Labor informality at the municipality level : A Small Area Estimation approach.

Andrea Otero Cortés
Central Bank of Colombia

Sarajevo, October 31st 2024

Disclaimer

The views and opinions expressed in this presentation are those of the authors and do not necessarily reflect the views or official policy of Banco de la República or its Board of Directors. Any errors are exclusively the responsibility of the authors.

This project is part of a larger research agenda on informality that will be published as an ESPE at the Central Bank of Colombia in February 2025. All the research projects that are part of the ESPE, including this one, have benefited from the discussion and feedback of more than 20 coauthors from the Central Bank of Colombia and other institutions (Universidad de los Andes and Fedesarrollo).

Table of Contents

Key features on labor informality in Colombia

- The aggregate labor informality rate has been consistently above 50%, but it has decreased over time with important regional variations.

- Usual findings:
  - High negative correlation between education and informality.
  - High negative correlation between firm size and informality: larger firms have low labor informality rates, while small and micro firms have high rates.

- No statistical differences between informality rates for men and women.
- U-shaped pattern between labor informality and age.

- Most of the labor informality in Colombia is due to informal self-employment.

Importantly: Survey data only allows us to measure labor informality at the main capital cities, aggregate rural areas and at the national level. Admin records only allow us to capture formal employment, but not informal workers.

Definitions

- **Labor informality**: Two measures:

a. Legalistic definition: Based on the payment of mandatory pension contributions.

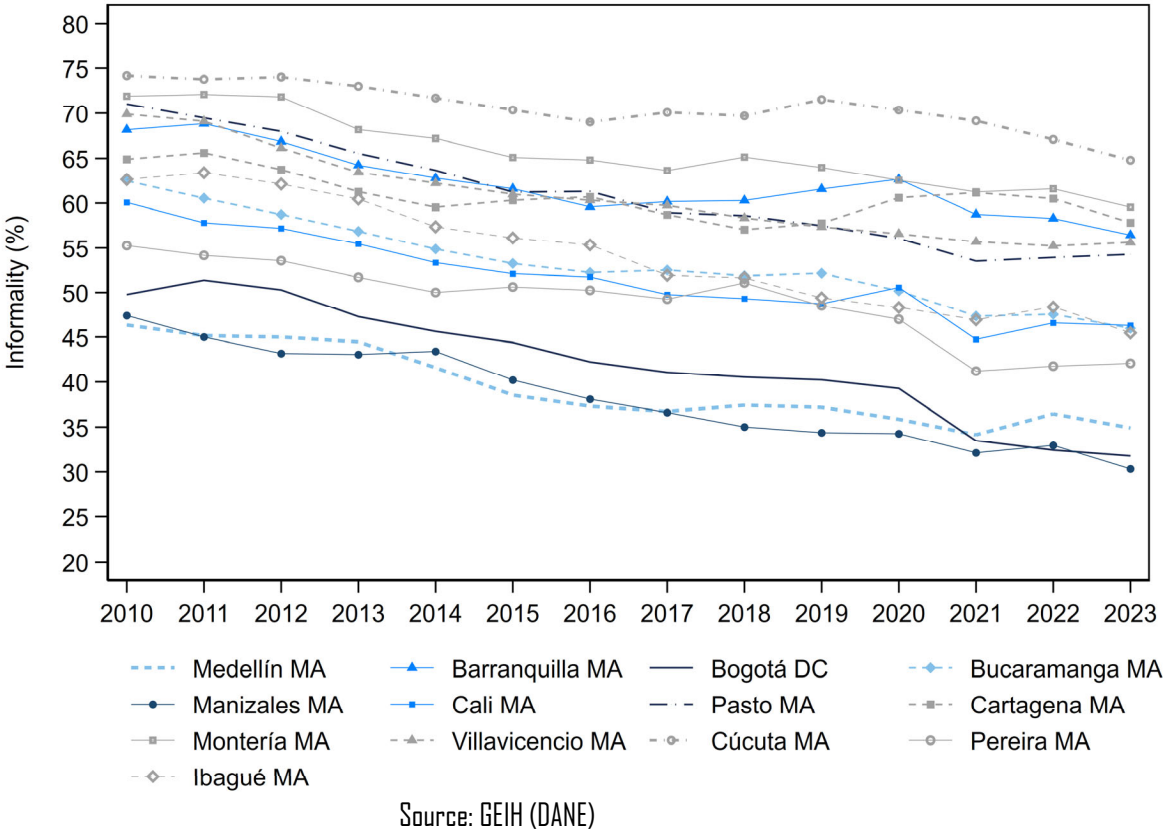
a. DANE definition: All salaried employees or domestic workers who do not have health or pension contributions. Informal workers also include all unpaid workers, self-employed workers and employers classified in the informal sector, which consists of economic units without commercial registration and formal accounting, such as firms with up to 5 employees.

- **Firm informality:** Two measures:

a. Based on a multidimensional IMIE index that takes into cosideration four components (informal if IMIE>50%):
1. Entry costs (Registro Unico Tributario and Registro Mercantil).
2. Taxes (accounting, corporate income tax, value-added tax and manufacturing industry and commerce taxes).
3. Inputs (social security contributions and payroll taxes).
4. Product (operating costs).

b. Based only on paying registration costs.

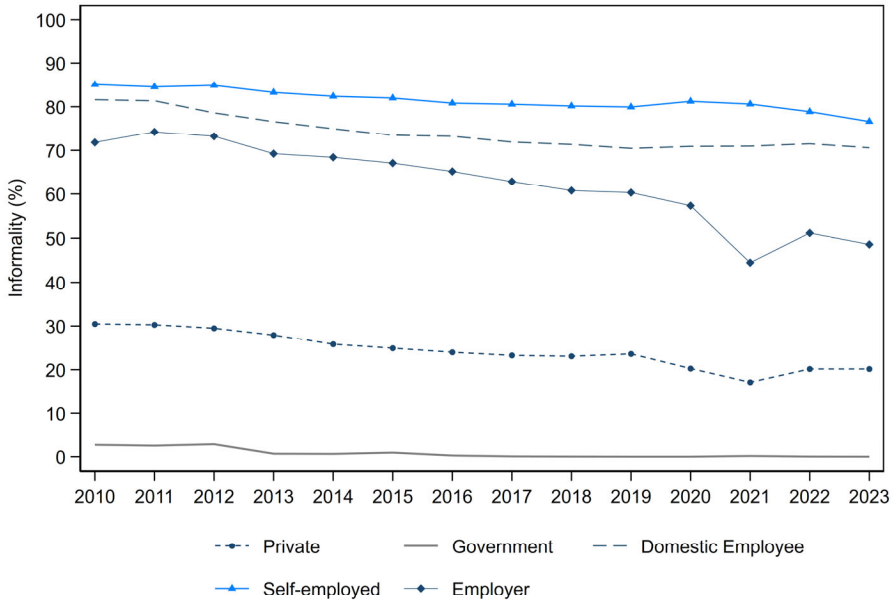
Labor informality rate in main cities and metro areas, 2010-2023.



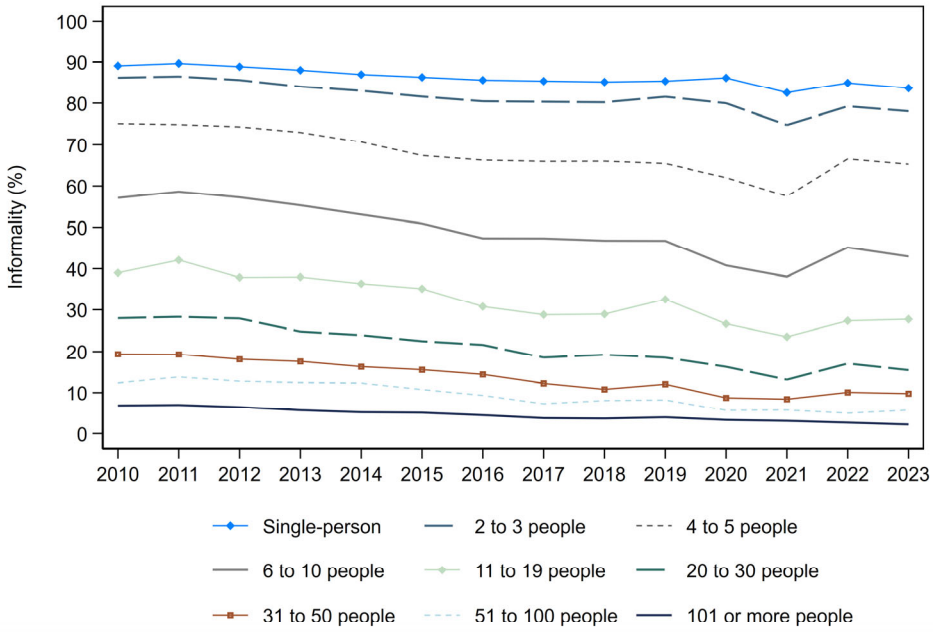
Source: GEIH (DANE)

- In 2023, the average labor informality rate in the main cities of Colombia was 56,7%, but there has been a steady decline in the share of informal workers in th past 13 years.

Labor informality rate by job occupation, 2010-2023



Labor informality rate by firm size, 2010-2023



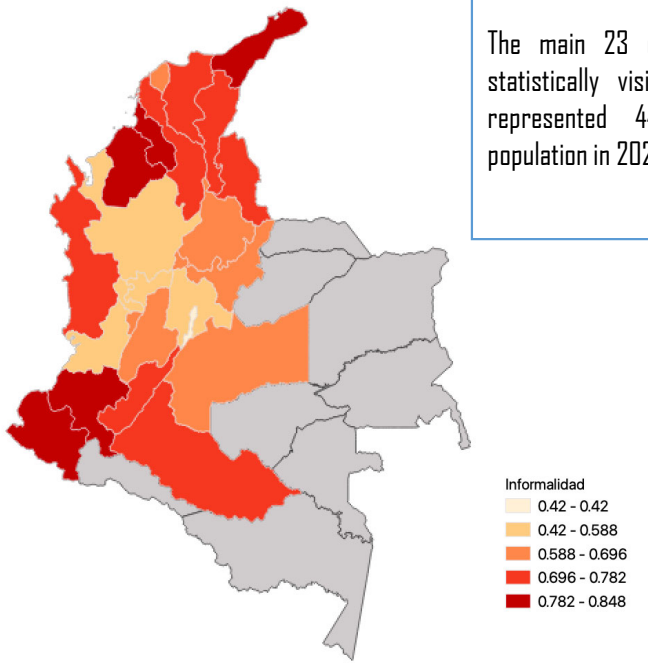
Source: GEIH (DANE)

- Labor informality in Colombia is mostly due to informal self-employment, rather than informal employees. Informal self-employed workers represent 70%-75% of all informal workers.
- Larger firms exhibit lower labor informality, while micro and small businesses are the ones with the highest labor informality rates.

Labor informality at the municipality level : A SAE approach.

Motivation

- In Colombia, we do not have representative survey data at the municipality level, outside from the main capital cities to estimate labor informality rates.
- Small area estimation (SAE) techniques can be an alternative to measure labor informality rates for statistically "invisible" populations.



The main 23 capital cities, which are statistically visible in the GEIH survey, represented 44% of the Colombian population in 2021.

Methodology: Transformed Fay-Herriot estimators

- Fay-Herriot (FH) estimators are designed for continuous variables, therefore we need to use a modified version in order to guarantee that the estimations are within a [0,1] range .

- To use the transformed FH estimator as proposed by Schmid et al. (2017), we first need to compute the basic FH estimator.
  - In this case, we need: (1) a sampling model, and (2) a linking model, which allows us to obtain a direct estimator for statistically observable populations and a synthetic estimator for all the municipalities for which we do not have stastically representative data.

Methodology: Fay-Herriot estimators

Population Division

- The total population P is divided into a small, mutually exclusive areas.
- Information is only available for a sample of n units from the population.

Sampling

- The sampled units from an area are indexed with m, and the non-sampled ones with r.

Variable and Weights

- The target variable y is continuous (e.g., income).
- Each sampled unit has a weight w , reflecting its importance.

Estimator Formula

- A weighted average of the observed values y in each area m is used, considering their weights w.

$$\hat{\theta}_m^{Direct} = \frac{\sum_{i=1}^{n_m} w_{mi} y_{mi}}{\sum_{i=1}^{n_m} w_{mi}} \qquad (I)$$

Assuming the existence of a continuous target variable y, y_{mi} corresponds to the one unit variable i ∈ N in an area $m \in a$ and their corresponding sampling weights are w_{mi}.

Methodology: Fay-Herriot estimators

- $\hat{\theta}_m^{Direct}$ estimates biased values of the areas for which the sample is not representative. To obtain unbiased estimators, the Fay-Harriot model proposes a two-stage method:

First Stage: Sampling Model

- Direct estimates are obtained for each area based on observed data.

$$\hat{\theta}_m^{Direct} = \hat{\theta}_m + \epsilon_m \qquad (2)$$

Second Stage: Linking Model

- A linear regression is fitted at the area level using covariates.

$$\hat{\theta}_m = X_M^T \beta + u_m \qquad (3)$$

- X_M^T: Vector of relevant auxiliary variables for the area.
- β : Model coefficients.
- u_m : Random error at the area level.
- ϵ_m : Sampling error.

Methodology: Fay-Herriot estimators

Linear Mixed Model

- From (2) and (3) the mixed model is derived:

$$\hat{\theta}_m^{Direct} = X_M^T \beta + u_m + \epsilon_m \quad (4)$$

Best Linear Unbiased Predictor (BLUP) of $\hat{\theta}_m$

$$\hat{\theta}_m^{BLUP} = \mathrm{X_M^T}\hat{\beta} + \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}(\hat{\theta}_m^{Direct} - \mathrm{X_M^T}\hat{\beta}) = \gamma_m \hat{\theta}_m^{Direct} + (1 - \gamma_m)\mathrm{X_M^T}\hat{\beta} \quad (5)$$

- γ_m represents the shrinkage factor of an area m and is defined as $\sigma_u^2\left(\sigma_u^2 + \sigma_{\epsilon_m}^2\right)^{-1}$.
- σ_u^2 is the between-area variance.
- $\sigma_{\epsilon_m}^2$ is the within-area error variance.

Methodology: Transformed Fay-Herriot Estimators

In order to transform the basic FH estimator, we use the methodology proposed by Schmid et al (2017) due to the type of distribution followed by the variable of interest, informality.

Steps:

- Transform the direct estimators with the inverse sine function: $v_m = f\left(\hat{\theta}_m^{Direct}\right) = sin^{-1}\left(\sqrt{\hat{\theta}_m^{Direct}}\right)$

- The sample variance of v_m is approximated by $\sigma_e^2 = \frac{1}{4\tilde{n}_m}$. Where $\tilde{n}_m$ represents the effective sample.
- The classic FH equation (5) is estimated for v_m y $\sigma_{\epsilon_m}^2$
- The estimators obtained from the previous step are transformed back to their original scale:

$$\hat{\theta}_m^{FH-transf} = f^{-1}\left(\hat{\theta}_m^{FH}\right) = sin^2\left(\hat{\theta}_m^{FH}\right) \qquad \text{For } m = 1, \dots, a.$$

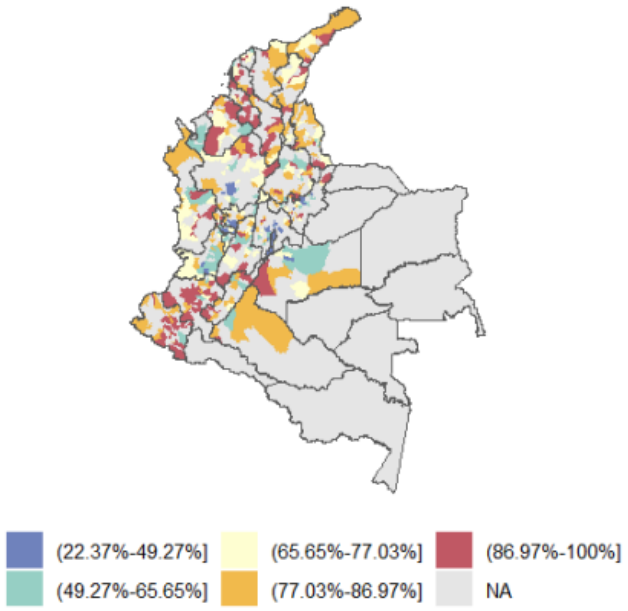
Lastly, we also compute confidence intervals for the $\hat{\theta}_m^{FH-transf}$ estimators, which are obtained through resampling (bootstrap).

Data: Survey data on labor informality (GEIH)

- The sampled units are obtained from the GEIH, which is representative at the national, total urban and total rural level.
- The GEIH is collected through a sampling process, which on average surveys individuals in 436 municipalities per month across Colombia.

Representative for:

- 24 capital cities + metro areas.
- 23 departments.



- Note: The survey is collected in other municipalities but it is not statistically representative.

Data: Geographical variables

- This set of variables capture differences in location, access to markets, topography, and institutions.

Descriptive statistics of geographical variables

Variable	Mean	Standard deviation	Min	Max
Years since creation of the municipality	145.7	110.4	9	491
Altitude ($masl$)	1,134.3	919.8	1	3,35
Area (km^2)	1,019.3	3,203.8	15	65,674
Distance to department capital (km)	81.5	60.6	0.0	493.1
Distance to main market (km)	68.2	100.7	0.0	913.2
Distance to Bogotá (km)	320.9	192.6	0.0	1,228.1

Sources: DANE, PILA, SPE, Panel municipal CEDE.

Data: Demographic variables

- They aim to capture differences in population size, population structure by gender and age, and the urbanization process in each municipality.

Descriptive statistics of demographic variables

Variable	N	Mean	Standard deviation	Min	Max
		2011			
Population	1,118	39,803	242,871	240	7,152,656
Dependency ratio	1,118	88.4	19.3	44.9	253.6
Rurality	1,118	57.8	23.9	0.1	100.0
		2016			
Population	1,120	41,797	249,327	262	7,300,918
Dependency ratio	1,120	82.8	19.2	43.4	228.6
Rurality	1,120	56.9	23.9	0.1	100.0
		2021			
Population	1,113	45,81	268,572	311	7,823,334
Dependency ratio	1,113	79.0	15.0	45.4	190.4
Rurality	1,113	55.1	23.8	0.05	100.0

Sources: DANE, PILA, SPE, Panel municipal CEDE.

Data: Socio-economic variables

Descriptive statistics of socio-economic variables – 2011 & 2021

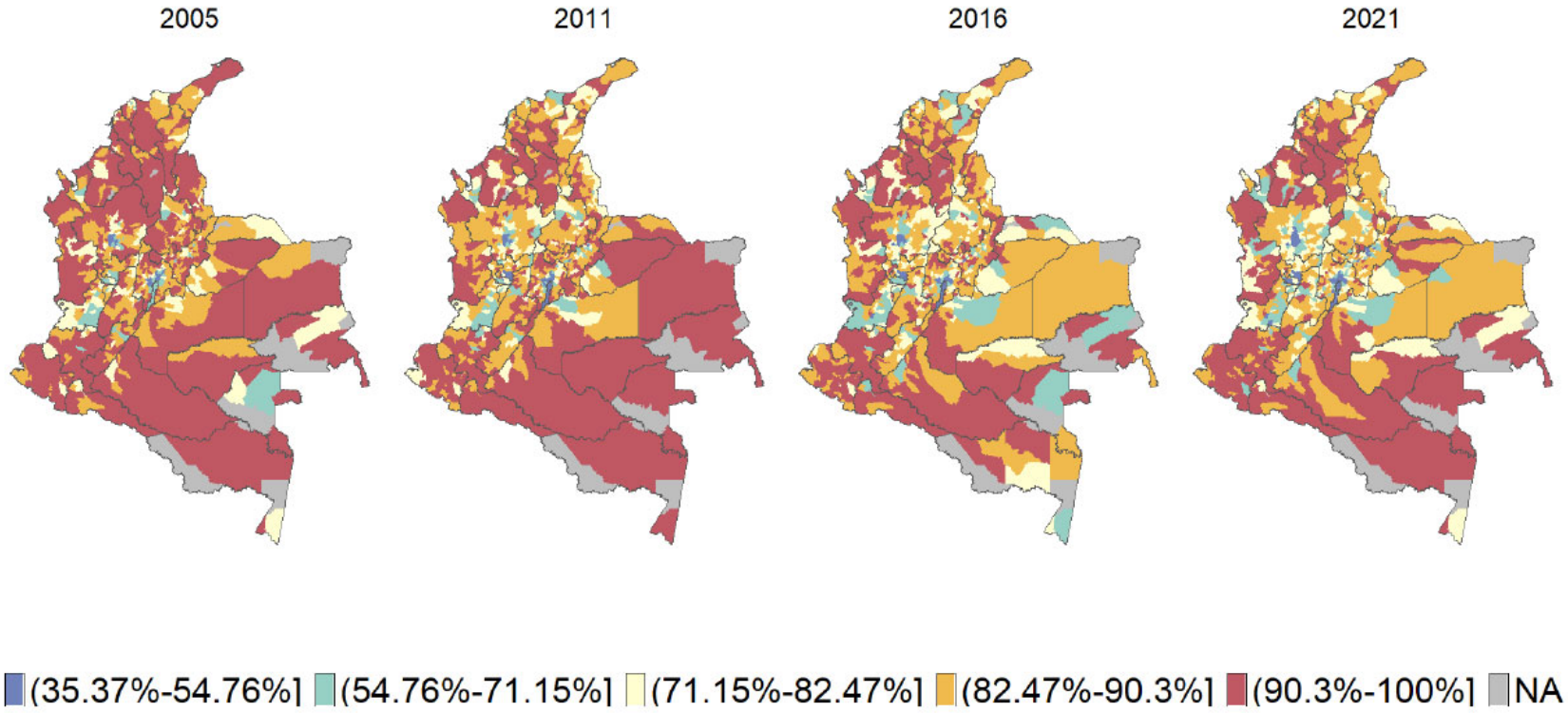
Variable	N	Mean	Standard deviation	Min	Max
		2011			
PILA Rate	1,118	4.8	5.2	0.1	46.8
Average Wages	1,118	777,203	263,981	275,093	1,839,899
Rural WA Ratio	1,118	58.2	23.9	0.1	100.0
Female WA Ratio	1,118	49.2	2.3	25.6	54.8
Pop. Subsidized Regime Ratio	1,103	73.4	21.2	0.1	140.0
Average SABER 11 Score	1,108	210.4	11.6	164.8	245.0
Average Spanish SABER 11 Score	1,108	42.9	3.3	29.0	53.6
Average Mathematics SABER 11 Score	1,108	43.2	3.6	23.0	54.0
Primary Sector Value Added	1,118	96.8	521.2	0.0	14,360.7
Secondary Sector Value Added	1,118	104.0	867.0	0.1	24,737.2
Tertiary Sector Value Added	1,118	301.9	3,476.2	0.4	110,786.2
Total Value Added Per Capita	1,118	1.1	2.0	0.1	43.2

Variable	N	Mean	Standard deviation	Min	Max
		2021			
PILA Rate	1,113	7.5	7.1	0.4	53.8
Average Wages	1,113	1,079,424	291,064	830,331	2,949,186
Virtual Vacancy Rate	1,077	0.9	2.1	0.003	27.2
Rural PET Ratio	1,113	55.5	23.8	0.05	100.0
Female WA Ratio	1,113	49.4	2.2	31.4	55.2
Pop. Subsidized Regime Ratio	1,113	63.8	19.7	0.5	137.1
Average SABER 11 Score	1,107	233.6	20.3	160.9	286.7
Average Spanish SABER 11 Score	1,107	49.5	3.9	34.2	60.5
Average Mathematics SABER 11 Score	1,107	47.5	4.7	30.8	60.3
Primary Sector Value Added	1,113	141.8	466.1	0.0	9,346.2
Secondary Sector Value Added	1,113	166.3	1,190.5	0.1	32,074.1
Tertiary Sector Value Added	1,113	659.7	7,435.8	1.2	235,361.8
Total Value Added Per Capita	1,113	1.7	2.0	0.2	23.4

Sources: DANE, PILA, SPE, Panel municipal CEDE, ICFES.

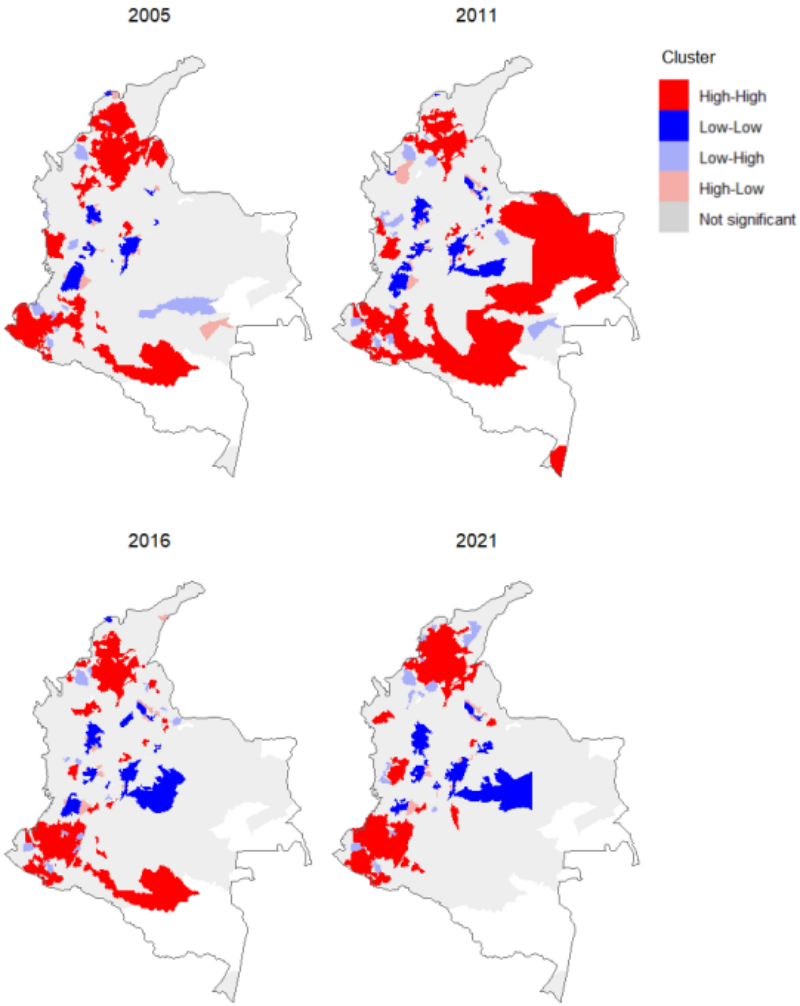
Results

Estimated labor informality rate at the municipality level – Fay-Herriot model



- There is high spatial concentration of labor informality. A center-periphery pattern arises.
- However, there has been a reduction in labor informality rate in most of the regions of the country.

Cluster analysis: Local indicators of spatial association



- We use a queen contiguity matrix for proximity, which considers municipalities as neighbors if they share a vertex or side.
- 2005 LISA indicators are directly computed using Census data.
- There are persistent hot-spots around the Caribbean region and the southern portion of the country.
- There is suggestive evidence that HH clusters are different than the other areas as they have lower quality of education (as per Saber 11 tests), newer political institutions and they farther from main wholesale markets.

Robustnes check: Importance of auxiliary variables

- We use a permutation approach to evaluate the importance of each covariate. This method measures the impact of the permutation of each characteristic on the model square error (MSE). Significance is quantified as the percentage change between the original mean square error (MSE) and the MSE after permutation.

	2011		2016		2021	
Ranking	Variable	Value	Variable	Value	Variable	Value
1	Subsidized regime proportion	2.41	Subsidized regime proportion	4.18	Subsidized regime proportion	2.80
2	PILA rate	1.24	PILA rate	2.47	PILA rate	2.52
3	Secondary sector (VA)	0.97	Virtual vacancies rate	2.18	Virtual vacancies rate	0.85
4	Percentage of women aged 24 to 54	0.73	Secondary sector AV	0.95	Distance to Bogota	0.77
5	Women WAP proportion	0.56	Percentage of women aged 24 to 54	0.88	Secondary sector AV	0.69
6	Percentage of men aged 24 to 54	0.55	Secondary sector AV	0.79	Percentage of men aged 24 to 54	0.69
7	Altitud	0.38	Percentage of men aged 24 to 54	0.71	Base salary	0.64
8	Base salary	0.38	Women WAP proportion	0.70	Distance to market	0.48
9	Rurality	0.37	Base salary	0.67	Distance to main market	0.48
10	Distance to main market	0.34	Rurality	0.60	Tertiary sector AV	0.44

Concluding remarks

- Labor informality at the municipality level is highly concentrated in Colombia and persistently high.
  - There are hot spot clusters around the Caribbean and the southern-Pacific region.
  - Such areas also exhibit particular characteristics, such as lower quality of education, less years since creation of the municipality and they are less connected from main wholesale markets.
- Notably, there is a reduction in informality after the pandemic concentrated in municipalities with lower levels of informality.
- The determinants of informality seem to change through years and across the space. Thus, regional/national models might be necessary.
- The transformed FH model showed a slightly better performance compared to other models, such as mixed effects random forest models (MERF), but there is still room for testing other ML approaches.

Annex: Mixed-effects random forest estimator (MERF)

- Following Krenmmair and Schmid (2022), a model with a structure similar to FH, but with a flexible functional form is followed:

  - $\hat{\theta}_m^{RF} = f(\hat{X}_m) + Z_m\hat{\vartheta}_m + \epsilon_m$
  - In this methodology a form F as a function of X is followed using a random forest algorithm.
  - The uncertainty estimate (MSE) is obtained via nonparametric bootstrap.

- The MERF algorithm follows these steps:
  - Initialise $b = 0$, while setting the random effect parts to zero (ϑ).
  - Set $b *= b + 1$ and update $f(X_m)$ and ϑ.
  - Obtain $y_b^* = y - Z\hat{\vartheta}_b$
  - Follow a standard random forest algorithm for dependent variable y_b^* and X.
  - Obtain out of sample predictions $f(\hat{X}_m)$ and fit it in the standard LMM function.
  - Pull out the variances components to get the random effects ϑ.
  - ** $\vartheta_b = H_b Z' V_b^{-1}(y - f(X)_b^{OOB})$
  - loop steps $(b + 1)$ until convergence.

- The random forest algorithm partitions the data such that a learning method is ensembled based on multiple classifications.

ESPE - labor and business informality

Editor:

- Andrea Otero Cortés (CEER)

Co-authors :

Karina Acosta, Luis Eduardo Arango, Danilo Aristizábal (Uniandes), Oscar Ávila, Oscar Becerra (Uniandes), Cristina Fernández (Fedesarrollo), Luz Adriana Flórez, Luis Armando Galvis, Anderson Grajales, Catalina Granda, Franz Hamann, Juliana Jaramillo, Carlos Medina, Jesús Morales, Leonardo Morales, Juan José Ospina, Christian Posso, José David Pulido y Mario Ramos

Thank you!!!