables dentro del random forest. En esta interpretación, se analizan las correlaciones no lineales de algunas variables importantes en relación con el Coeficiente de Gini Sintético, para explorar más a fondo los datos a nivel de Departamento y a lo largo del tiempo. La Sección 4.4 presenta la estimación con factores variables y la comparación entre modelos, la importancia de los factores variables que se concentran principalmente en la acumulación de capital humano y la dinámica poblacional. La Sección 4.5 presenta las dinámicas no lineales entre el Coeficiente de Gini Sintético y los regresores previamente analizados donde emergen prácticamente los mismos patrones. Finalmente, la Sección 4.6 revisa brevemente la distribución geográfica de la Desigualdad de Ingresos Sintética para ambos modelos, el de Efectos Fijos Dominantes y el de Factores Variables.

4. Proceso general

Dado que CEDE (2023) es una de las fuentes de información que consolida los datos municipales de Colombia durante una cantidad significativa de años, seleccioné estos datos como el núcleo de las estimaciones empíricas. Dicho esto, los datos de CEDE para esta versión del estudio (1.02.2025) componen siete paneles de información. Estos están relacionados con:

1. Agricultura y tierra
2. Buen gobierno
3. Características generales (2022)
4. Características generales (2023)
5. Conflicto y violencia
6. Educación
7. Salud y servicios

Desafortunadamente, para todos estos temas no todos los datos están disponibles. En particular, en el panel de Características Generales (ambos años), el coeficiente de Gini solo presenta información para el año 2005 a nivel municipal. El Gini, entonces, falta para los años restantes, consolidando una brecha en nuestro conocimiento sobre el comportamiento de la desigualdad de ingresos en el país.

4.1 Transformaciones de datos

El primer paso fue identificar dentro de los siete paneles de información cuál contenía el mayor número de observaciones disponibles. Identifiqué que el panel 2) Buen Gobierno contenía la mayor información disponible (N=43541), por lo que mantuve este panel como la base central para el proceso de fusión.

El segundo paso fue seleccionar las claves de unión para la gran fusión entre paneles. Como es bien sabido, los códigos municipales (llamados DIVIPOLA) y los años son los adecuados para la tarea posterior de estimar variables a nivel municipal. El resultado de esta fusión creó un panel de 43541 observaciones con 2718 variables.

El tercer paso fue retener en este panel agregado la información entre los años 2000 y 2020. Después de suprimir observaciones bajo esta condición, el panel resultante contenía 23563 observaciones. Como algunos paneles tenían las mismas variables (por lo tanto, repetidas), se ejecutó un algoritmo para eliminar duplicados. Esto dejó un panel de 2654 variables y 23563 observaciones.

El cuarto paso fue contar los valores faltantes para todas las variables y luego calcular el porcentaje de valores faltantes por variable. Se aplicó una condición estricta en este punto: retener solo las variables con un 1 por ciento o menos de valores faltantes. Esto para asegurar, en lo posible, la existencia de información real (sin imputación de datos). Además, se eliminaron algunas variables de tipo carácter vacías (pero no faltantes) [7]. El panel resultante en este punto contenía 48 variables y 23563 observaciones.

El quinto paso fue convertir todas las variables de tipo carácter a variables numéricas, lo que dejó nuevamente algunas variables vacías pero no faltantes [8]. En este punto, se aplicó una condición final para retener solo casos completos en el panel. El panel resultante contenía 44 variables y 23082 observaciones.

El sexto paso fue recuperar la información única del Gini disponible solo en el año 2005, la cual estaba dentro del panel de Características Generales 2023. Luego se fusionó de nuevo con el panel limpio sin valores faltantes. Este panel final se usó para técnicas de estimación. El panel para este proceso contenía 45 variables y 23082 observaciones.

4.2 Estimación mediante aprendizaje automático

El software y los programas utilizados para estimar los diferentes modelos de aprendizaje automático son Ridgeway and Ridgeway (2004), Kuhn (2008), Wickham (2011), Therneau et al. (2015), Wickham et al. (2019), Yarberry and Yarberry (2021), donde primero se verificó la estructura para asegurar que todas las variables o "características" fueran numéricas

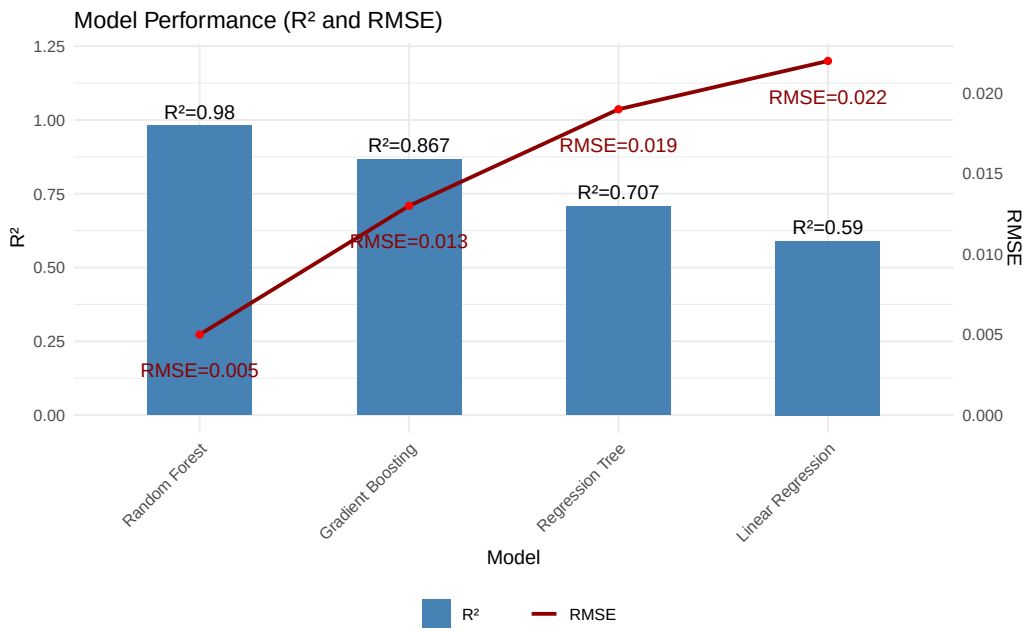
7. Esto se aplicó a las variables que contenían "" en el panel y estaban relacionadas con las variables "DF2 categórica, DF2 doinicial, DF2 rango, categoría", las cuales pertenecen al panel de buen gobierno pero fueron consideradas no vacías. Por lo tanto, se eliminaron.
8. Las variables "depto, provincia, municipio" y "act adm" eran identificadores incompletos de los municipios, no relevantes ya que el código de cada uno se encontraba en otras variables.

(incluyendo la conversión de las categóricas). El segundo paso para la imputación mediante aprendizaje automático fue identificar los datos donde el Gini existe y donde falta.

El tercer paso fue definir los predictores (44 variables) y la variable objetivo (el Gini). Luego se estableció una semilla para entrenar los modelos mediante validación cruzada. El número de particiones seleccionadas para este proceso fue de cinco para entrenar el algoritmo con el enfoque de validación cruzada. Dado que algunas variables eran esencialmente duplicadas y que algunos identificadores también estaban incluidos en los paneles, se realizó un proceso de limpieza para dejar un único tipo de identificador de ubicación geográfica.

El cuarto paso fue estimar los modelos de regresión lineal, árbol de regresión, bosque aleatorio (Random Forest) y gradiente boosting para investigar cuál era el más adecuado en los datos de entrenamiento. Los resultados del paso 4 se encuentran en la figura 13, que contiene las medidas clásicas de R^2 y $RMSE$ para resultados continuos.

Figure 13: Desempeño de estimación de modelos



Fuente: Elaboración propia

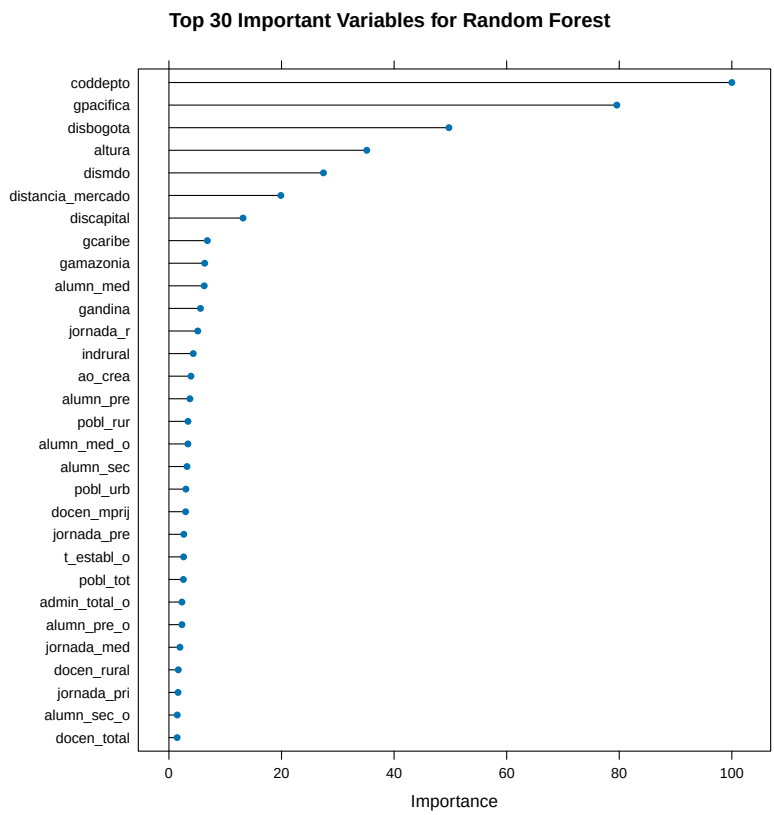
Estas estimaciones sugieren que el enfoque de bosque aleatorio (Random Forest) supera al resto de los modelos. En particular, exhibe un $R^2 = 98.67\%$ con un $RMSE = 0.005$, lo que sugiere un resultado preciso incluso después de la validación cruzada con cinco particiones.

Con este mejor modelo (el bosque aleatorio o Random Forest), se imputan los datos del índice de Gini para todo el panel sin valores faltantes ($K = 45, N = 23082$), incluyendo la variable objetivo. El panel contiene, por lo tanto, información de acerca 1091 municipios para los años 2000 y 2020.

Un resultado adicional importante es la relevancia de las variables en la estimación del

bosque aleatorio (Random Forest), que se muestran en la figura 14.

Figure 14: Importancia de las características



Fuente: Elaboración propia

4.3 Interpretación de las variables dentro del bosque aleatorio (Random Forest)

Al observar la figura 14, es posible identificar un patrón en la escala de importancia para la estimación del bosque aleatorio (Random Forest). En particular, la variable inicial "coddepto" se refiere al clúster geográfico de las divisiones administrativas departamentales de Colombia. Esta es la división administrativa siguiente a la Nación. La variable "gpacifica" está relacionada con la ubicación de la "región pacífica" de Colombia. Luego se encuentra la distancia hacia Bogotá (la capital de Colombia), la altitud, la distancia lineal al mercado de alimentos más importante cercano (en Km), la distancia lineal al municipio donde se encuentra el mercado alimentario más grande, la distancia lineal a la capital del departamento, seguida de otras variables de ubicación fija, como las regiones Andina, Amazonia y Caribe. Este primer conjunto de variables tiene algo en común: están relacionadas con las ubicaciones físicas geográficas del municipio.

En el segundo conjunto de variables importantes se encuentran características socioeconómicas que involucran dinámicas poblacionales, temas escolares y educativos. Este conjunto de variables incluye la densidad rural de los municipios, el número total de horarios escolares (clases), el número total de estudiantes en educación media, el número total de estudiantes en preescolar, el número total de estudiantes en educación secundaria, la población urbana y total del municipio, el número de docentes, el número de instalaciones educativas y el número de horarios escolares para la educación primaria.

Interpreto la selección de las características fijas o "semi" fijas, como los clústeres geográficos y las distancias, como una forma de "efectos fijos" desde la perspectiva econométrica. En general, existen algunas características invariantes en el tiempo que explican una parte significativa de la heterogeneidad de los individuos, lo que implica la presencia de dichos efectos fijos constantes. Esto se ajusta a las categorías de distancia y clústeres geográficos estáticos. Más allá de esta interpretación, podrían incluirse características institucionales invariantes en el tiempo, como el desempeño de las autoridades locales, la infraestructura física constante y las vías de comunicación o acceso a otros municipios, en particular, la capital de Colombia, Bogotá, y la capital de los departamentos. Las interconexiones de mercado también juegan un papel significativo aquí.

Según la figura 14, estos efectos fijos son los principales impulsores de la desigualdad entre los municipios colombianos. La pregunta de qué y cómo cambiarán las estimaciones se abordará en la siguiente sección. Por ahora, utilizando estos datos, es interesante ver cómo la desigualdad de ingresos podría estar relacionada con las variables continuas utilizadas en las estimaciones del bosque aleatorio (Random Forest), para entender mejor la caja negra dentro de los datos sintéticos y qué patrones pueden identificarse empíricamente.

La figura 15 muestra algunas de las interrelaciones no lineales entre el coeficiente de Gini sintético y algunos de los principales regresores continuos (o características) utilizados en el modelo de bosque aleatorio (Random Forest). Surgen patrones claros de estas dispersiones: a) Cuanto mayor sea la distancia del municipio con respecto a la capital del país (Bogotá), mayor será la desigualdad de ingresos. b) Cuanto mayor sea la distancia del municipio con respecto al mercado alimentario más grande cercano, mayor será la desigualdad de ingresos. c) Un mayor tamaño de los horarios escolares en áreas rurales tiende a generar una mayor desigualdad hasta un punto de inflexión, donde los horarios escolares son lo suficientemente grandes como para invertir el patrón, resultando en una disminución de la desigualdad (relación en forma de U invertida). d) Existe una relación decreciente entre la desigualdad y el número de estudiantes de educación media, sin embargo, en algún punto, cuando los estudiantes se conglomeran, la desigualdad comienza a crecer nuevamente. e) Cuando la proporción de población rural supera a la población urbana, implica una creciente desigualdad. f) Como consecuencia, existe una relación positiva y casi lineal entre la población rural y la desigualdad de ingresos. g) La población urbana tiene un efecto decreciente en las estimaciones de la desigualdad de ingresos; sin embargo, cuando la población urbana se masifica, la desigualdad comienza a crecer nuevamente (relación en forma de U). h) La población total imita el comportamiento de la población urbana; por lo tanto, cuando la masificación surge, la desigualdad vuelve a crecer.

Estos patrones son consistentes con la literatura económica [9], donde, por ejemplo, las

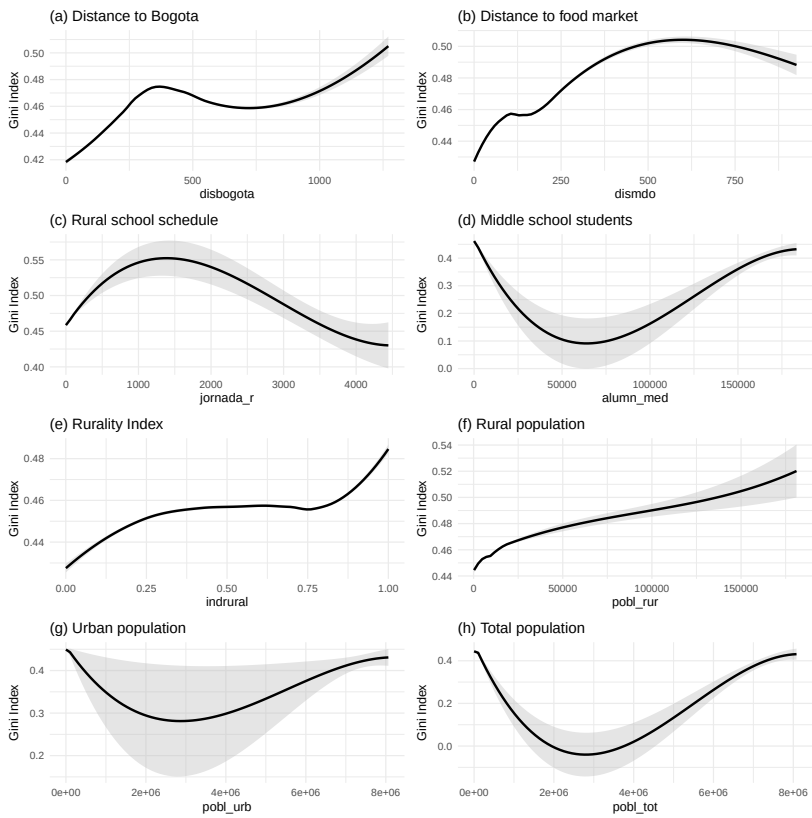
9. En particular, esto se alinea con los estudios de Paas and Schlitte (2008), Rey (2004), Salvati (2016), Kühn (2015), Oppido et al. (2023), Riveros-Gavilanes (2023), Lee and Lee (2018), Castelló-Climent and Doménech (2021), Lee and Vu (2020) relativo al comportamiento de estas variables observables y la desigualdad en el ingreso.

regiones o zonas periféricas tienden a sufrir desigualdad de ingresos debido a la falta de interconexión de mercados, servicios públicos y provisión de bienes. Por esta razón, los patrones de la figura 15 relativos a los paneles (a) y (b) son fuertes, capturando la distancia relativa a la capital del país y también la distancia lineal a los mercados alimentarios centrales. Según este argumento, las dinámicas poblacionales, como la concentración rural y la población rural, parecen estar correlacionadas positivamente con la desigualdad sintética, como se observa en los paneles (e) y (f). Finalmente, las dinámicas no lineales de acumulación de capital humano también son muy interesantes de analizar. El hecho de que los horarios escolares rurales tengan una forma de U invertida refleja algunas de las dinámicas internas relacionadas con las oportunidades y la asistencia escolar, que tenderán a disminuir la desigualdad si son lo suficientemente altos. En particular, cuando los horarios escolares satisfacen la demanda educativa en los territorios rurales, es más probable encontrar una reducción en la desigualdad sintética. Por otro lado, desde el lado de la demanda, como se observa en el panel (d), cuando el número de estudiantes de educación media aumenta lo suficiente, estos enfrentan una restricción en el acceso a la educación, lo que genera un aumento en la desigualdad sintética. Finalmente, las dinámicas poblacionales relativas a la población urbana (g) y la población total (h) muestran una relación en forma de U, lo que implica que, cuando la masificación de los municipios ocurre, la desigualdad sintética tiende a aumentar.

A continuación, se muestra el comportamiento del índice de Gini sintético producido por el bosque aleatorio (Random Forest) a nivel departamental y su evolución en el tiempo en las figuras 16 y 17. Las barras de error del gráfico implican que no hay variación significativa a nivel intra-departamental. Esto es consistente con las variables más importantes utilizadas en el bosque aleatorio (Random Forest), que son esencialmente "fijas". El departamento con la mayor desigualdad de ingresos sintética es Chocó, con un Gini de casi 0.54, mientras que la menor desigualdad de ingresos se refleja en Bogotá, con un Gini cercano a 0.43. Esto no es una sorpresa, ya que Chocó se ha caracterizado históricamente por la falta de oportunidades, la concentración masiva de ingresos y la distancia a la capital. Por otro lado, Bogotá se convierte en el lugar menos desigual en las estimaciones con los factores fijos. Las áreas periféricas, incluyendo Cauca, Amazonas, Vaupés, Guainía, Guaviare y Caquetá, también siguen el patrón de alta desigualdad de ingresos.

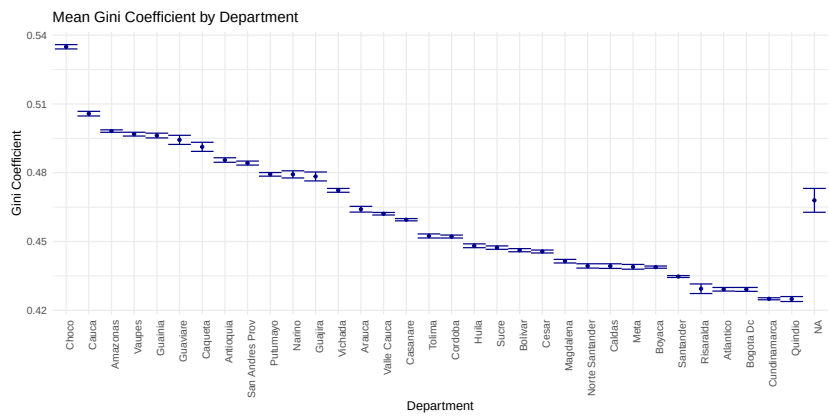
La evolución anual de la desigualdad de ingresos sintética producida por el bosque aleatorio (Random Forest) no muestra cambios significativos. De hecho, la desigualdad sintética parece rondar consistentemente un Gini de 0.457 y un 0.454, reflejando un promedio de desigualdad de ingresos estancado para cada año. En el mejor de los casos, ha disminuido solo un 0.002 del Gini en un lapso de 20 años.

Figure 15: Dispersión del Índice de Gini Sintético y características



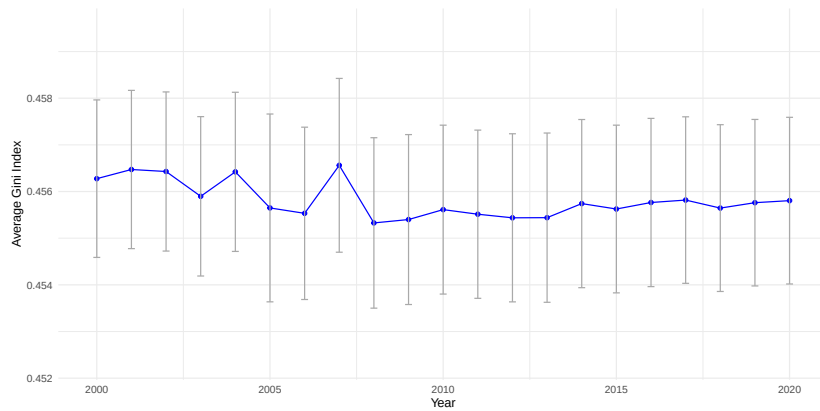
Fuente: Elaboración propia

Figure 16: Índice de Gini Sintético a nivel Departamental



Fuente: Elaboración propia

Figure 17: Evolución anual del Índice de Gini Sintético



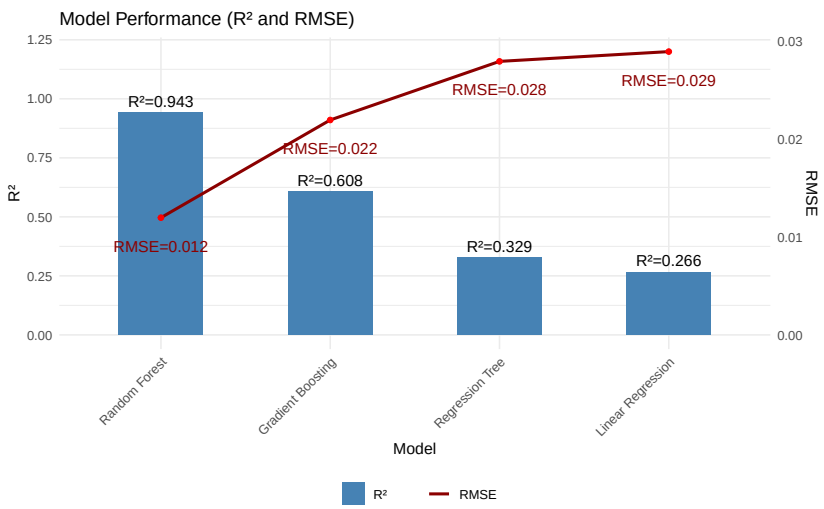
Fuente: Elaboración propia

4.4 Estimación con factores variables en el tiempo

Considerando la interpretación anterior, y en particular el patrón mostrado en la figura 17, uno podría preguntar cómo podría cambiar el coeficiente de Gini sintético al dejar en el Random Forest un conjunto de regresores puramente variables. En particular, es fácil identificar los regresores fijos y "semi-fijos"[10] de la figura 14.

Con esta idea, reestimo el modelo Random Forest excluyendo los regresores fijos y "semi-fijos"[11]. Tras este proceso, el desempeño de los modelos de aprendizaje automático se presenta en la figura 18, donde es visible que el Random Forest nuevamente supera al gradient boosting, a los árboles de decisión de regresión y a la regresión lineal. Las métricas para el Random Forest son un $R^2 = 94.3$ y un $RMSE = 0.012$.

Figure 18: Desempeño del modelo con regresores variables



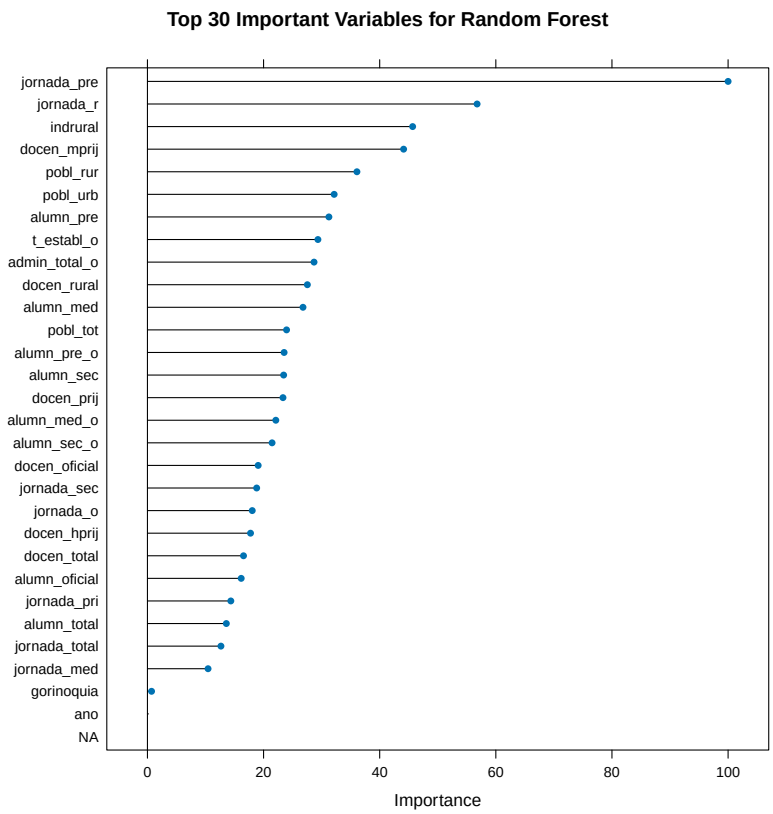
Fuente: Elaboración propia

Mientras que el primer modelo "Dominante en efectos fijos" estima con una precisión de $R^2 = 98.67\%$, el modelo "Dominante en factores variables" estima la precisión en $R^2 = 94.3\%$. Sorprendentemente, la pérdida es de aproximadamente 4.37%, lo cual se interpreta como una disminución no significativa. Ahora, ¿qué tipo de regresores dominan los modelos de factores variables? Inspeccionemos esto en la siguiente figura, donde se presenta la importancia de las variables.

El conjunto de factores que impulsan el Random Forest ahora está relacionado con las características socioeconómicas. En particular, la acumulación de capital humano es

10. El término semi-fijo se utiliza aquí para referirse a aquellas variables continuas que no varían significativamente. Por ejemplo, la distancia lineal a la capital, esta variable, aunque es continua por naturaleza, no cambia con el tiempo, constituyendo un efecto fijo.
11. Específicamente se eliminan los siguientes efectos fijos: "codmpio", "codprovincia", "codmdo", "coddepto", "gpacifica", "disbogota", "altura", "dismdo", "discapital", "gcaribe", "gamazonia", "gandina", "ao cre", "ao crea", "distancia mercado", "mercado cercano", donde en particular las últimas cuatro variables son esencialmente efectos semi-fijos, ya que se refieren al año de creación del municipio y la distancia al mercado de alimentos más cercano (que se repite con "mercado cercano"). Por tanto, dejo solo una medida de distancia a los mercados de alimentos.

Figure 19: Importancia de las variables con factores variables



Fuente: Elaboración propia

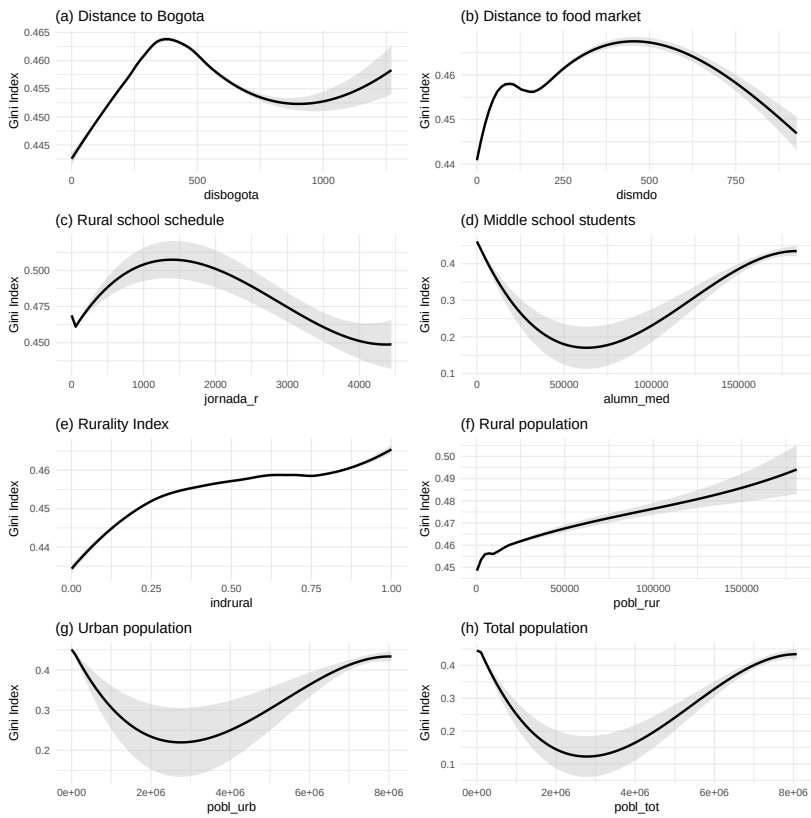
uno de los principales impulsores junto con las dinámicas poblacionales. La etapa inicial de acumulación de capital humano, referida a los horarios de clases de preescolar, es la variable más importante para este Regression Forest, seguida por los horarios de clases rurales. La tercera variable importante está relacionada con el índice rural, seguida por el número de maestras en preescolar. La población rural y urbana siguen en importancia. Luego el número de estudiantes en preescolar, el número de instalaciones educativas, el número de personal administrativo en dichas instalaciones y el número de profesores en áreas rurales, seguidos por el número de estudiantes en secundaria. La población total también entra aquí junto con más indicadores de capital humano.

4.5 Estimación de aprendizaje automático con factores variables

Considerando que el modelo con efectos fijos dominantes podría potencialmente impulsar resultados diferenciales, esta sección presentará el comportamiento del Coeficiente de Gini Sintético con el modelo de factores variables. Para hacer algunas comparaciones, las relaciones no lineales se presentan con las mismas variables de los primeros modelos. Esto para inspeccionar alguna posible robustez de los resultados y patrones. La figura 20 muestra los

patrones relativos del Coeficiente de Gini Sintético con los regresores del modelo con efectos fijos dominantes. A primera vista, es sorprendente que los patrones sean prácticamente los mismos; sin embargo, la magnitud de las estimaciones difiere ligeramente.

Figure 20: Dispersión del Índice Gini Sintético y factores - Factores variables

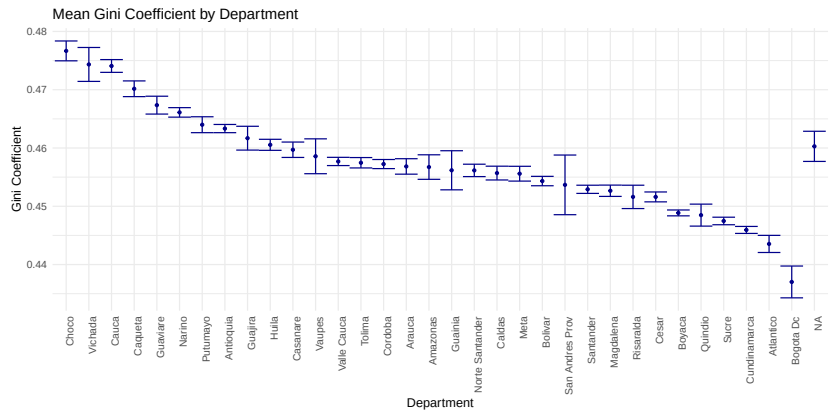


Fuente: Elaboración propia

El mismo patrón existe, como se muestra en la figura 20 y se describe en la sección anterior. Sin embargo, la magnitud es lo que podría cambiar significativamente según los valores en el eje y. Como este comportamiento no se puede observar fácilmente en los gráficos anteriores, las siguientes figuras 21 y 22 presentan también la variación a través de Departamentos y la evolución a lo largo del tiempo del Índice Gini Sintético bajo el modelo de factores variables dominantes. Como era de esperar, en contraste con el modelo de efectos fijos dominantes, el modelo de factores variables presenta una mayor volatilidad a través de los Departamentos y el tiempo. Esto se debe a que los impulsores son más "rígidos" en el primer modelo al sintetizar el coeficiente de Gini a lo largo del tiempo. Por otro lado, existe más variabilidad al confiar en los regresores variables. Sin embargo, algunos hallazgos consistentes existen. Chocó es, de hecho, nuevamente el Departamento más desigual de Colombia, pero ahora, en lugar de Bogotá, Quindío parece ser el mejor en términos de desigualdad sintética. A lo largo del tiempo, se observan dinámicas diferen-

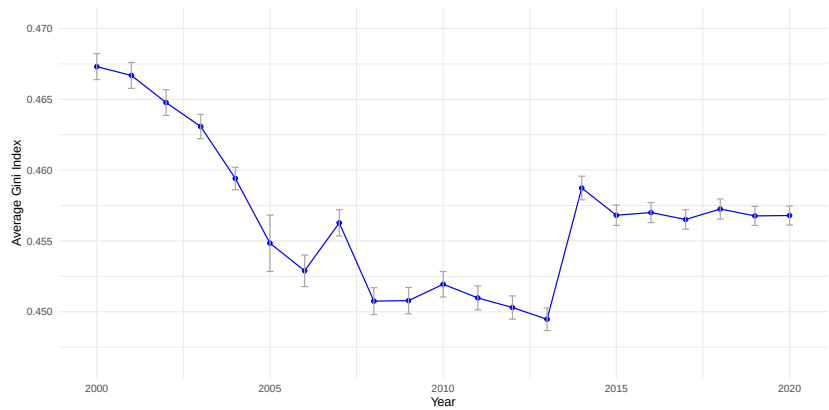
ciales: inicialmente, la desigualdad es de aproximadamente 0.469, pero se reduce a 0.459 con algunos aumentos significativos en los años 2006 y 2013. Mientras que en el modelo de efectos fijos dominantes la desigualdad inicial comenzó en 0.4567 y disminuyó a 0.4557. Esto confirma la "rigidez" del modelo de efectos fijos dominantes.

Figure 21: Índice Gini Sintético a nivel Departamental



Fuente: Elaboración propia

Figure 22: Evolución anual del Índice Gini Sintético

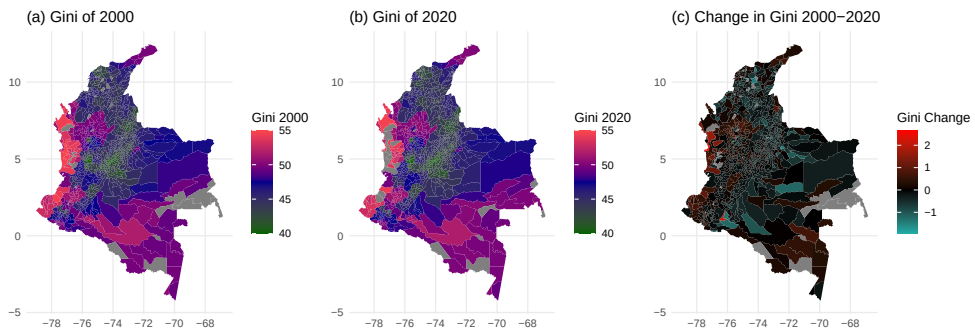


Fuente: Elaboración propia

4.6 Análisis Geográfico

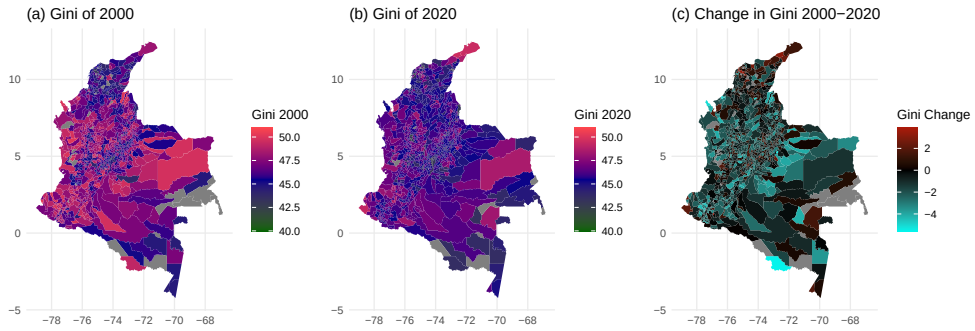
Considerando ambas imputaciones, el Coeficiente de Gini Sintético presenta una evolución relativamente asimétrica entre el modelo de Efectos Fijos Dominantes y el modelo de Factores Variables. Para presentar estas asimetrías, las figuras 23 y 24 muestran la distribución geográfica del Coeficiente de Gini Sintético para ambos modelos. Un aspecto crucial a destacar es que el modelo de Efectos Fijos Dominantes presenta una evolución conservadora de la desigualdad de ingresos, con una mejora apenas perceptible. Mientras tanto, el modelo de Factores Variables muestra algunas mejoras en Colombia.

Figure 23: Índice Gini Sintético con Efectos Fijos Dominantes



Fuente: Elaboración propia

Figure 24: Índice Gini Sintético con modelo de Factores Variables



Fuente: Elaboración propia

Las figuras contienen un conjunto de tres paneles: el panel (a) muestra la estimación del Gini Sintético para el año 2000, el panel (b) hace lo mismo para el año 2020, y el panel (c) muestra el cambio absoluto. Las áreas negras del último panel representan un cambio nulo en el lapso de 20 años, las áreas rojas muestran un empeoramiento donde la desigualdad aumentó, y las áreas azul claro representan posibles mejoras en la desigualdad

de ingresos. Este marco de distribución concluye que el modelo de Factores Variables presenta la evolución más optimista de la desigualdad de ingresos en Colombia, ya que el panel (c) presenta zonas más claras en contraste con el modelo de Efectos Fijos Dominantes.

Un resultado divergente se encuentra en las áreas donde aumentó la desigualdad de ingresos. Por ejemplo, en las estimaciones conservadoras del modelo de Efectos Fijos Dominantes, el occidente de Colombia, cerca de la región pacífica, se ha deteriorado en términos de su desigualdad de ingresos. En contraste, el modelo de Factores Variables muestra que la parte norte de Colombia ha empeorado su condición de desigualdad de ingresos.

4.7 Estadísticas Descriptivas

Esta sección solo presenta las tablas de estadísticas descriptivas. Como la parte final del estudio en este idioma.

Table 3: Estadisticas descriptivas del Gini modelo Dominante de Efectos Fijos

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 4: Estadisticas descriptivas del Gini modelo Dominante en Factores Variantes

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

4.8 Conclusiones

Este estudio, en un esfuerzo por contribuir al problema de la falta de datos sobre la desigualdad de ingresos en Colombia, ha sintetizado dos conjuntos de datos que estiman el coeficiente de Gini a nivel municipal entre los años 2000-2020, mediante técnicas de aprendizaje automático. En particular, esto resultó en dos modelos estimados mediante la técnica de Bosques Aleatorios (Random Forests), que fue el mejor modelo entre las estimaciones, superando a las técnicas de boosting de gradiente, árboles de regresión y enfoques de regresión lineal. Los dos modelos utilizando Bosques Aleatorios se describen de la siguiente manera: 1) Un modelo de Efectos Fijos Dominantes, denominado así por la importancia de los predictores/características que lo componen, principalmente basado en un conjunto de variables fijas o semi-fijas. 2) Un modelo Dominante en Factores Variables, como una alternativa al modelo de Efectos Fijos Dominantes, estimado principalmente con predictores/características que varían en el tiempo. Ambos modelos presentan un buen desempeño en sus métricas, con un $R^2 = 98.67\%$ para el modelo de Efectos Fijos Dominantes y un $R^2 = 94.3\%$ para el modelo Dominante en Factores Variables. Por lo tanto, la estimación de estos modelos para el resto del panel se denomina Coeficiente/Índice de Gini Sintético.

El modelo de Efectos Fijos Dominantes presenta el Coeficiente de Gini Sintético con cambios mínimos/conservadores/rígidos, donde la desigualdad en promedio ha disminuido de 45.63 en el año 2000 a 45.58 en 2020. En contraste, el modelo Dominante en Factores Variables presenta mayor variabilidad a lo largo del tiempo, comenzando con una desigualdad media de 46.73 que se ha reducido a 45.67. Todas las estimaciones reflejan que Colombia no ha mejorado significativamente su condición de desigualdad en los ingresos. Además, existen heterogeneidades presentes en los municipios de Colombia. En particular, para ambas estimaciones, el Departamento de Chocó es el más desigual del territorio colombiano. El hecho de que Colombia no haya mejorado significativamente la situación de desigualdad de ingresos durante este período de tiempo plantea interrogantes sobre el nivel de mejora del bienestar en el país [12].

En el modelo Dominante en Factores Variables, se destaca que la composición de los predictores/variables para estimar el índice Sintetico de Gini tiene dos categorias esenciales relacionadas con la acumulación de capital humano (tanto variables de oferta como de demanda) y dinámicas poblacionales. La relación no lineal entre el índice sintetico Gini estimado, y la información de estos regresores, confirma varios patrones encontrados en la literatura económica.

La versión pública disponible del Coeficiente de Gini Sintético para el modelo de Efectos Fijos Dominantes es:
https://docs.google.com/spreadsheets/d/1jc1c-X1aum8GkfrsZHOec1gz_1Qo9w_S/edit?usp=drive_link&ouid=118158209086311183140&rtpof=true&sd=true

12. En particular, porque la desigualdad de ingresos ha estado negativamente correlacionada con los niveles de bienestar entre las economías, ver Dagum (1990), Clark and Kavanagh (1996), Abdel-Rahman and Wang (1997), Coburn (2015), Kim (2017), Riveros-Gavilanes (2021), Wildowicz-Szumarska (2022), Coady et al. (2022), Riveros-Gavilanes et al. (2022), Yang and Tang (2023), y Sologon et al. (2023)

La versión pública disponible del Coeficiente de Gini Sintético bajo el modelo de Factores Variables es:

https://docs.google.com/spreadsheets/d/1JUt93lBzp3S_kgWnU4msuDJ2LQBGsbVq/edit?usp=drive_link&ouid=118158209086311183140&rtpof=true&sd=true

References

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

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

Castelló-Climent, A. and Doménech, R. (2021). Human capital and income inequality revisited. *Education Economics*, 29(2):194–212.

CEDE (2023). Panel municipal cede, centro de estudios sobre el desarrollo económico. `https://datoscede.uniandes.edu.co/catalogo-de-datos/`.

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

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

Coburn, D. (2015). Income inequality, welfare, class and health: A comment on pickett and wilkinson, 2015. *Social science & medicine*, 146:228–232.

Dagum, C. (1990). On the relationship between income inequality measures and social welfare functions. *Journal of Econometrics*, 43(1-2):91–102.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gond, V. K., Dubey, A., and Rasool, A. (2021). A survey of machine learning-based approaches for missing value imputation. In *2021 third international conference on inventive research in computing applications (ICIRCA)*, pages 1–8. IEEE.

Hong, S. and 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.

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.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, 28:1–26.

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

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.

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

Lee, K.-K. and Vu, T. V. (2020). Economic complexity, human capital and income inequality: a cross-country analysis. *The Japanese Economic Review*, 71(4):695–718.

Lin, W.-C. and Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509.

Lin, W.-C., Tsai, C.-F., and Zhong, J. R. (2022). Deep learning for missing value imputation of continuous data and the effect of data discretization. *Knowledge-Based Systems*, 239:108079.

Oppido, S., Ragozino, S., and Esposito De Vita, G. (2023). Peripheral, marginal, or non-core areas? setting the context to deal with territorial inequalities through a systematic literature review. *Sustainability*, 15(13):10401.

Paas, T. and Schlitte, F. (2008). Regional income inequality and convergence processes in the eu-25. *Scienze regionali: Italian Journal of regional Science: 7, supplemento 2, 2008*, pages 29–49.

Rey, S. J. (2004). Spatial analysis of regional income inequality. *Spatially integrated social science*, 1:280–299.

Ridgeway, G. and Ridgeway, M. G. (2004). The gbm package. *R Foundation for Statistical Computing, Vienna, Austria*, 5(3).

Riveros-Gavilanes, J. M. (2021). Estimación de la función de bienestar social de amartya sen para américa latina. *Ensayos de Economía*, 31(59):13–40.

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

Riveros-Gavilanes, J. M., Al Akayleh, F., Oduniyi, O., Samuel, A. H., and Hassan, S. M. (2022). On the welfare trends: A view from the sen's social welfare function. Technical report, M&S Research Hub institute.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

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.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Seu, K., Kang, M.-S., and Lee, H. (2022). An intelligent missing data imputation techniques: A review. *JOIV: International Journal on Informatics Visualization*, 6(1-2):278–283.

Sologon, D. M., Doorley, K., and O'Donoghue, C. (2023). Drivers of income inequality: what can we learn using microsimulation? *Handbook of Labor, Human Resources and Population Economics*, pages 1–37.

Sullivan, T. R., Lee, K. J., Ryan, P., and Salter, A. B. (2017). Multiple imputation for handling missing outcome data when estimating the relative risk. *BMC medical research methodology*, 17:1–10.

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

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.

Wang, S., Li, B., Yang, M., and 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*, pages 67–72. Springer.

Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 3(2):180–185.

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 tidyverse. *Journal of open source software*, 4(43):1686.

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.

Xue, J. (2023). Review on data imputation methods in machine learning. In *Journal of Physics: Conference Series*, volume 2646, page 012034. IOP Publishing.

Yang, X. and Tang, W. (2023). Additional social welfare of environmental regulation: the effect of environmental taxes on income inequality. *Journal of Environmental Management*, 330:117095.

Yarberry, W. and Yarberry, W. (2021). Dplyr. *CRAN recipes: DPLYR, stringr, lubridate, and regex in R*, pages 1–58.