Integrating survey and administrative data to predict informality status in Colombia: imputation in March and April 2020 during COVID-19 pandemic

Authors: Juan Oviedo-Arango[1♠], José Lobo-Camargo [2♠], Anderson Leal-Vélez[3♥], Cristhyan Naranjo-Puertas[4♣], Juan Ordoñez-Herrera[5♣], Andrés García-Suaza[6♦]

Abstract

The COVID-19 pandemic created important challenges to produce statistical information. For the labor market indicators, the National Statistics Offices (NSO) often rely on household surveys that were negatively affected by the pandemic because of reduced questionnaires. This article proposes different prediction algorithms of the informality status within the household surveys for the March-April 2020 period, in which it was not possible to collect all the characteristics required to determine this variable. The matching of the household survey (GEIH) and the social security register (PILA) in Colombia allow us to exploit a novel source of variation which enhances the prediction for these two critical months.

Keywords: missing data, household survey, informality, COVID-19, machine learning, administrative registers; integration; linking records

JEL codes: E26; C45; C53; C81

[1♠] DANE, General Director;
[2♠] DANE, Professional in Administrative Records for statistical purposes.
[3♥] DANE, Professional at IT division; **Presenter; e-mail: Alealv@dane.gov.co**
[4♣] DANE, Professional in Administrative Records for statistical purposes.
[5♣] DANE, Advisor at General Directory
[6♦] Universidad del Rosario. External advisor, DANE.

1. INTRODUCTION

During the second quarter of 2020, confinement and social distancing policies were imposed globally with the aim of containing the spread of COVID-19. This situation has brought important challenges to the National Statistical Offices (NSO's) to monitor in real time the main economic and social variables in order to support timely decision-making. In this context, continuous statistical operations, such as household surveys, which are usually collected monthly through face-to-face interviews, were greatly affected, and therefore had to be redesigned to maintain the data collection process. In fact, UNSTATS (2020) reports that 96% of INEs stopped collecting face-to-face data partially or totally.

In the case of the labor market, the Great Integrated Household Survey (GEIH, for its acronym in Spanish), as the main characterization instrument, faced challenges of collecting data during the confinement, particularly in the months of March and April. The data collection process was modified to maintain a subset of indicators that allowed tracking the behavior of the labor market. These modifications have been implemented in a significant number of countries and follow the suggestions of ILO (2020). The March and April shorter questionnaires lack the question about the employee numbers of the employer that is required to calculate the official informality rate of Colombia, hence the time series of informality status that are more than ten years old are missing for these two months.

Therefore, a machine learning algorithm, Random Forest, is estimated on the data set of the GEIH employed for 2019 and the first semester of 2020. The data include a set of socioeconomic variables and a novel indicator of a deterministic match between GEIH and the social security register PILA. This imputation made it possible to estimate the informality rate for the months of March and April 2020.

The United Nations Economic Commission for Europe (UNECE) highlights the importance of using machine learning methodologies in the official statistics production. In particular, the UNECE points out that National Statistical Offices can make use of these methods to: (i) make inference, (ii) correct the unit of non-response, (iii) impute the non-response, (iv) measure the error of a model, and (iv) make predictions of the near future (UNECE, 2018).

In this sense, the work conducted by DANE, contributes to; provide evidence on the measurement errors associated with the collection mode and on the learning / recall effect. To do this, a large-scale household survey is analyzed, and the redesign of the Colombian household survey is exploited as an experimental scenario to understand the impact of these two measurement errors. For this purpose, adequate counterfactuals are constructed through the integration of survey data with administrative data. In addition, given the context of the pandemic, additional exercises were conducted to assess the impact of changes in data collection on sampling and the prevalence of non-response.

This document is organized into five sections, the first is this introduction. The second describes the tools and ideas proposed to tackle the challenges of household surveys in the COVID-19 pandemic. The third section explains the data and methodology applied in the imputation procedure. In the fourth, the main results are presented. Finally, section five presents the conclusions.

2. HOUSEHOLD SURVEYS IN THE COVID-19 PANDEMIC

The quality of the information collected through traditional operations such as household surveys (HS) has been strongly affected by the restrictions imposed by the confinement since the end of March of 2020. This has led to consider possible biases in the self-reporting of economic activities in the face of confinement and the impossibility of performing normal work.

The ILO has issued a series of suggestions to ensure monitoring of the behavior of the labor market and reduce possible bias. These suggestions include considering the situation of absence of the employed, which could have increased as a result of lockdowns. Moreover, the importance of monitoring the employed population who have seen their working hours reduced, and the unemployed discouraged by the possibility of finding employment. This monitoring can be supported by alternative data sources such as surveys of establishments or administrative records (PILA in the case of Colombia).

In this monitoring process, it is recommended to make a disaggregation that considers:

• People absent from work due to absence, duration and pay (as applicable);

• Employees who work more / less hours than usual, as well as the reasons;

• People outside the labor force by degree of connection to the labor market and for the reasons for not seeking or not being available to work;

• People who recently lost a job due to job termination, reasons and general characteristics of their last job position (occupation, branch of activity, status in occupation).

In addition, the operatives have been forced to implement telephone interviews, which may increase biases associated with the greater probability of not locating some population groups due to the lack of a telephone line or the availability of contact information. These biases make it necessary to carry out a household replacement selected to a survey or to make an adjustment through the responses to the information collected.

3. DATA AND METHODOLOGY

3.1. Data

The Great Integrated Household Survey (GEIH) and the social security administrative register (PILA, for its acronym in Spanish) are the two sources of data that allow us to estimate the impact on the use of the collection method and quantify the relevance of the recall bias. The GEIH collects the main variables that allow characterizing the Colombian labor market on a monthly basis- among them the unemployment, participation, and occupation rate. This survey collects information from 20,700 households on average per month and generates representative information for geographic domains of the main 23 cities and for the rest, and, in addition, for each one of the 23 cities. The GEIH structure is modular and includes modules of sociodemographic information, as well as variables of employment characteristics, including the income level. For the proposed exercises, the data corresponding to the period January 2019 to June 2020 are used.

Furthermore, in order to have a reference measurement of the income of the employed, information from the Statistical Register of Labor Relations (RELAB) produced by DANE, based on the

PILA record, is exploited. The RELAB contains monthly information of the employee-employer relationship of a subset of employees with payments to the social security system. The matching of GEIH and RELAB is carried out using the personal identifiers of the register system of Colombia. The matching allows us to assess the biases between the telephone and face-to-face survey controlling for the possible heterogeneities between geographic domains such as the 23 cities, the rest urban and dispersed rural areas.

The objective of the proposed model is to forecast the informality status, which is a dichotomous variable that takes the value of 1 in case the employed person is informal and 0 if the person is formally employed. For official statistics purposes, the informal employed are the people who, during the reference period, were in one of the following situations:

1. Private employees and workers who work in establishments, businesses or companies that employ up to five people in all their agencies and branches, including the employer and/or its partner.
2. Unpaid family workers in companies with five or less workers.
3. Unpaid workers in companies or businesses of other households.
4. Domestic employees in companies with five or less workers.
5. Day laborers in companies with five or less workers.
6. Self-employed workers who work in establishments with up to five people, except professional freelancers.
7. Employers in companies with five or fewer workers.
8. Government employees are excluded.

This variable is calculated for the period between January 2019 and February 2020, and the months of May and June 2020 where the questionnaire has the information required to calculate the informality status. The inclusion of the last two months is crucial because the estimated model then has information on the changes in employment patterns that could have resulted from the COVID-19 pandemic.

The data are splitted into three groups: i. training, corresponding to 70% of the sample; ii. testing, for the remaining 30%; and iii. the imputation period with the data of workers in the 23 main cities in the months of March and April 2020. According to the official informality definition, there are groups of workers for which the informality status is known with the available information. Government employees as well as professional self-employed workers are both formal workers, therefore they are excluded from the sample and a value of 0 is assigned in the dichotomous variable. Thus, the sample consists of 276,419, corresponding to the 23 main cities.

The algorithm exploits the variation in socioeconomic and employment characteristics of the individuals to forecast the informality status of a worker. In particular, the variables sex, age, city, occupation, economic activity, occupational position, and the registration of novelties such as vacations or suspensions are considered. In addition, the month and year of the survey report are included to control for seasonal patterns in the labor market.

Furthermore, the matching of GEIH and RELAB make available significant additional variation to impute the informality status. It is common to consider that RELAB records correspond to the formal component of employment, so that an employee in GEIH that is also found in RELAB is very likely to have a job under formality status. In fact, Table 1 shows that, for the periods of January and April 2019 and January 2020, a large percentage of the employed persons that match between GEIH and RELAB are formally employed. The non-match of GEIH-RELAB has a high percentage of informal workers, almost 80%. This is an important remark since it means that information from administrative records may be relevant to determine the informality status.

Table 1. Percentage of informality in the GEIH-RELAB match

	jan-19		apr-19		jan-20	
	Non-RELAB	RELAB	Non-RELAB	RELAB	Non-RELAB	RELAB
Formal	22,14	77,36	21,66	78,63	20,75	78,51
Informal	77,86	22,64	78,34	21,37	79,25	21,49

Source: DANE, GEIH-RELAB

Thus, among the individual characteristics, a dichotomous variable that indicates whether the GEIH worker is also among the RELAB workers in the same month is considered. This indicator, crucial in the classification exercise as we will see later, constitutes an important and novel statistical application of administrative registers to enhance a household survey. Moreover, the employment situation in RELAB is also included in the data and an indicator that validates the quality of the identification of the respondents is included from GEIH since lesser quality inhibits the potential match with RELAB and could be related to the informality status. Finally, considering the impact of COVID-19 on employment patterns, a dichotomous variable that takes the value of 1 during the pandemic period[7] is included.

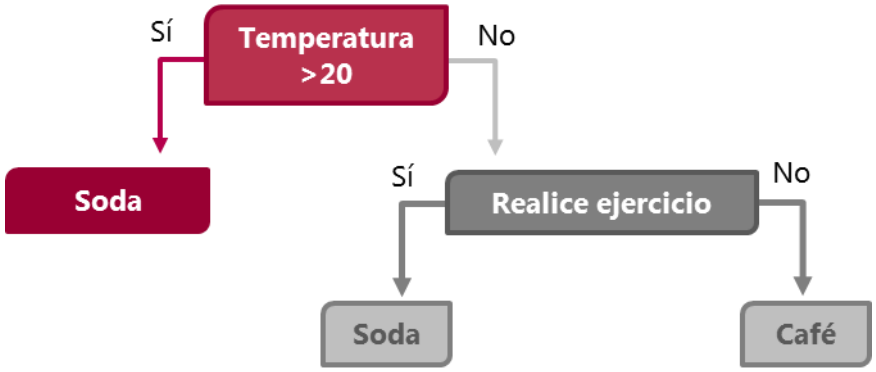
3.2. Methodology

We implemented machine learning algorithms which are increasingly being used in economic and labor market analysis (Gerunov 2014 and Athey and Imbens, 2019), among them the Random Forest (RF) algorithm is a refinement of the decision trees algorithm (see James et al, 2013 and Lantz 2015). The decision tree algorithm separates records into subsets with a higher level of homogeneity, which can be measured through an entropy index. In this way, the algorithm builds a series of simple decisions, like sentences of the form *if… then…* in a computational algorithm.

[7] The months of March to June 2020

To illustrate how the algorithm works, suppose you want to decide what drink to have for breakfast. Such a decision depends on several factors such as the ambient temperature and the activities you do in the morning. In this way, if the ambient temperature is high or you decide to exercise, you will prefer a cold drink, e.g., soda. In other cases, you would prefer coffee. Then, the choice of the drink has the temperature as the first criterion, and the physical activity as the second criterion. This sequence of decisions can be represented in a tree as shown in Figure 1. The same logic can be implemented on the observable characteristics of individuals to determine if a worker is formal or informal.

Figure 1. Example of the Decision Tree Learning Mechanism



Under this same perspective, the decision trees algorithm is trained from a set of records that allow inferring patterns and establishing rules for classifying them as informal employees. This same principle applies to RF that is part of the decision tree assembly models. An RF is a set of decision trees that makes the classification task more robust through the random choice of data sets and variables (see Breiman, 2001, and Lantz, 2015). In this way, the classification exercise does not depend only on the construction of a tree, but also allows to generalize patterns on the data without easily falling into the memorization of these by the algorithm.

For the RF training, 70% of the sample is considered while the remaining 30% is used to analyze the predictive power of the estimated model. To evaluate the performance of the algorithm, metrics such as the F1 *score*[8] and the level of precision are used.

[8] In statistics the F-Value is a measure of precision calculated as a weighted average of precision (proportion of those correctly classified as informal) and recall (proportion of informal classified correctly), where an F1 score reaches its best value at 1 and worst score at 0.

4. RESULTS

The implementation of the RF algorithm requires the selection of parameters that determine the machine learning process. A priori, the selection of these parameters is arbitrary, however, from the comparison of scenarios (or cross-validation process) it is possible to determine the parameter settings that generate the best performance. In this case, these parameters refer to the maximum percentage of variables that each tree trains and the number of classifiers or trees.

The best configuration for the algorithm corresponds to the inclusion of all the available information based on different exercises of selection of variables. The implementation of the cross-validation technique suggested the estimation of 134 decision trees using 26.8% of the variables available in each of these. Table 2 presents the classification error obtained in the training and testing samples, where it can be observed that the model has a good level of performance.

Table 2. Training error and test of the RF algorithm

Training error	11,52%
Testing error	11,47%

Source: RELAB-GEIH, author's calculation

A deeper analysis of the performance of the algorithm can be carried out through the confusion matrix presented in Table 3. This matrix shows the level of success of the model and how the classification errors are distributed within the test data. The algorithm possibly does not present a bias in the classification errors, that is, the errors tend to be distributed in a relatively uniform way among formal employees who are classified as informal, and vice versa.

Table 3. Confusion matrix of the test data of RF algorithm

			Prediction values	
Observed values	**Categories**		**Formal**	**Informal**
			0	**1**
	Formal	**0**	32.360	5.174
	Informal	**1**	5.545	50.221

Source: Source: RELAB-GEIH, author's calculation

The model metrics such as Precision, Recall and F1 score validate the results observed in the confusion matrix, i.e. the algorithm has good performance since these metrics are close to one. The F1 score is an average of the previous metrics and (see Table 4) shows a satisfactory performance of the algorithm.

Table 4. RF algorithm metrics

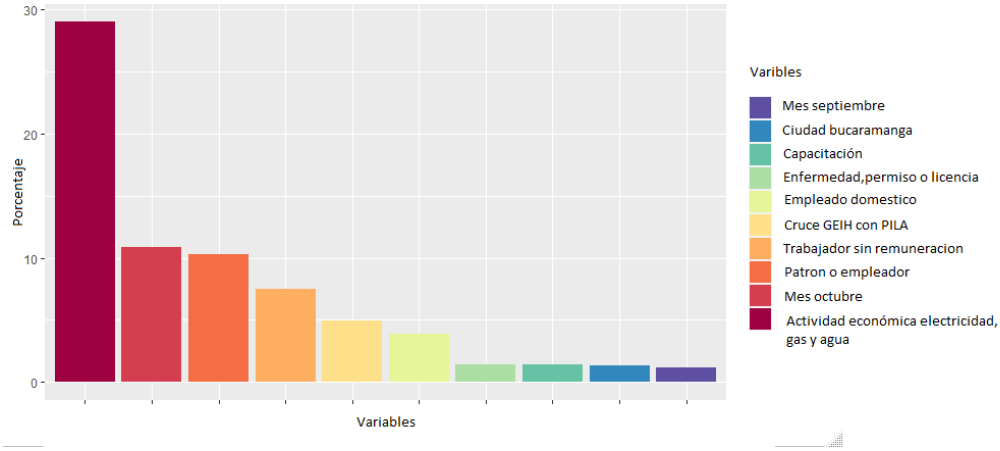
Precision	0,8537
Recall	0,8622

F1 score	0,8579

Source: RELAB-GEIH, author's calculation

An interesting attribute of this algorithm, which cannot be generalized to other machine learning algorithms, is that it offers the possibility of inferring the contribution of each of the variables in the classification process. The results, presented in Figure 2, indicate that the five variables with the highest incidence in the classification of the employed between formal and informal are the economic activity electricity, gas, and water; October; occupational positions of employer and unpaid family worker, and the RELAB-GEIH matching indicator variable.

Figure 2. Relative importance of variables in the RF algorithm forecast



Source: RELAB-GEIH, author's calculation

The final RF algorithm is applied to the months of March and April. This allows not only to carry out the analyzes at the microdata level and correlate informality with characteristics of the individuals, but also to recover the series of the informality rate (see table 5).

Table 5. Imputation of informality status

		mar-19	abr-19	mar-20 (e)	abr-20 (e)
13 main cities	Formal	5.761.320	5.563.060	5.490.990	4.327.579
	Informal	4.999.454	5.068.967	4.317.637	3.335.724
	Informality rate (%)	46,46%	47,68%	44,02%	43,53%
23 main cities	Formal	6.236.682	6.061.929	5.926.923	4.678.564
	Informal	5.693.664	5.744.017	4.903.411	3.724.109
	Informality rate (%)	47,72%	48,65%	45,27%	44,32%

Source: GEIH, author's calculation

5. CONCLUSIONS

The integration of GEIH, a household survey, and RELAB, a statistical register, enhance the imputation of the informality status in March and April 2020. This application allows us to overcome some of the challenges in the recollection stage of the survey due to the COVID-19 pandemic. The imputation process generates the micro-data of the informality status that is essential to any analysis of the effects of the COVID-19 pandemic in the informality. The imputation process also fills the two months gap in a long time series that is key for the monitoring of the labor market in a developing country such as Colombia.

The results also contribute to the improvement of official statistics using administrative registers and a household survey in an integrated manner. Statistical operations that exploit administrative registers constitutes a long-standing trend in National Statistics Offices. The administrative registers now not only are used in the construction of the statistical framework from which a random sample for the surveys is taken but also serves in the imputation and analysis stage of the statistical process.

6. REFERENCES

- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685-725.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Gerunov, A. (2014). Big data approaches to modeling the labor market.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Lantz, B. (2015). Machine learning with R. Packt Publishing Ltd.
- OIT (2020). COVID-19: Orientaciones para la recolección de datos de las estadísticas del trabajo. https://ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/ publication /wcms_745104.pdf
- Yung, Wesley, et al. "The Use of Machine Learning in Official Statistics." UNECE Machine Learning Team report (2018).
- UNECE (2018). Two-phase and double machine learning for data editing and imputation