DOCUMENTOS DE TRABAJO SOBRE

ECONOMÍA REGIONAL Y URBANA

Informalidad municipal en Colombia

Por:
Karina Acosta
Juliana Jaramillo-Echeverri
Daniel Lasso
Alejandro Sarasti-Sierra

Núm. 327
Junio, 2024

Centro de Estudios Económicos Regionales (CEER) - Cartagena

Informalidad municipal en Colombia

Karina Acosta[†] Juliana Jaramillo-Echeverri[‡] Daniel Lasso[§]
Alejandro Sarasti-Sierra[¶] [‖]

Resumen

Se estima que más del 50 % de la población laboral en Colombia pertenece al sector informal, un fenómeno persistente durante las últimas tres décadas. A pesar de la amplia literatura sobre la informalidad laboral y sus determinantes a nivel nacional o en las principales áreas urbanas, las tasas de informalidad municipales permanecen inexploradas en el país, debido a la falta de disponibilidad y calidad de los datos. En general, la información necesaria para medir la informalidad subnacional, ya sea a través del tamaño de la empresa, la afiliación al régimen contributivo o la existencia de un contrato escrito, es escasa o incompleta, lo que dificulta una estimación directa. En este trabajo se propone un ejercicio de medición para avanzar en el estudio de la informalidad en Colombia, estimando la informalidad laboral municipal entre 2005 y 2021. Los resultados muestran que, aunque la informalidad es persistentemente alta, está fuertemente concentrada. Además, se observa que, aunque la informalidad cayó paulatinamente entre 2005 y 2016 en todos los municipios, aquellos con tasas de informalidad más altas experimentaron un retroceso en estas ganancias en 2021.

Palabras clave: informalidad, estimaciones en áreas pequeñas, clústeres, LISA, Colombia
Clasificación JEL: J46, O17, O54, R23, C38

[†]Investigadora Júnior, Centro de Estudios Económicos Regionales (CEER). Contacto: kacostor@branrep.gov.co

[‡]Investigadora Júnior, Centro de Estudios Económicos Regionales (CEER). Contacto: jjaramec@branrep.gov.co

[§]Facultad de Economía, Universidad de Los Andes. Contacto: df.lasso@uniandes.edu.co

[¶]Facultad de Economía, Escuela de Ingenieros de Antioquia. Contacto: sarasti1907@gmail.com

[‖]Agradecemos a Luz Adriana Flórez, Jaime Bonet, Javier Pérez, Luis A. Galvis y a los participantes del seminario de la Gerencia Técnica por sus valiosos comentarios. Agradecemos también la asistencia de Adriana Ortega por su asistencia en esta investigación. Todos los errores y omisiones son nuestros.

Municipal informality in Colombia

Karina Acosta[†] Juliana Jaramillo-Echeverri[‡] Daniel Lasso[§]
Alejandro Sarasti-Sierra[¶‖]

Abstract

It is estimated that more than 50 % of the labor force in Colombia belongs to the informal sector, a persistent phenomenon over the last three decades. Despite extensive literature on informality and its determinants at the national level or in the main urban areas, municiapl informality rates remain unexplored in the country due to the lack of availability and quality of data. In general, the information necessary to measure sub-national informality, whether through firm size, affiliation to social security, or the existence of a written contract, is scarce or incomplete, making direct estimation difficult. This study proposes a measurement exercise to contribute to the study of informality in Colombia, estimating municipal informality between 2005 and 2021. The results show that, although informality is persistently high, it is strongly concentrated. Furthermore, it is observed that, although informality gradually declined between 2005 and 2016 in all municipalities, those with higher informality rates experienced a setback in these gains in 2021.

Keywords: informality, small area estimation, clusters, LISA, Colombia
JEL Codes: J46, O17, O54, R23, C38

[†]Junior Researcher, Center for Regional and Economic Studies. Contact info: kacostor@branrep.gov.co

[‡]Junior Researcher, Center for Regional and Economic Studies. Contact info: jjaramec@banrep.gov.co

[§]Faculty of economics, Universidad de Los Andes. Contact info: df.lasso@uniandes.edu.co

[¶]Faculty of economics, Escuela de Ingenieros de Antioquia. Contact info: sarasti1907@gmail.com

[‖]We thank Luz Adriana Flórez, Jaime Bonet, Javier Pérez, Luis A. Galvis, and the participants of the internal seminar for their valuable comments. We also thank Adriana Ortega for her research assistance. All errors and omissions are ours.

1. Introducción

¿Cuánto es la informalidad laboral en los municipios en Colombia y cómo ha cambiado a lo largo del periodo 2005-2021? En 2018 el Departamento Administrativo Nacional de Estadística (DANE) estimaba que la informalidad laboral -medida como la no afiliación a seguridad social- en el país era cercana al 60 % (figura 1), mientras que para el trimestre de octubre a diciembre del 2023, el DANE calculaba que la proporción de ocupados informales en Colombia era de 55.5 %. Esta cifra se ha mantenido más o menos constante desde la década de 1980 y posiciona al país como uno de los países con mayores tasas de informalidad en múltiples dimensiones. No obstante, esta cifra agregada hace invisible la heterogeneidad de las regiones en Colombia. En la literatura, la mayoría de las investigaciones se han centrado en medir y estudiar la informalidad nacional o en las principales áreas urbanas (Bernal, 2009; Cárdenas y Mejía, 2007). Por su parte, las tasas de informalidad laboral municipales permanecen inexploradas debido a la falta de disponibilidad y a la calidad de los datos. Esto en la medida en que los datos sobre informalidad subnacional son escasos o incompletos -principalmente por la dificultad de realizar muestreos- lo que dificulta una estimación directa. El presente trabajo contribuye al estudio de la informalidad laboral en Colombia, estimando la distribución espacial de la informalidad municipal para los años 2005, 2011, 2016 y 2021.[1]

Dado que la informalidad laboral se puede medir en función de diferentes variables, como el tamaño de empresa, afiliación al régimen contributivo y contar con un contrato laboral escrito, entre otros, en este documento se define como informales a los trabajadores que no cumplan con el criterio de estar cotizando a un fondo de pensiones.[2] Esta definición se toma de la principal fuente de indicadores laborales en Colombia, la Gran Encuesta Integrada de Hogares (GEIH), realizada con una periodicidad mensual. La GEIH permite la obtención de información agregada nacional, departamental, cabecera-resto y las principales ciudades del país.[3]

No obstante, esta encuesta es insuficiente para obtener estimaciones sobre otras subpoblaciones (dominios) de interés como municipios con muestras no representativas (Petersen, Minkkinen, y Esbensen, 2005). Por otra parte, los censos nacionales que son la principal fuente para obtener indicadores municipales confiables dada su

[1] Se realizan ejercicios cuatrianulaes de las estimaciones, por eso se identifican estos años.

[2] Para una explicación detallada sobre las diferentes definiciones de informalidad laboral véase Bernal (2009); LaboUR (2018) y la tabla 5 en el apéndice.

[3] La GEIH es una encuesta realizada por el Departamento Administrativo Nacional de Estadística (DANE).

representatividad y alcance, presentan la desventaja de ser costosos, tienen una alta infrecuencia y en muchos casos no cuentan con preguntas que permitan estimar indicadores socioeconómicos de interés, como la informalidad. Otras fuentes, como la Planilla Integrada de Liquidación de Aportes (PILA), donde se registran todos los aportes a seguridad social que hacen los empleados formales, presenta errores en los registros, en particular, en algunos municipios hay problemas con los códigos municipales o el número de registrados en un municipio dado es muy pequeño o muy grande con respecto a la población en edad de trabajar. Esto dificulta el uso de estas fuentes como una medida exacta de la informalidad municipal, pero resulta en una variable con alta capacidad predictiva de la misma.

Figura 1: Proporción estimada de ocupados informales en Colombia

Fuentes: Elaboración propia basada en la información del DANE (2018)

Como alternativa para la estimación de dominios invisibles -la informalidad municipal- con una adecuada precisión, han surgido un conjunto de herramientas estadísticas denominadas "Estimación de Área Pequeña" (SAE por sus siglas en inglés). Estos modelos se clasifican en dos grandes grupos: modelos de áreas y de unidad (Rao y Molina, 2015). Los modelos de área se basan en el análisis de zonas geográficas como distritos censales, municipios, departamentos, entre otros. Por su parte, los modelos de unidad apalancan la información de microdatos para generar estimaciones directas de las características de individuos dentro del área de análisis.

Este estudio emplea el conjunto de herramientas SAE con modelación a nivel de áreas para estimar la informalidad municipal en diferentes años en Colombia. Estas permiten obtener estimaciones indirectas de la variable de interés (informalidad en este caso) a partir de información auxiliar proveniente de diferentes fuentes como la PILA, el Servicio Público de Empleo (SPE), cuentas nacionales, proyecciones poblacionales, entre otros. Asimismo, la precisión de los estimadores dependen de la calidad

de las variables auxiliares y de una adecuada especificación. Por esta razón este estudio compara el comportamiento de dos modelos área-SAE: un modelo basado en los clásicos modelos lineales mixtos (LMM por sus siglas en inglés) siguiendo la metodología de Schmid, Bruckschen, Salvati, y Zbiranski (2017) y modelos más recientes que incorporan algoritmos de aprendizaje de máquinas, utilizando la propuesta de Krennmair y Schmid (2022) ajustada. Finalmente, este artículo realiza un análisis de asociación espacial de la informalidad para identificar conglomerados de municipios con alguna cercanía con baja y alta informalidad. De esta forma se podrá analizar la distribución espacial de los municipios y generar alertas para las autoridades territoriales. Para realizar este análisis se hace uso de la metodología de indicadores locales de asociación espacial (LISA por sus siglas en inglés)(Anselin, 1995).

Este artículo contribuye al estudio de la informalidad laboral en Colombia al estimar la informalidad municipal para cuatro años entre el 2005 y el 2021 y analiza los cambios, tendencias y patrones. Esto permite entender la heterogeneidad subnacional que ocultan los indicadores nacionales o agregados departamentales. Igualmente, contribuye al comparar los resultados en un caso práctico de los modelos tradicionales de la literatura de SAE con nuevos modelos adaptados de la literatura de Aprendizaje de Máquinas. Asimismo, deriva análisis de estadísticos espaciales para identificar las regiones y conglomerados municipales que conjuntamente tienen niveles de informalidad altos, lo cual resulta informativo para el desarrollo de política pública. Para esto, consolidamos un panel con información municipal que cuenta con variables demográficas, socioeconómicas y geográficas, al igual que las estimaciones directas de informalidad provenientes de las encuestas de hogares.

Los resultados indican una informalidad laboral por promedio simple en el 2021 del 77 %[4] y un promedio ponderado (por población) del 61 %. No obstante, se encuentra una destacable heterogeneidad en la distribución de la informalidad en los municipios de Colombia. Se estima que la informalidad municipal varía entre el 25 % y el 97 %, con una distribución asimétrica y sesgada hacia una mayor informalidad. Del mismo modo, se encuentra que la informalidad en general ha disminuido entre el 2005 y el 2021. En particular, se calcula una reducción -en promedio y estadísticamente significativa- de 12.5 puntos porcentuales en la informalidad entre 2005 y 2021. La mayor parte de esta disminución sucedió entre los años 2005 y 2016. En contraste, para 2021 se encuentra una tendencia regresiva para los municipios más informales, quienes parecen haber aumentado la informalidad post-pandemia. Del mismo modo,

[4]Promedio simple de las estimaciones municipales.

el análisis de los clústeres espaciales (LISA) identifica municipios contiguos con altos/bajos niveles de informalidad. Principalmente, en el suroccidente colombiano y el sur de la región Caribe. Igualmente, la distribución de la informalidad laboral en Colombia sigue un patrón centro-periferia, con un clúster de baja informalidad en la región andina.

La siguiente sección presenta una breve revisión de la literatura sobre informalidad en Colombia y una justificación del la definición de informalidad que se usará a lo largo de este documento. La segunda sección describe las variables auxiliares que se consideran tienen poder predictivo para estimar la informalidad. La tercera sección presenta la metodología tanto de SAE como de LISA. Finalmente, la cuarta sección discute los resultados de las estimaciones temporales y espaciales de la informalidad y la quinta presenta las conclusiones.

2. Revisión de literatura: informalidad laboral en Colombia

Una de las características más distintivas del mercado laboral colombiano es su elevado nivel de informalidad. En las últimas tres décadas, se ha observado que las tasas de informalidad oscilan entre el 50 % y el 60 % en las 23 principales ciudades y zonas metropolitanas del país, siendo además un fenómeno con una fuerte persistencia (Bustamante, 2011; Ruffer y Knight, 2007; Mora y Muro, 2017). Estos porcentajes se calcularon conforme a la definición operativa que el DANE utilizó hasta el año 2018 con información recogida por la GEIH. La informalidad era definida como la proporción de personas ocupadas que cumplían con alguno de los siguientes criterios: trabajar como empleado o como cuenta propia en establecimientos con menos de 5 personas, incluido el empleador o socio y excluyendo a los profesionales independientes; ser jornalero o peón; ejercer labores sin remuneración o ser empleado doméstico (DANE, 2009).

Sin embargo, definir el concepto de informalidad puede resultar complejo debido a la amplitud de fenómenos, enfoques y situaciones que lo pueden afectar. De acuerdo con Ariza y Retajac (2021), el trabajo informal comprende "actividades económicas que no están reguladas por el estado, a menudo caracterizadas por la ausencia de contratos formales, seguridad social y protecciones legales". Esta definición subraya la naturaleza desfavorable del empleo informal, destacando su falta de supervisión regulatoria y de redes de seguridad social.

Celín Camargo y cols. (2023) proporcionan una perspectiva más amplia, sugiriendo que la informalidad puede ser vista a través de dos lentes principales: el enfoque legalista, que se centra en el cumplimiento de las condiciones legales del empleo, y el enfoque estructural (utilizado por el DANE hasta 2018), que considera la informalidad como un subproducto de estructuras económicas y sociales que excluyen a ciertas poblaciones de la economía formal, centrando su atención en las características de las empresas, como su tamaño o sector productivo. Esta doble perspectiva ayuda a comprender la complejidad de la informalidad, indicando que no es meramente una cuestión legal, sino también profundamente arraigada en el tejido socioeconómico. De igual manera, Galvis-Aponte (2012) continúa esta idea comparando ambos enfoques y concluyendo que existe una correlación mayor a 0.8 entre ellos. Es decir, que ambas medidas están fuertemente asociadas. A la hora de obtener métricas de informalidad, sin embargo, el enfoque estructuralista subestima la informalidad sobre todo debido a la importancia que se le da a los cuenta propia.

En 2018, el DANE ajustó la definición de informalidad, considerando informales a los trabajadores sin cotizaciones en salud y pensiones, incluyendo trabajadores familiares sin remuneración, trabajadores por cuenta propia y patrones o empleadores en unidades de hasta cinco personas (DANE, 2018). Este ajuste fue motivado por el crecimiento de la evidencia en contra de un enfoque puramente estructuralista. Múltiples estudios demostraron que el enfoque estructuralista no estaba correlacionado con puntos clave de la idea generalizada de informalidad, como la calidad del empleo, la vulnerabilidad que produce y la adhesión a la normativa nacional. Por ejemplo, Bernal (2009) en su investigación comparó 27 definiciones de informalidad y concluyó que la basada en cotización a salud o pensión por parte del trabajador es la que generaba mayor correlación con todos los criterios de informalidad.[5]

Continuando la misma línea de investigación, Guataquí, García, y Rodríguez (2011) formularon dos definiciones de informalidad: una fuerte y una débil. La fuerte define como informal al trabajador que no recibe todos los beneficios que se encuentran en la ley. La débil considera informal al trabajador sin cotización a salud. Con estas definiciones llegó a conclusiones y tasas de informalidad muy similares a las halladas por Bernal (2009). Adicionalmente, concluyó que el enfoque del DANE tenía una fuerte contradicción, debido a que pretende medir un elemento de la demanda laboral mediante una herramienta de medición de oferta laboral (la GEIH).

[5]Véase también Flórez (2002).

Aunque para el 2018 hubo un cambio de definición, aún existe una importante limitación para las mediciones de informalidad en el país. Esta encuesta por la cual se obtienen datos del mercado laboral (GEIH) solo es estadísticamente representativa para las zonas urbanas de los 23 municipios más grandes del país, dejando de lado a gran parte de la población colombiana. Estudios previos como el hecho por el laboratorio laboral de la universidad del Rosario han demostrado que algunas zonas rurales del país presentan tasas de informalidad hasta 30 puntos porcentuales más altas, llegando hasta el 80 % (LaboUR, 2018). Esta investigación tiene como objetivo principal afrontar este problema empleando metodologías que permitan aproximar las tasas de informalidad subnacionales en la amplitud del territorio nacional.

3. Metodología

En este documento utilizamos un conjunto de técnicas estadísticas de estimación en áreas pequeñas (SAE, por sus siglas en inglés). Estas técnicas consisten en la estimación de parámetros en subpoblaciones de interés cuando la información disponible por encuestas no son representativas para estos grupos. Las áreas no solamente hacen referencia a zonas geográficas, como un municipio o una unidad censal, también puede indicar clases por características demográficas como sexo, edad, etnia, entre otros. Usualmente, las encuestas están diseñadas para tener una muestra representativa para áreas o grupos específicos. Por ejemplo, por medio de la GEIH se pueden obtener estimaciones para las 23 principales ciudades, regiones y algunos departamentos de Colombia, pero su muestra no es suficiente para hacer estimaciones de todos los municipios de Colombia ni unidades censales.

Dentro del campo de SAE se identifican dos grandes grupos de modelos: modelos de área y modelos de individuo. En el primero la unidad de análisis son las áreas, mientras que en el segundo se modelan hogares o individuos. En este documento realizamos estimaciones del primer tipo. A los modelos de referencia tipo área-SAE también se les conocen como estimadores Fay-Herriot porque fue introducido inicialmente por los autores Fay y Herriot (1979). Además de estimar el modelo básico de Fay-Harriot, este documento estima un modelo Fay-Herriot transformado, como se propone en Schmid y cols. (2017), y un modelo de bosques aleatorios de efectos fijos, formulado por Krennmair y Schmid (2022). Uno de los valores agregados de este documento es la comparación de los estimadores resultantes de estos modelos, donde el modelo clásico Fay-Harriot se propone como un punto de referencia (*benchmark*). El

objetivo último de este estudio es estimar las tasas de informalidad municipal para los años 2011, 2016 y 2021 y analizar los cambios temporales comparándolos con la información censal disponible para el 2005. Sin embargo, previo a las estimaciones para todos los años, en la Section 5.1 comparamos los resultados de las metodologías descritas a continuación para 2021. La metodología con las mejores estimaciones se aplica posteriormente a los años restantes.

3.1. Estimadores Fay-Herriot

Asumamos que una población P se compone de a áreas pequeñas mutuamente excluyentes en las cuales se distribuye su población total N. Adicionalmente, se tiene información de n unidades de muestreo de la población N. Las unidades muestreadas se indexarán con m y las no muestreadas con r. Asumiendo la existencia de una variable objetivo continua y, y_{mi} corresponde a la variable de una unidad $i \in N$ en una área $m \in a$ y sus pesos de muestreo correspondientes son w_{mi}. Un estimador de la media poblacional de la variable y para cada área m estará dado por:

$$\hat{\theta}_m^{\text{Directo}} = \frac{\sum_{i=1}^{n_m} w_{mi} y_{mi}}{\sum_{i=1}^{n_m} w_{mi}} \tag{1}$$

Por definición, $\hat{\theta}_m^{\text{Directo}}$ estima valores sesgados de las áreas para las cuales la muestra no es representativa. Para obtener estimadores corregidos de las áreas observadas, el modelo de Fay-Harriot propone una estimación de dichos estimadores en dos etapas. Asimismo, se obtiene una estimación sintética para las áreas no muestreadas. La primera etapa o modelo de muestreo obtiene estimadores a partir de las estimaciones directas. La segunda etapa (modelo de enlace o linking model) alimenta la especificación con covariables a nivel de área. La primera etapa se define por:

$$\hat{\theta}_m^{\text{Directo}} = \hat{\theta}_m + \epsilon_m \tag{2}$$

La segunda etapa se determina por una regresión lineal a nivel de área:

$$\hat{\theta}_m = X_m^T \beta + u_m \tag{3}$$

X_m^T denota un vector de variables auxiliares relevantes para áreas que son explicativas de los valores directos observados. La especificación asume que los errores

de muestreo ϵ_m y los efectos aleatorios u_m son independientes y siguen una distribución normal, $\epsilon_m \sim N(0, \sigma^2_{\epsilon_m})$ y $u_m \sim N(0, \sigma^2_u)$. De (2) y (3) se deriva el modelo mixto lineal:

$$\hat{\theta}^{\text{Directo}}_m = X^T_m \beta + u_m + \epsilon_m \tag{4}$$

El mejor predictor empírico lineal insesgado (BLUP, por sus siglas en inglés) de $\hat{\theta}_m$ (4) con el modelo Fay-Harriot (FH) está dado por:

$$\hat{\theta}^{\text{FH}}_m = \hat{\gamma}_m \hat{\theta}^{\text{Directo}}_m + (1 - \hat{\gamma}_m) X^T_m \beta \tag{5}$$

En (5), $\hat{\gamma}_m$ define el valor de contracción (*shrinkage factor*) de una área m y se deriva de $\sigma^2_u(\sigma^2_u + \sigma^2_{\epsilon_m})^{-1}$. Por su parte, los efectos aleatorios estimados estarán dados por $\hat{u}_m$ y se definen como $\hat{\gamma}_m(\hat{\theta}^{\text{Directo}}_m - X^T_m \beta)$. Cabe resaltar que para las áreas sin muestra (r) el estimador FH descrito por (5) no está definido, debido a que no se puede obtener un estimador directo. Para estas áreas, los estimadores sólo se obtienes de la fracción sintética de (5), $\hat{\theta}^{\text{FH}}_{m,r} = X^T_m \beta$.

En la literatura se han sugerido varias opciones para obtener el error cuadrático medio del estimador $\hat{\theta}^{\text{FH}}_m$. Debido a los problemas de sobre-contracción de las propuestas iniciales, como se sugiere en Li y Lahiri (2010), en este documento utilizamos el estimador de máxima verosimilitud del componente de varianza sugerido por estos mismos autores. Li y Lahiri (2010) proponen una verosimilitud ajustada de σ^2_u: $L_{adj}(\sigma^2_u) = \sigma^2_u L(\sigma^2_u)$. Para este documento, $L(\sigma^2_u)$ representa la función del perfil de verosimilitud, pero este también puede representar la función de verosimilitud residual.

3.2. Estimadores Fay-Herriot transformado

Una de las limitaciones del modelo Fay-Harriot clásico es que se diseñó para variables continuas. Por ello, estimaciones con el modelo FH no producen necesariamente resultados acotados dentro de un rango. Este hecho es relevante al trabajar con variables restringidas dentro de intervalos específicos como la proporción de trabajadores en condición de informalidad, que se encuentra dentro de un rango $[0, 1]$. Así, es indeseable que las estimaciones estén fuera de estos valores lógicos de la proporción. En el contexto de modelos área-SAE se han propuesto diferentes alternativas

para la estimación de proporciones; todas asociadas con la transformación de la variable a modelar (Franco y Bell, 2015; Liu, Lahiri, y Kalton, 2014; López-Vizcaíno, Lombardía, y Morales, 2015)

En este estudio seguimos el modelo sugerido por Schmid y cols. (2017) (Arc-FH), donde se realizan cuatro pasos para obtener las estimaciones de área: (i) se transforman los estimadores directos con una función seno inversa, $\nu_m = f(\hat{\theta}_m^{direct}) = \sin^{-1}(\sqrt{\hat{\theta}_m^{direct}})$, (ii) la varianza muestral de ν_m se aproxima por $\sigma^2_{\epsilon_m} = 1/4\tilde{n}_m$. $\tilde{n}_m$ representa la muestra efectiva, o división de la muestra por un estimado de efecto de diseño, (iii) se estima la ecuación clásica de FH (ecuación 5) para ν_m y $\sigma^2_{\epsilon_m}$. (iv) Los estimadores obtenidos del paso (iii) se transforman nuevamente a su escala original:

$$\hat{\theta}_m^{FH-transf} = f^{-1}(\hat{\theta}_m^{FH}) = \sin^2(\hat{\theta}_m^{FH}) \qquad para\ m\ =\ 1, ..., a. \tag{6}$$

Finalmente, los estimadores $\hat{\theta}_m^{FH-transf}$ transformados se complementan con los intervalos de confianza de las estimaciones "transformados hacia atrás"(tasa de informalidad), los cuales se obtienen a través de remuestreo (*bootstrap*) como se presenta en Casas-Cordero Valencia, Encina, y Lahiri (2016).[6]

3.3. Estimadores con bosques aleatorios de efectos mixtos

Entre las críticas a los modelos clásicos de estimación SAE presentados anteriormente, se encuentra que estos siguen una especificación lineal mixta (LMM). Por lo cual, estos son sensibles a los errores de especificación de las variables auxiliares y a las formas funcionales que a priori se asumen toman estas mismas. Alternativamente se han introducido propuestas que involucran métodos de aprendizaje de máquinas que utilizan modelos no-lineales y no-paramétricos. Entre estas opciones resalta los bosques aleatorios de efectos mixtos (MERF por sus siglas en inglés) como se presenta en Krennmair y Schmid (2022). Los autores, a través de simulaciones evidencian que, al comparar los estimadores de MERF con los métodos tradicionales, la flexibilidad de los árboles permite una precisión entre 40 y 50 % mayor de las estimaciones, especialmente, ante procesos no lineales en los datos que pueden ser sujetos de sesgos de especificación.

[6]Para mayores detalles sobre la metodología, la transformación y las propiedades del estimador, referirse a Schmid y cols. (2017).

Krennmair y Schmid (2022) proponen un modelo con una estructura similar a los modelos clásicos de LMM como se introduce en Battese, Harter, y Fuller (1988) para estimar medias de áreas, siguiendo la siguiente forma funcional:

$$\hat{\theta}_m^{RF} = f(\hat{X}_m) + Z_m\hat{\vartheta_m} + \epsilon_m \tag{7}$$

Donde $f(X_m)$ representa una relación $f()$ entre las variables auxiliares (X) y $\hat{\theta}_m$. Por su parte, $Z_m\vartheta_m$ describe la parte lineal del modelo y captura las dependencia por efectos aleatorios, donde Z_m define efectos aleatorios de áreas-municipios específicos. Adicionalmente, se asume que los efectos aleatorios de cada área (ϑ_m) y los errores de las mismas (ϵ_m) siguen las siguientes formas funcionales: $\vartheta_m \sim N(0, H)$ y $\epsilon_m \sim N(0, R)$. H y R representan sus respectivas matrices de varianza y covarianza.

A diferencia del modelo clásico, donde $f()$ se define por una función lineal $f(X) = X\beta$, en la propuesta sugerida por Krennmair y Schmid (2022) se especifica un bosque aleatorio.[7] Asimismo, la media estimada para áreas muestreadas (s) está descrita por la ecuación 7, mientras que la media de las áreas no muestreadas será $\hat{\theta}_m^{RF} = f(\hat{X}_m)$. A diferencia de estos autores, las unidades utilizadas en la estimación de (7) son municipios en lugar de individuos.

Para la estimación de $f()$, el algoritmo de efectos mixtos para bosques aleatorios sigue dos pasos en ciclo hasta que el algoritmo llega a converger: i) estima la función de bosques aleatorios convencional, asumiendo que los términos de los efectos aleatorios son correctos (fija ϑ_m en cero) y ii) estima efectos aleatorios utilizando "predicciones fuera de bolsa"(*out-of-bag*) resultante del bosque aleatorio.[8] Específicamente, las predicciones fuera de bolsa utilizan las observaciones no utilizadas para la construcción del sub-árbol de cada bosque.[9]

3.4. Indicadores locales de asociación espacial – LISA

Para determinar si la informalidad se concentra en ciertos municipios del país medimos la autocorrelación espacial. Esta estadística mide el grado de dependencia entre las observaciones en el espacio geográfico y proporciona una medida de la

[7] Para una revisión de bosques aleatorios véase Biau y Scornet (2016).

[8] Para una descripción detallada del algoritmo véase Krennmair y Schmid (2022).

[9] Esta aproximación cuenta con dos hiperparámetros que tienen que ser ajustados para mejorar la predicción: i) el número de árboles y el número de candidatos para la separación de los nodos. Estos son calibrados utilizando validación cruzada

correlación de una variable en diferentes lugares del espacio. La autocorrelación puede ser positiva, negativa o nula. La positiva implica que lugares parecidos comparten valores similares de la variable de interés. La negativa, por el contrario, implica que los lugares tienden a estar rodeados de vecinos que tienen valores disímiles. El tercer resultado es la autocorrelación nula e indica que la ubicación de los puntos de datos están distribuidos aleatoriamente en el espacio.

Para detectar una autocorrelación espacial estadísticamente significativa en los datos, se utilizan los indicadores locales de asociación espacial - LISA (Anselin, 1995). La prueba LISA se basa en la hipótesis nula de que la variable de interés, en este caso las estimaciones de informalidad, se distribuye aleatoriamente entre los municipios. Cuando el "valor p" es estadísticamente significativo con un nivel de confianza del 95 %, se puede rechazar la hipótesis nula y establecer clústeres positivos o asociaciones de áreas negativas. Los lugares que tuvieron una autocorrelación espacial significativa se definen como clústeres alto-alto, que son lugares donde la informalidad toma valores altos y están rodeados de lugares con informalidad alta; clústeres bajo-bajo, que incluyen lugares con informalidad baja con vecinos con informalidad baja; clústeres alto-bajo, que son lugares con informalidad alta rodeados de lugares con informalidad baja, y clústeres bajo-alto, que son lugares con informalidad baja, rodeados de lugares con informalidad alta. En cuanto a la proximidad espacial, utilizamos una matriz de contigüidad reina para definir la vecindad entre lugares y elegimos una continuidad de primer orden. La matriz de contigüidad representa las relaciones de vecindad entre los municipios. La matriz reina considera como vecinos a todos los municipios que comparten un vértice o un lado, lo que significa que dos municipios se consideran vecinos si están adyacentes directamente o comparten al menos una esquina.

4. Datos

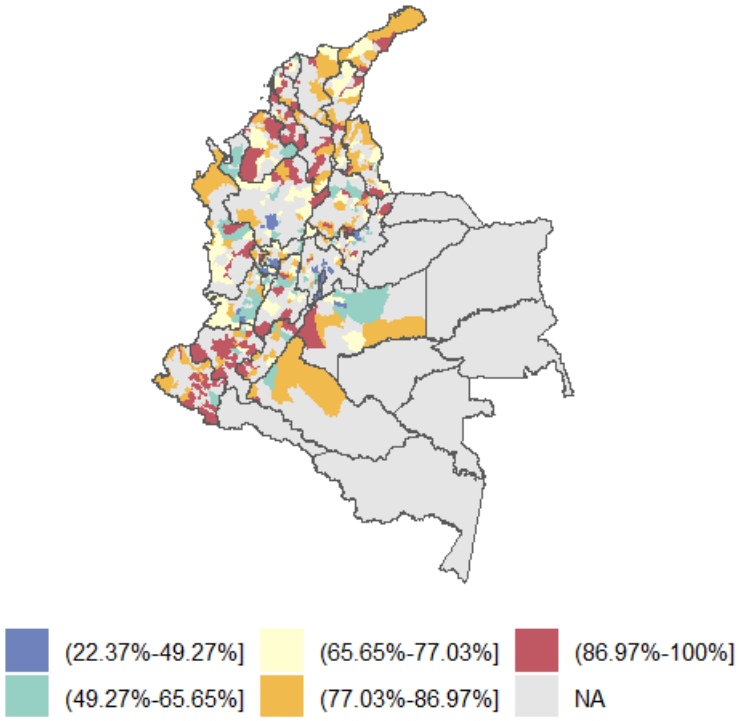
Para estimar las tasas de informalidad municipales construimos una base de datos que incorpora unidades con muestra de donde obtenemos nuestra variable objetivo (informalidad) y variables auxiliares de municipios. Las variables auxiliares incluyen información de aspectos geográficos, demográficos y económicos. Estos tres ejes porque consideramos esencial la inclusión de variables que no solo reflejen la desigualdad regional que define a Colombia, sino también las características sociales del país que, presumiblemente, están intrínsecamente ligadas al fenómeno de la informa-

lidad laboral. No obstante, la inclusión de algunas de estas variables está limitada a la disponibilidad de datos y a la calidad de estos. A continuación presentamos un resumen de estos grupos variables, así como la intuición detrás de la selección de estas.

4.1. Gran Encuesta Integrada de Hogares

Las unidades muestreadas se obtienen de la GEIH del DANE, la cual cuenta con una cobertura nacional que posibilita la obtención de resultados representativos tanto para zonas urbanas como rurales. La figura 2 muestra los municipios para los que la GEIH recoge información, no necesariamente estadísticamente representativa. Esta encuesta abarca cinco grandes regiones (Región Atlántica, Región Oriental, Región Central, Región Pacífica y Región Bogotá), 23 departamentos, las 13 principales ciudades con sus áreas metropolitanas y 11 ciudades intermedias y además la GEIH se recolecta en 436 municipios de Colombia. La GEIH se destaca como una de las fuentes de información más sólidas para la medición del mercado laboral en Colombia, dado que recopila datos a nivel individual y de hogar. Incluye preguntas relacionadas con la ocupación, el tipo de empleo y la afiliación a pensiones. Como se mencionó anteriormente, esta información permite estimar la informalidad en 24 áreas metropolitanas del país de manera representativa, y para otros municipios de forma no representativa. Estos datos son, por ende, la piedra angular para calcular la variable objetivo.

Figura 2: Informalidad observada en la GEIH, 2021



Fuentes: Departamento Administrativo Nacional de Estadística (DANE).

4.2. Variables geográficas

Las variables geográficas que compilamos tienen como objetivo capturar diferencias en ubicación, acceso a mercados, topografía y también instituciones. La variable *años desde creación del municipio* mide la duración en años desde la creación del municipio hasta el año de la observación, proporcionando una perspectiva temporal de su historia. La variable *altura* se refiere a la elevación sobre el nivel del mar del municipio, suministrando información topográfica. En Colombia la altura municipal varía desde un metro sobre el nivel del mar a los 3,350 metros sobre el nivel del mar. Por otro lado, *área* representa la extensión territorial del municipio en kilómetros cuadrados, indicando su tamaño geográfico. En cuanto a *distancia a capital del departamento*, *distancia a mercado principal* y *distancia a Bogotá*, son medidas de distancia lineal expresadas en kilómetros que representan la proximidad del municipio a la capital del departamento, al mercado mayorista de alimentos más importante y al centroide de Bogotá, respectivamente. Estas variables son esenciales para comprender la ubicación geográfica y la accesibilidad del municipio en relación con puntos clave en su entorno. Adicionalmente, se incluyen en la base de datos variables dummy que se-

ñalan la pertenencia del municipio a las respectivas regiones geográficas de Colombia. La tabla 1 presenta el resumen estadístico de los datos geográficos compilados.

Cuadro 1: Resumen estadístico variables geográficas

Variable	Media	Desviación estándar	Min	Max
Años desde creación del municipio	145.7	110.4	9	491
Altura	1,134.3	919.8	1	3,350
Área	1,019.3	3,203.8	15	65,674
Distancia a capital departamento	81.5	60.6	0.0	493.1
Distancia mercado principal	68.2	100.7	0.0	913.2
Distancia a Bogotá	320.9	192.6	0.0	1,228.1

Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Panel municipal CEDE.

4.3. Variables demográficas

Las variables demográficas tienen como objetivo capturar diferencias en el tamaño de la población, la estructura poblacional por género y edades, así como el proceso de urbanización en cada municipio.

La variable *población* hace referencia a las proyecciones poblacionales totales según el DANE. Esta cifra representa la cantidad absoluta de habitantes en el municipio. Además, se incluyeron los porcentajes de distribución por grupo de edad y género que proporcionan una visión detallada de la estructura demográfica del municipio. La *tasa de dependencia* es un indicador demográfico que refleja la relación entre la población que no está en edad laboral activa en comparación con la población en edad de trabajar. Se expresa como una razón, se calcula dividiendo la suma de la población menor de 15 años y mayor de 65 años entre la población en edad laboral (de 15 a 64 años) y se interpreta como el número de personas en edad de dependencia por cada 100 personas en edad productiva. Esta medida proporciona información sobre la carga que los dependientes (niños y personas mayores) representan para la fuerza laboral activa. Es importante anotar que en Colombia en algunos municipios de la Amazonía la tasa de dependencia alcanza hasta un 250 %. En cuanto a la variable *ruralidad*, se refiere al grado de ruralización del municipio. Esta variable se expresa como porcentaje y se calcula como la proporción de personas dentro del municipio que viven en áreas rurales. La tabla 2 presenta el resumen estadístico de los datos demográficos compilados para 2011, 2016 y 2021.

Cuadro 2: Resumen estadístico variables demográficas

Variable	N	Media	Desviación estándar	Min	Max
			2011		
Población	1,118	39,803	242,871	240	7,152,656
Tasa de dependencia	1,118	88.4	19.3	44.9	253.6
Ruralidad	1,118	57.8	23.9	0.1	100.0
			2016		
Población	1,120	41,797	249,327	262	7,300,918
Tasa de dependencia	1,120	82.8	19.2	43.4	228.6
Ruralidad	1,120	56.9	23.9	0.1	100.0
			2021		
Población	1,113	45,810	268,572	311	7,823,334
Tasa de dependencia	1,113	79.0	15.0	45.4	190.4
Ruralidad	1,113	55.1	23.8	0.05	100.0

Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Panel municipal CEDE.

4.4. Variables económicas

Finalmente, se compilan variables económicas que miden las condiciones productivas de un municipio, así como su fuerza laboral y su demanda y oferta de trabajo. La variable *tasa de PILA* representa la razón de empleados formales con respecto a la población total.[10] Esta variable se calcula como el promedio anual (del conteo mensual) de individuos registrados en la PILA dividido por la población total del municipio. Este indicador proporciona una aproximación a la formalidad del empleo en un municipio. La variable *salarió promedio*, que en este contexto podría representar el costo de la formalidad, se calcula como el promedio anual (del conteo mensual) de los salarios registrados en la PILA. Por su parte, la variable *tasa vacantes virtuales* indica la relación de vacantes de empleo con respecto a la población municipal, se obtiene mediante el conteo de vacantes registradas en el Servicio Público de Empleo (SPE). Este indicador mide la disponibilidad de empleo en una zona específica. Se calcula además la proporción de la población rural económicamente activa

[10] Alternativamente, estimamos la tasa PILA como la razón entre empleados formales registrados y la Población en Edad de Trabajar. Sin embargo, la razón por población mostró resultados predictivos ligeramente mejores, por lo que es nuestra tasa preferida

en comparación con la población total en edad de trabajar (*proporción PET rural*), así como la proporción de la población femenina económicamente activa (*Proporción PET mujeres*). De igual manera se calcula la razón entre la población afiliada al régimen subsidiado de salud sobre la población total del municipio (*Proporción régimen subsidiado*), como un indicador que permite hacer una aproximación de las personas no cotizantes. La variable *valor agregado* es una medida económica que representa el valor en millones de pesos creado por el municipio y proviene de la serie municipal del DANE que se encuentra dentro de las cuentas nacionales departamentales. A su vez, *VA sector primario*, *VA sector secundario* y *VA sector terciario* representan el valor agregado específico de las actividades primarias, secundarias y terciarias, respectivamente y provienen de la misma fuente. Estos indicadores ayudan a entender la contribución de cada sector a la economía municipal. Asimismo, *VA per cápita* se calcula dividiendo el valor agregado municipal entre la población municipal. La literatura de informalidad ha hecho manifiesta la importancia de otras variables económicas relevantes como el número de años de educación. No obstante, no fue posible introducir esta variable en las estimaciones, ya que solo se encuentra disponible para los años censales en Colombia. Asimismo, incluimos variables que aproximan a la educación como la información sobre los resultados en el SABER 11 y la cobertura educativa [11]. No obstante, estas variables no se incluyeron en las estimaciones finales por dos razones. Primero, la información no está disponible para áreas no municipalizadas, lo que implica la exclusión de 20 municipios de la muestra total. Segundo, los modelos no tuvieron un mejor rendimiento con la inclusión de estas variables.

La tabla 3 presenta el resumen estadístico de los datos demográficos compilados para 2011, 2016 y 2021.

[11] Corresponden a los datos del Sistema de Información Nacional de Educación Básica y Media (SINEB) y representa un indicador de la capacidad de la infraestructura escolar para atender a la población en edad escolar con base en las edades teoréticas.

Cuadro 3: Resumen estadístico variables económicas

Variable	N	Media	Desviación estándar	Min	Max
		2011			
Tasa de PILA	1,118	4.8	5.2	0.1	46.8
Salario promedio	1,118	777,203	263,981	275,093	1,839,899
Proporción PET rural	1,118	58.2	23.9	0.1	100.0
Proporción PET mujeres	1,118	49.2	2.3	25.6	54.8
Proporción régimen subsidiado	1,103	73.4	21.2	0.1	140.0
Puntaje promedio global SABER 11	1,108	210.4	11.6	164.8	245.0
Puntaje promedio lectoescritura SABER 11	1,108	42.9	3.3	29.0	53.6
Puntaje promedio matemáticas SABER 11	1,108	43.2	3.6	23.0	54.0
Valor agregado sector primario	1,118	96.8	521.2	0.0	14,360.7
Valor agregado sector secundario	1,118	104.0	867.0	0.1	24,737.2
Valor agregado sector terciario	1,118	301.9	3,476.2	0.4	110,786.2
Valor agregado per capita	1,118	1.1	2.0	0.1	43.2
		2016			
Tasa de PILA	1,120	6.1	6.5	0.1	58.7
Salario promedio	1,120	897,662	262,163	683,817	2,364,164
Tasa vacantes virtuales	1,022	0.5	1.2	0.004	12.5
Proporción PET rural	1,120	57.2	23.9	0.1	100.0
Proporción PET mujeres	1,120	49.3	2.2	28.8	54.9
Proporción régimen subsidiado	1,103	67.5	19.6	0.03	129.2
Puntaje promedio global SABER 11	1,111	246.1	18.6	185.4	305.3
Puntaje promedio lectoescritura SABER 11	1,111	50.1	3.3	37.0	60.9
Puntaje promedio matemáticas SABER 11	1,111	48.4	4.6	33.3	65.0
Valor agregado sector primario	1,120	87.5	266.4	0.0	5,759.5
Valor agregado sector secundario	1,120	152.3	1,133.9	0.1	31,963.1
Valor agregado sector terciario	1,120	463.3	5,262.6	0.7	167,314.2
Valor agregado per capita	1,120	1.3	1.3	0.2	16.4
		2021			
Tasa de PILA	1,113	7.5	7.1	0.4	53.8
Salario promedio	1,113	1,079,424	291,064	830,331	2,949,186
Tasa vacantes virtuales	1,077	0.9	2.1	0.003	27.2
Proporción PET rural	1,113	55.5	23.8	0.05	100.0
Proporción PET mujeres	1,113	49.4	2.2	31.4	55.2
Proporción régimen subsidiado	1,113	63.8	19.7	0.5	137.1
Puntaje promedio global SABER 11	1,107	233.6	20.3	160.9	286.7
Puntaje promedio lectoescritura SABER 11	1,107	49.5	3.9	34.2	60.5
Puntaje promedio matemáticas SABER 11	1,107	47.5	4.7	30.8	60.3
Valor agregado sector primario	1,113	141.8	466.1	0.0	9,346.2
Valor agregado sector secundario	1,113	166.3	1,190.5	0.1	32,074.1
Valor agregado sector terciario	1,113	659.7	7,435.8	1.2	235,361.8
Valor agregado per capita	1,113	1.7	2.0	0.2	23.4

Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE, Instituto Colombiano de Educación Superior (ICFES).

4.5. Pre-selección de variables auxiliares y restricción de muestra

Debido a la amplia disponibilidad de covariables (alrededor de 39) en los modelos Arc-FH, se restringió este conjunto siguiendo a varios autores como Ha, Lahiri, y Parsons (2014) y Schmid y cols. (2017). Para ello seguimos tres pasos. En primer lugar, restringimos la muestra al 50 % de municipios más grandes con base en su tamaño de muestra [12]. Con este subgrupo de la información estimamos una regresión simple con la variable $\sin^{-1}(\sqrt{y})$ como variable dependiente. Estas estimaciones asumen que los pesos de la muestra pueden ser ignorados debido a su baja variabilidad. Por último, se aplica un Criterio de Información de Bayes (BIC) con *stepwise* para penalizar la inclusión de variables y definir un modelo parsimonioso.[13] Estos pasos se siguieron para cada uno de los años independientemente, y los modelos finales resultan en un R^2 de alrededor de 80 % en 2021. Además, esta selección elimina covariables altamente correlacionadas entre sí.

En adición a la restricción de las variables de los modelos, también acotamos el número de municipios incluidos en las estimaciones. En la actualidad, Colombia cuenta con 1,122 municipios. No obstante, excluimos nueve municipios debido a la presencia de datos atípicos. En particular, utilizamos como referencia una de nuestras variables clave (conteo de personas en PILA/población) para la delimitación de los datos. Los municipios donde esta variable superaba el 100 % fueron eliminados. Asimismo, excluimos los municipios cuya razón PILA/población superaba a la ciudad capital más alta. Bajo estas condiciones, los municipios como Tarapacá, El Encanto, Mapiripana y Cacahual en Amazonas, Agua de Dios en Córdoba, San Joaquín y San Miguel en Santander, Puerto Carreño en Vichada y San Andrés y Providencia en el Archipiélago se excluyeron de las estimaciones.

Adicionalmente, con el propósito de mantener la consistencia en el tiempo de los resultados que se presentan a continuación, se procedió a excluir un conjunto de 20 zonas que carecían de información sobre el número de personas inscritas en el régimen subsidiado para los años 2011 o 2016.[14] La ausencia de estos datos compro-

[12] La muestra se restringe asumiendo que la variabilidad muestral de los estimadores directos de los municipios con mayor muestra es reducida, lo cual reduciría errores de pre-selección de modelos basados en técnicas estándar.

[13] En la aplicación de BIC utilizamos la función step del paquete emdi en R.

[14] Para el 2011 estos municipios y áreas no municipalizadas son: Tuchín en Córdoba, Puerto Alegría, Miritíi - Paraná en Amazonas y La Guadalupe, Cachahual y Morichal en Guainía. Para el 2016 estos son: La Chorrera, La Pedrera, La Victoria, Puerto Arica y Puerto Santander en Amazonas, Barrancominas, San Felipe, Puerto Colombia, y Pana Pana en Guainía, y Yavaraté en Vaupés.

metía la aplicabilidad de los métodos de estimación de la informalidad propuestos en este estudio. En consecuencia, el análisis se llevó a cabo utilizando una muestra de 1,103 municipios para los años 2011 y 2016, y de 1,113 municipios para el año 2021. No obstante, a fines de mantener una comparabilidad consistente, los resultados se presentan considerando únicamente 1,093 municipios para los cuales tenemos información durante todo el periodo.

5. Resultados

5.1. Comparación de modelos

La figura 3 presenta los resultados de las estimaciones de informalidad municipal para el 2021. El panel A reporta las estimaciones usando el modelo Fay-Herriot transformado, el panel B las estimaciones usando bosques aleatorios con efectos fijos (MERF) y en el panel C se muestra la diferencia en las predicciones entre ambos modelos. Las estimaciones del modelo Fay-Harriot convencional fueron excluídas debido a que se encontraban fuera de los rangos esperados. [15]

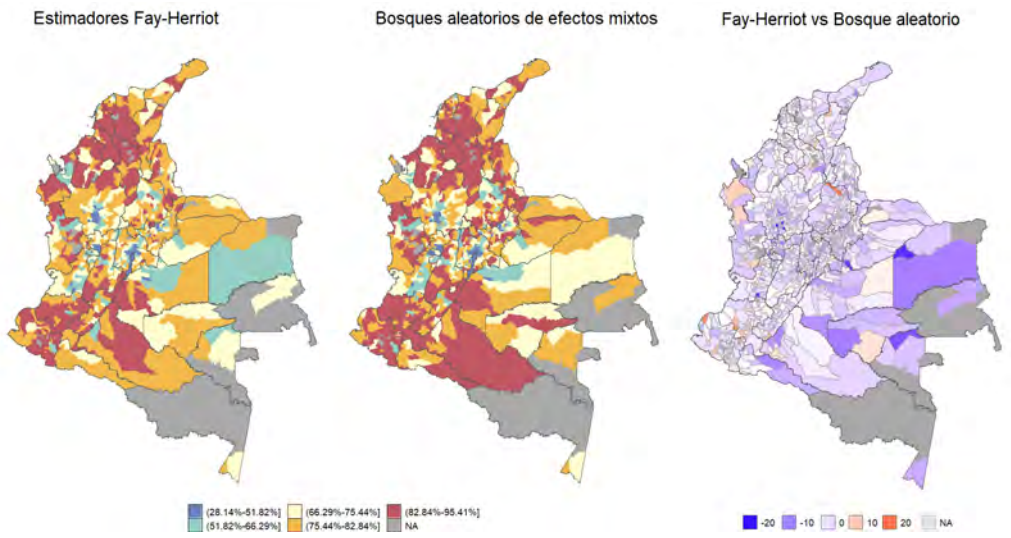
Los resultados de ambos modelos confirman una variación considerable en las tasas de informalidad a lo largo de la geografía del país, estimando una informalidad entre el 25 % (Sabaneta, Envigado, y Caldas) y el 97 % (Mercaderes, Talaigua Nuevo, y Toribío). Adicionalmente, para ambos métodos el *promedio simple* de informalidad municipal estimada es de 77 % con una desviación estándar de 10.5 puntos porcentuales.[16] Como se puede observar en la figura 3, existen sustanciales patrones regionales en las estimaciones, donde el Pacífico, Caribe y la Amazonía se destancan como las áreas con una mayor informalidad, estimada en promedio, de 30 puntos porcentuales mayores que las región Andina.

Por otra parte, el panel C en la figura 3 muestra la diferencia nominal entre las estimaciones por Fay-Herriot transformado y MERF para cada municipio de la muestra. La variabilidad de esta diferencia está acotada entre -20 y 20 puntos porcentuales en la tasa de informalidad. Esencialmente, estas diferencias son grandes en magnitud para la región Amazónica, y para algunos municipios del Pacífico, Caldas, Magdalena y Sucre, mientras que son pequeñas y casi inexistentes para varios en la región Andina, Bolívar y La Guajira. Una alta proporción de los municipios con una diferencia superior a 10 puntos porcentuales (pp) entre los modelos se encuentran

[15] Los resultados están disponibles previa solicitud a los autores.
[16] El valor del promedio simple no tiene en cuenta la representatividad o tamaño poblacional.

en los departamentos de Meta, Casanare, Amazonas, Guainía y Vaupés. Como se mencionó anteriormente, la calidad de las estimaciones no solo dependen de una correcta especificación del modelo, sino de la calidad de los datos auxiliares. Es posible entonces que las estimaciones para estas áreas estén afectadas por la calidad de su información auxiliar. Por ejemplo, algunos municipios de Vaupés aparecen en 2021 con una tasa de PILA en relación con la población total de 0.1 % y con menos de 1 % de la población vinculada al régimen subsidiado. Asimismo, varios municipios de Amazonas y de Guainía aparecen con tasas de vinculación en el régimen subsidiado de menos de 1 %. Finalmente, es importante considerar que la implementación de MERF actualmente no cuenta con la posibilidad de ponderar las estimaciones por la representatividad del muestreo municipal -medido por la cantidad de individuos encuestados en la muestra-. Esta diferencia en la corrección muestral puede también estar influyendo en las diferencias.

Figura 3: Comparación entre modelos de estimación de informalidad laboral, 2021



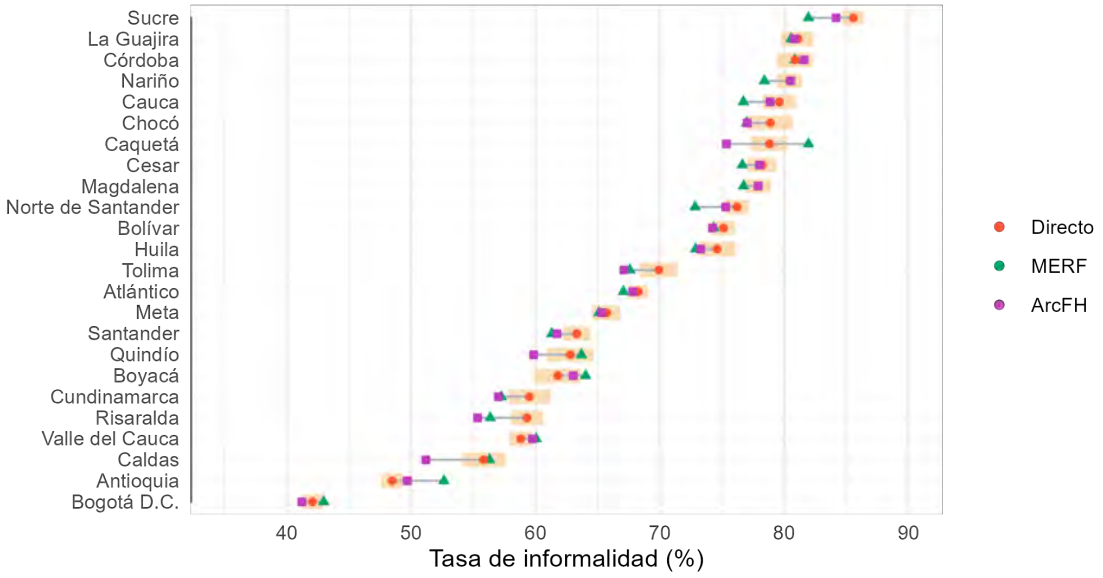
Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE.

La figura 4 tiene un doble propósito. Por una parte muestra la comparación entre las estimaciones departamentales utilizando la información municipal que arrojan ambos modelos y los estimados directos departamentales para los cuales la encuesta es representativa. Se incluye además intervalos de confianza del 95 % para las estimaciones directas, lo que permite una mejor comparación de los modelos.[17] Asimismo, la figura advierte de la alta dispersión entre departamentos de Colombia,

[17]Las figuras 12 y 13, presentan las gráficas para los demás años.

el cual coincide con la elevada variabilidad observada en los estimadores municipales. Los resultados departamentales para MERF y Fay-Herriot transformado en la Figura 4 exhiben el promedio ponderado de la informalidad municipal de cada departamento, donde la variable de ponderación es la población en edad de trabajar (PET). Una mejor aproximación de los promedios departamentales debería utilizar información de la población ocupada o población económicamente activa en cada municipio, pero esta información es inexistente en Colombia, por lo que aproximamos dicha población con la PET municipal. Pese a estas imprecisiones, la informalidad ponderada departamental de los modelos se asemejan en la mayoría de los departamentos a los valores directos estimados con la GEIH. En general, las estimaciones con los modelos Fay-Herriot transformados se acercan más a los estimadores directos departamentales que los estimadores MERF.

Figura 4: Estimaciones departamentales, 2021



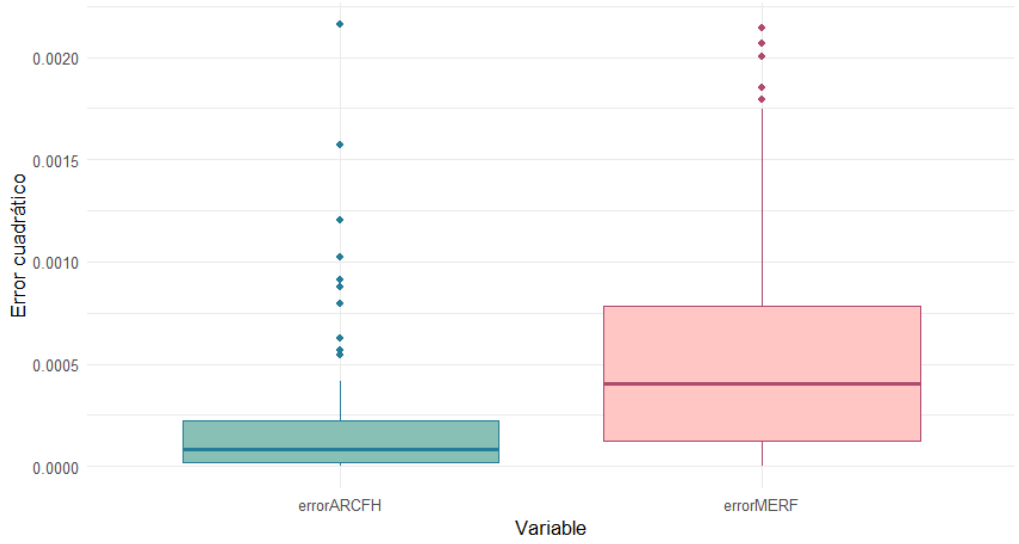
Notas: la información departamental correspondiente a los modelos SAE y MERF representan un estimado ponderado de los valores municipales. Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE.

A manera de comparación entre el comportamiento de los modelos utilizados, estimamos la raíz del error cuadrático medio empírico (empirical root mean squared error) siguiendo:

$$\text{RMSE}_{\text{emp}} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\hat{\mu}_d - \mu_d)^2} \tag{8}$$

Donde $\hat{\mu}_d$ define la media estimada del departamento d, como se describió, y μ_d es la media real del departamento d aproximada por la estimación directa de la GEIH. El sesgo relativo del modelo Fay-Herriot transformado y MERF son 0.0002 y 0.0005, respectivamente, como se muestra en la figura 5. Estos resultados sugieren que el comportamiento de Fay-Herriot transformado es ligeramente mejor que MERF, lo cual coincide con la evidencia de Krennmair y Schmid (2022). Por esta razón, nuestro modelo preferible es ArcFH y las estimaciones finales para todos los años resultan de estos modelos.

Figura 5: Error cuadrático medio empírico



Los resultados departamentales, así como los municipales, reflejan un evidente patrón de aglomeración regional. Si bien es cierto que todos los departamentos de Colombia presentan niveles de informalidad superiores que otras zonas del mundo, los departamentos con un mayor porcentaje de su población ocupada en esta condición se encuentra principalmente en el Caribe y el Pacífico. En el extremo opuesto se encuentran los departamentos centrales, donde Bogotá y Antioquia lideran los mejores indicadores.

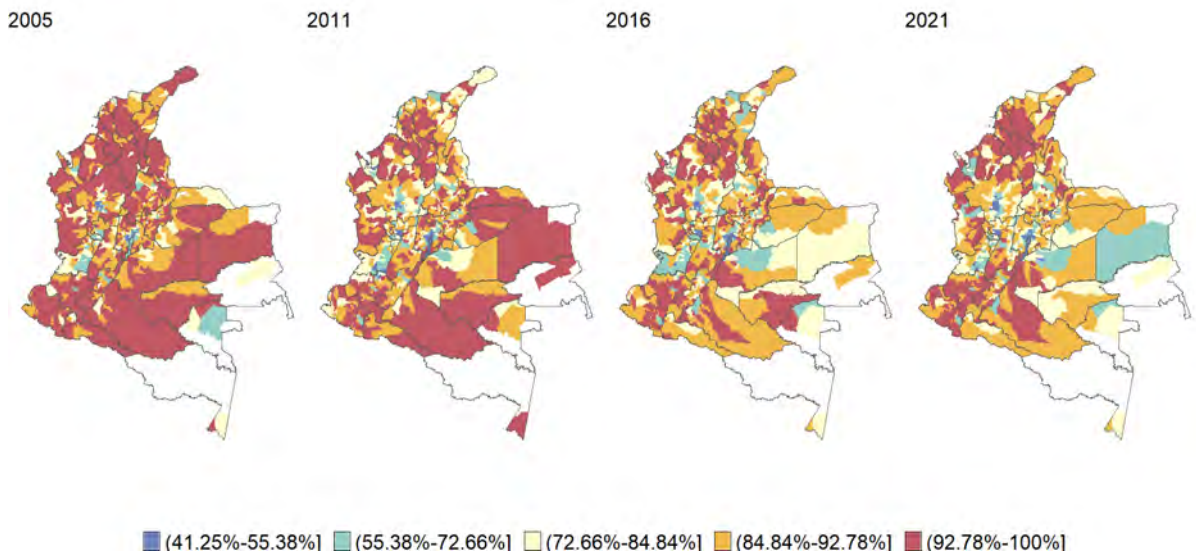
5.2. Cambios en el tiempo

La figura 6 muestra las estimaciones de informalidad directa usando la información del censo del 2005 y los resultados de las estimaciones usando el modelo Arch-FH para los años 2011, 2016 y 2021. Temporalmente se evidencia una caída paulatina de la informalidad en varios municipios y en las distintas regiones del país. Entre 2005 y 2021 la informalidad disminuyó en casi todos los municipios del país, pasando de un promedio simple de 89.7 % en 2005, a un 85 % en 2011, a un 78.8 % en 2016 hasta un 77.3 % en 2021.[18] La excepción se concentró en algunos municipios en Bolívar y Magdalena donde la informalidad permaneció más o menos en los mismos niveles del 2005. Por ejemplo en los municipios de María La Baja, Talaigua y Tiquisio en el departamento de Bolívar las tasas de informalidad pasaron de 94 %, 96 % y 95 % en 2005 a 99 %, 98 % y 96 % en 2021, respectivamente. Por otra parte, la informalidad parece haber disminuido más notablemente en La Guajira, Cesar y en la Orinoquía (Arauca, Casanare, Meta, y Vichada). En La Guajira, la informalidad promedio pasó de 90 % en 2005 a 81 % en 2021; en Cesar de 93 % a 83 % y en Orinoquía del 90 % en 2005 al 74 % en 2005. Vale la pena señalar, sin embargo, que la informalidad en estas regiones sigue siendo alta ubicándose en tasas superiores al 70 % y que los números se deben interpretar con precaución .

A pesar de la caída en las tasas de informalidad en la región de la Orinoquía, las mayores tasas de informalidad municipal se reportaron en los municipios del Meta tanto en 2005 como en 2021. Por ejemplo, los municipios de El Dorado, El Castillo, Fuerte de Oro y San Juan de Arama, entre otros, aparecen entre aquellos con mayores tasas de informalidad nacional, superiores al 85 % tanto en 2005 como en 2021. Por su parte, en 2005 y 2021, los municipios con menores niveles de informalidad aparecen en Cundinamarca y Antioquia. Por ejemplo, en 2005 las menores tasas de informalidad se estimaron en los municipios de Sabaneta y Envigado en Antioquia con un valor del 42 %. En 2021 estos mismos municipios lideraban el ranking de formalidad con tasas de informalidad cercanas al 25 %.

[18] El valor del promedio simple no tiene en cuenta el tamaño poblacional.

Figura 6: Informalidad laboral municipal, 2005, 2011, 2016 y 2021

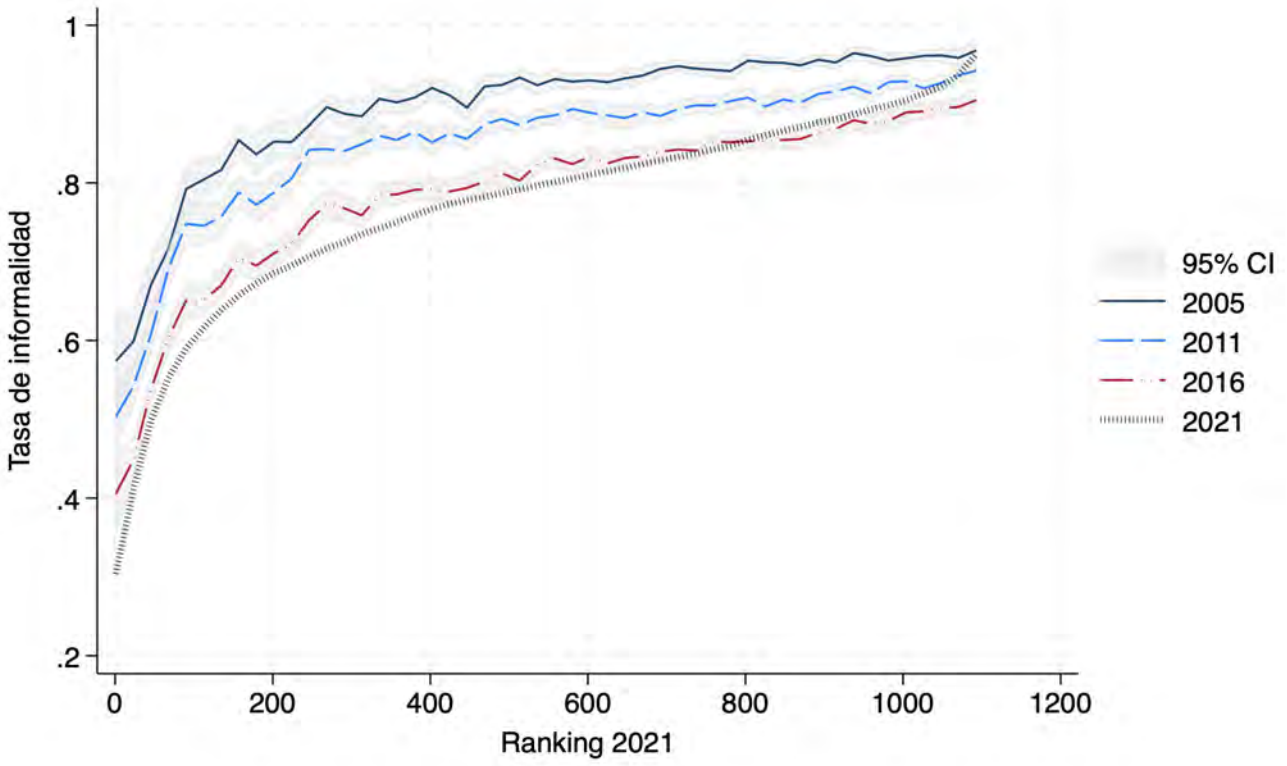


Notas: Para los años 2011, 2016 y 2021 las estimaciones son hechas usando el modelo Fay-Herriot transformado. Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE.

Los resultados en la figura 6 indican que los mayores cambios en informalidad se dieron entre 2005 y 2016 donde cayó 10.9 puntos porcentuales (pp). Adicionalmente, el número de municipios con informalidad mayor a 84.84 % pasó de 887 en 2005 a 369 en 2016. Las ganancias en el 2021 fueron menos significativas, y la informalidad se redujo solo 1 pp con respecto a la informalidad en 2016 y el número de municipios con informalidad mayor a 84.84 % pasó a 308.

La figura 7 corrobora estos resultados. En esta gráfica se muestra a todos los municipios organizados de acuerdo su nivel de informalidad en 2021. Se observa que, aunque los niveles de informalidad cayeron (en promedio) paulatinamente entre 2005 y 2016 para todos los grupos de municipios, aquellos con una alta tasa de informalidad experimentaron un retroceso en estas ganancias en 2021. La figura también evidencia una considerable desigualdad en la reducción de la informalidad a través del tiempo. En particular, se muestra que en el grupo con menores tasas de informalidad la disminución en la informalidad fue mayor; cerca de 20 pp entre 2005 y 2021. En contraste, los 400 municipios con informalidad más alta en 2021 vieron reducciones de alrededor de 10 pp y otros convergieron hacia sus tasas de 2005.

Figura 7: Cambios en la informalidad laboral municipal, 2005 - 2021



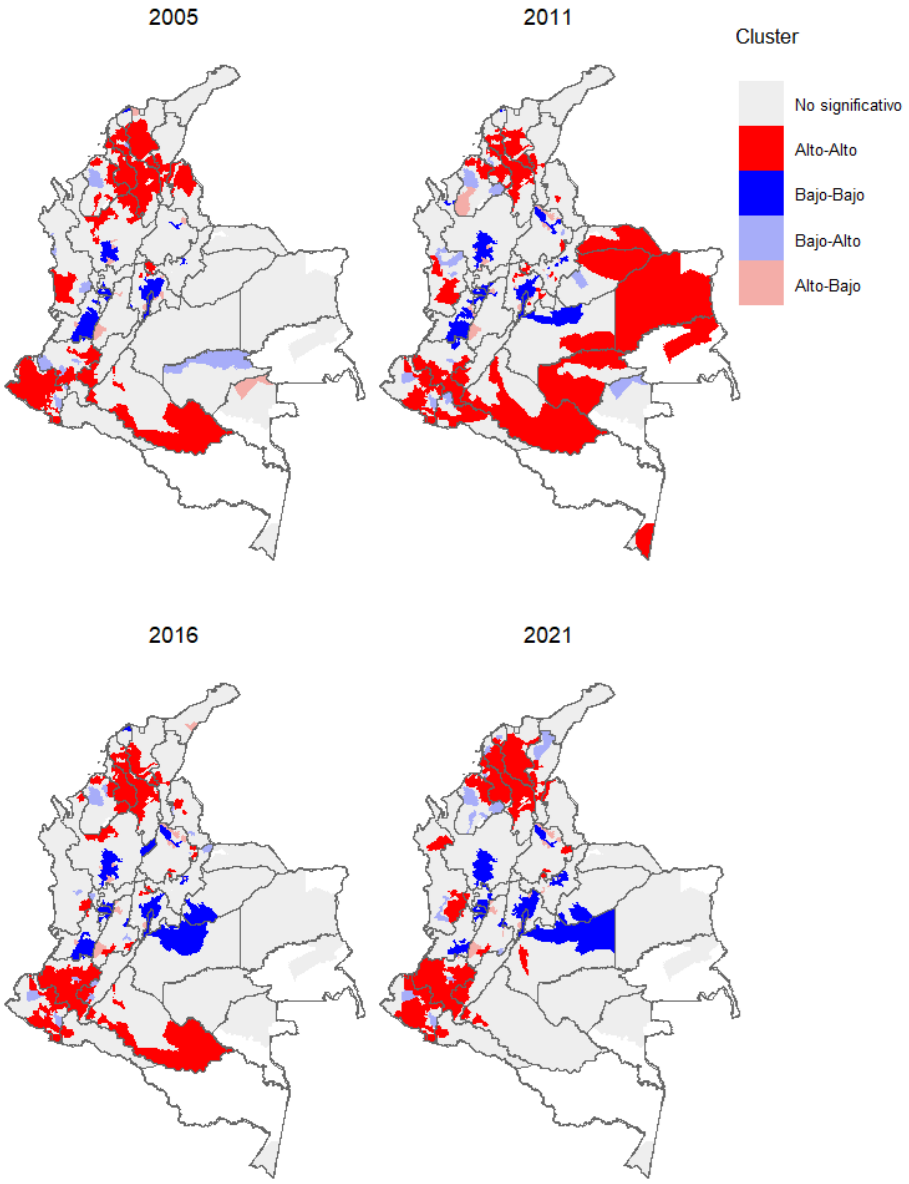
Notas: (1) Para los años 2011, 2016 y 2021 las estimaciones son hechas usando el modelo Fay-Herriot transformado. Los valores de 2005 corresponden a las estimaciones basadas en el Censo de población de 2005. (2) La gráfica traza un polinomio local suave para cada año. Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE.

5.3. Clústeres

En línea con estos resultados, la figura 8 presenta los indicadores locales de asociación espacial para los resultados de la informalidad en 2005 calculada de forma directa con información censal y para los años 2011, 2016 y 2021 estimadas mediante el modelo Arch-FH. Como puede observarse, la concentración de la informalidad es relativamente estable a través del tiempo y las altas tasas de informalidad en rojo se concentran en los departamentos de Bolívar, Cauca, Córdoba, Huila, Magdalena, Nariño y Sucre. En todos los años, los clústeres más grandes de alta informalidad se encuentran en Bolívar y Nariño. Por su parte, las zonas donde se concentran las menores tasas de informalidad (áreas en azul) son en los departamentos de Antioquia, Cundinamarca-Bogotá, Valle del Cauca y Santander. Estas zonas incluyen a lo que

se ha conocido históricamente como el trapecio de oro, o trapecio de prosperidad económica en Colombia.

Figura 8: Indicadores locales de asociación espacial



Nota: El mapa muestra los municipios con autocorrelación positiva y negativa para el test de LISA a un 95 % de significancia. Los municipios en gris tiene autocorrelación nula con sus vecinos.

De acuerdo a los resultados presentados previamente, Antioquia y Cundinamarca conforman los clústeres de formalidad más grandes del país en todos los años que se presentan en este documento. Los resultados también nos informan sobre

aquellos municipios con alta informalidad que están rodeados por municipios de baja formalidad (color rosa). De manera consistente observamos los casos de Pasca en Cundinamarca y Rioblanco en Tolima con esta característica. Asimismo, encontramos algunos municipios con bajas tasa de informalidad que están rodeados por municipios con altas tasas de informalidad (color azul celeste). A lo largo de los años, este fenómeno es prevalente en los municipios de Popayán (Cauca), Montería (Córdoba) y Pasto (Nariño), los cuales tienen niveles bajos de informalidad en comparación con sus vecinos más cercanos.

El análisis de los municipios cuya clasificación de clústeres ha sido estable a lo largo de los años analizados revela un patrón destacable. 151 municipios de 1,092 municipios con información para todos los años resultaron con clústeres significativos e inalterado en todos los años observados.[19] La mayor parte de los municipios con clústeres fijos a lo largo de los años se concentran en los clústeres alto-alto (52 %, equivalente a 78 municipios) y bajo-bajo (45 %, correspondiente a 68 municipios). Por su parte, los clústeres alto-bajo y bajo-alto agrupan, respectivamente, 2 y 3 municipios. Los municipios con clúster alto-alto fijos se distribuyen en los departamentos de Nariño (45 %), Bolívar (40 %), Magdalena (27 %), Sucre (23 %), Cauca (18 %), Huila (14 %), Córdoba (11 %), Caquetá (6,2 %) y Cesar (4 %). A su vez, los departamentos que agrupan los municipios con clústeres bajo-bajo estables y sus respectiva participación dentro de los municipios del departamento son Cundinamarca (24 %), Antioquia (18 %), Valle (14 %), Boyacá (1.6 %), Cauca (4.9 %), Risaralda (8.3 %), Santander (2.3 %), Bogotá, Caldas (3.7 %) y Quindío (8.3 %).

Las áreas con tasas de informalidad consistentemente altas rodeadas de alta informalidad tienen características notablemente diferenciadas del resto de áreas (Cuadro 4). De la comparación de las variables no consideradas en los modelos SAE, encontramos que las áreas subnacionales con clústeres altos-altos, en contraste con el resto de áreas, una calidad de la educación muy inferior (medida a través de las pruebas estandarizadas Saber), tienen menores años desde su creación y una alta distancia a los mercados mayoristas de alimentos más cercanos. Adicionalmente, aunque las diferencias son menos marcadas en comparación con el resto de áreas, reflejan elevadas tasas de dependencia demográfica, tasas de pobreza multidimensional, pobreza monetaria y monetaria extrema.

[19] La lista de municipios con clústeres fijos se encuentra en el apéndice 6.

Cuadro 4: Características generales de clústeres estables

	Clúster				
Variable	**NA**	**HH**	**LL**	**HL**	**LH**
Edad	142.7	127.6	236.3 ***	398.0 ***	98.5
Promedio Lectura Saber	47.4	46.5 ***	51.2 ***	52.4 ***	48.7
Promedio Matemáticas Saber	46.3	45.2 **	49.9 ***	51.5 **	47.1
Promedio Global Saber	229.6	225.6 **	246.3 ***	253.5 ***	233.1
Distancia mercado alimentos	58.6	77.4 ***	23.5 ***	0.0 *	31.2
Distancia a Bogotá	307.9	502.5 ***	164.2 ***	450.8 *	112.1 *
Tasa de dependencia	83.4	80.5 **	66.0 ***	63.2 ***	89.3
NBI	22.4	30.1 ***	6.0 ***	12.1	21.1
IPM	41.9	50.9 ***	15.7 ***	20.6 ***	45.6
Pobreza moderada	47.9	58.6 ***	22.2 ***	25.5 ***	55.0
Pobreza extrema	20.7	28.1 ***	5.5 ***	5.7 ***	25.6
N	941	78	68	3	2

Nota: (1) Las clases agrupan los municipios según clústeres estables entre 2005 y 2021. (2) ***p $<$ 0,01, **p $<$ 0,05, *p $<$ 0,1. La significancia estadística resulta de una prueba t-Student de las diferencias de medias entre cada clúster y NA (municipios que nunca son clasificados en un clúster o cuyo tipo de clúster cambió a lo largo del periodo de análisis).

5.4. Importancia de variables

La figura 9 ilustra la importancia de las variables auxiliares en la predicción de la informalidad en los 1,113 municipios evaluados para el 2021 y en el apéndice, las figuras 10 y 11, presentan las gráficas para los demás años. En dicha representación, se observa que las variables con mayor capacidad predictiva respecto a la informalidad siguen un patrón intuitivo. Consistentemente encontramos que las dos variables más importantes en la predicción de la informalidad son la proporción de la población vinculada al régimen subsidiado y la tasa de PILA en relación con la población total. Este resultado es importante pues evidencia la capacidad del modelo de identificar variables altamente correlacionadas con la informalidad. Así mismo, estas variables están asociadas a la definición de informalidad evaluada en este documento, como lo es no cotizar a un fondo de pensiones. En tercer lugar, encontramos la tasa de vacantes virtuales, información que parte del Servicio Público del Empleo, y que está asociada a la oferta disponible de trabajo formal en el municipio, aunque, otras ofertas de empleo formal pueden provenir de fuentes no analizadas.

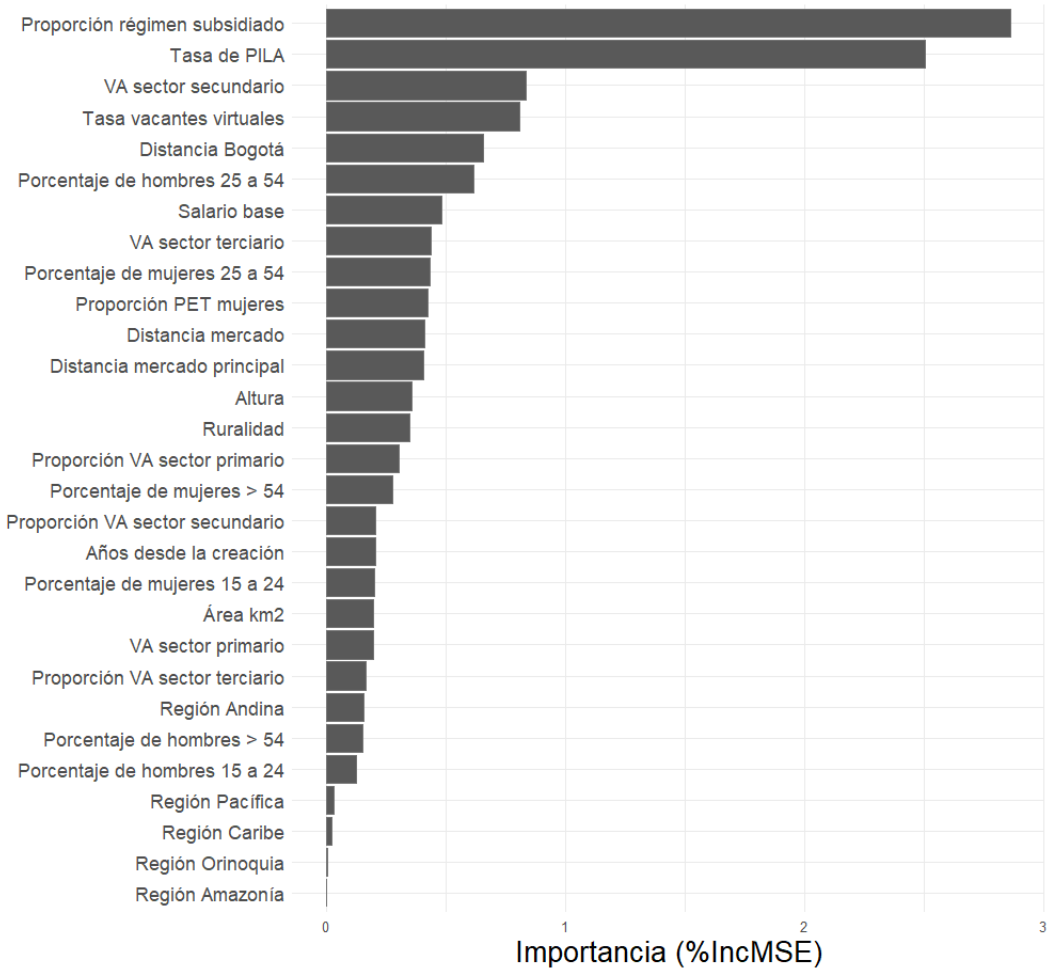
Además, la actividad económica, en particular la vinculada al sector secundario, muestra una alta capacidad predictiva, posiblemente porque requiere mano de

obra calificada, aunque este resultado valdría la pena explorarlo en futuras investigaciones. En relación con la oferta de trabajo el porcentaje de la población entre 25 y 54 años resultan ser importante en la estimación de la informalidad. Asimismo, el salario promedio que mide el costo de la informalidad aparece como una variable relevante en la predicción de la informalidad. Finalmente, se resalta la importancia de las variables que capturan el acceso a mercados que tienen los municipios. Por ejemplo, la distancia de estos a Bogotá, al mercado y a la capital más cercana del departamento son factores determinantes. A través del ejercicio también se pudo identificar que con estas 7 principales variables se explica el 70 % del comportamiento de la informalidad. Reduciendo la dimensionalidad y encontrando determinantes con alto poder predictivo de la misma.

Estos resultados están en línea con investigaciones previas sobre el mercado laboral en Colombia y sus determinantes. Por ejemplo, Arango y Flórez (2020) muestran que, durante 1984-2015, la tasa de desempleo estuvo determinada por el salario mínimo, la tasa de vacantes, la actividad económica y la proporción de trabajadores sin educación universitaria, entre otros.[20] Arango y Flórez (2021) concluyen que el efecto del salario mínimo en la informalidad es significativo aunque señalan que el efecto puede ser regionalmente heterogéneo (Arango, Flórez, y Guerrero, 2020).

[20]Por su parte, Tamayo (2008) encuentra que el papel de la participación de los jóvenes en el mercado laboral es un factor determinante en el desempleo.

Figura 9: Importancia de las variables auxiliares para la estimación por MERF, 2021



Nota: importancia de las variables estimada a través de la ganancia en la reducción del error cuadrátio medio de la adición de cada variable en el modelo

6. Conclusiones

Múltiples políticas públicas requieren de información socioeconómica desagregada. Por ejemplo, al abordar las desigualdades en el mercado laboral, en particular en términos espaciales, es esencial contar con información más precisa que la proporcionada por encuestas representativas. Como alternativa para facilitar estimaciones de informalidad en subpoblaciones ”invisibles”, en este documento hacemos uso del colectivo de herramientas metodológicas que ofrecen las estimaciones de áreas pequeñas. Emplear estos métodos nos permiten tener un mayor nivel de granularidad de la tasa de informalidad laboral en Colombia, lo cual posibilitaría intervenciones de política basada en lugares con un mayor nivel de precisión.

Aunque reconocemos que los modelos utilizados pueden tener imperfecciones, las técnicas de Estimación de Áreas Pequeñas pueden ser una alternativa para observar, medir y entender parcialmente indicadores laborales para poblaciones para los que no existe información censal. En este documento utilizamos avances relativamente recientes de SAE que intentan prevenir potenciales errores de especificación como la propuesta de Arc-FH y MERF. Pese a la existencia de algunos valores discrepantes entre estos modelos, en términos generales, convergen en conclusiones similares. Cuando desagregamos las tasas de informalidad municipal, se observa que pese a una mejora lenta y progresiva a lo largo de los años, persiste una profunda y alta heterogeneidad al interior de Colombia. Asimismo, pese a esta alta dispersión, incluso los municipios con las más bajas tasas de informalidad tienen indicadores comparativamente altos en el ámbito internacional.

Los resultados revelan patrones regionales importantes, con unas tasas de informalidad que varían entre el 25 % y el 97 %, y donde la mayor parte de los municipios han mantenido tasas de informalidad superiores al 50 % desde 2005. Sin embargo, los resultados también evidenciaron una reducción entre 2005 y 2021 en las tasas de informalidad para casi todos los municipios, aunque esta reducción estuvo altamente concentrada geográficamente. Los departamentos mayoritariamente Caribes y Pacíficos concentran las tasas de informalidad más altas del país. En este grupo se destaca Sucre, La Guajira y Córdoba de la región Caribe, así como Nariño, Cauca y Chocó del Pacífico. No obstante, del análsis subnacional más detallado se infiere que existen zonas del país que no solo han tenido tasas de informalidad prevalentemente altas, sino que también están rodeadas de alta informalidad a lo largo de los años. Se distinguen en este grupo el sur del Caribe, específicamente municipios del departamento de Bolívar, Magdalena y Sucre, así como municipios de Nariño y Cauca en la región Pacífica. Los municipios con persistencia de elevada informalidad relativa y rodeados de alta informalidad a lo largo de los años tienen características diferenciadas del resto de municipios que ameritan un estudio más detallado para poder identificar rutas de desarrollo alineadas con la reducción de este fenómeno, relacionado con otros indicadores de desarrollo como los ingresos y la pobreza.

Asimismo, de este estudio se concluye que la mayor reducción en informalidad sucedió entre el 2005 y el 2016 y que la pandemia desaceleró la disminución de las tasas de informalidad. Además, los municipios con una alta prevalencia de informalidad son los que experimentaron una mayor deterioro de estos indicadores con la pandemia. Este es el primer estudio que hace estimaciones municipales para

Colombia, por lo que consideramos que puede ser un paso introductorio para la discusión sobre la concentración geográfica de las imperfecciones en el mercado laboral Colombiano.

Referencias

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, *27*(2), 93–115.

Arango, L. E., y Flórez, L. A. (2020). Determinants of structural unemployment in colombia: a search approach. *Empirical Economics*, *58*(5), 2431–2464.

Arango, L. E., y Flórez, L. A. (2021). Regional labour informality in colombia and a proposal for a differential minimum wage. *The Journal of Development Studies*, *57*(6), 1016–1037.

Arango, L. E., Flórez, L. A., y Guerrero, L. D. (2020). Minimum wage effects on informality across demographic groups in colombia. *Borradores de Economía; No. 1104*.

Ariza, J., y Retajac, F. A. (2021). Composición y evolución de la informalidad laboral en colombia durante el periodo 2009-2019. *Apuntes del CENES*, *40*(72), 115–148.

Battese, G. E., Harter, R. M., y Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36.

Bernal, S. (2009). The informal labor market in colombia: Identification and characterization. *Desarrollo y sociedad*(63), 145–208.

Biau, G., y Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227.

Bustamante, J. P. (2011). Los retos de la economía informal en colombia. *Notas fiscales*, *9*, 1–36.

Casas-Cordero Valencia, C., Encina, J., y Lahiri, P. (2016). Poverty mapping for the chilean comunas. *Analysis of poverty data by small area estimation*, 379–404.

Celín Camargo, Y., Ramírez Buitrago, C. F., Torres Gómez, E. E., López González, M., Castrillón Gaviria, C. C., Gómez Muñoz, W. A., ... Argüello, R. (2023). *Aproximaciones formales a la informalidad*. Editorial Universidad del Rosario.

Cárdenas, M. S., y Mejía, C. M. (2007). Informalidad en colombia: Nueva evidencia. *Working Papers Series - Documentos de Trabajo*(35).

DANE. (2009). *Metodología informalidad gran encuesta integrada de hogares - geih* (Inf. Téc.). Departamento Administrativo Nacional de Estadística.

DANE. (2018). *Fichas metodológicas indicadores marco 2018 geih* (Inf. Téc.). Departamento Administrativo Nacional de Estadística.

Fay, R. E., y Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical*

Association, *74*(366a), 269–277.

Flórez, C. E. (2002). The function of the urban informal sector in employment: Evidence from colombia 1984-2000. *Documento cede*, *4*, 1–61.

Franco, C., y Bell, W. R. (2015). Borrowing information over time in binomial/logit normal models for small area estimation. *Statistics in Transition new series*, *16*(4), 563–584.

Galvis-Aponte, L. A. (2012). Informalidad laboral en las áreas urbanas de colombia. *Documentos de Trabajo Sobre Economía Regional y Urbana; No. 164*.

Guataquí, J. C., García, A. F., y Rodríguez, M. (2011). El perfil de la informalidad laboral en colombia. *Perfil de coyuntura económica*(16), 91–116.

Ha, N., Lahiri, P., y Parsons, V. (2014). Methods and results for small area estimation using smoking data from the 2008 national health interview survey. *Statistics in Medicine*, *33*, 3932–3945.

Krennmair, P., y Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *71*(5), 1865–1894.

LaboUR, O. L. (2018). Perfil actual de la informalidad laboral en colombia: estructura y retos. *Universidad del Rosario*, *6*.

Li, H., y Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, *101*(4), 882–892.

Liu, B., Lahiri, P., y Kalton, G. (2014). Hierarchical bayes modeling of survey-weighted small area proportions. *Survey Methodology*, *40*(1), 1–13.

López-Vizcaíno, E., Lombardía, M. J., y Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *178*(3), 535–565.

Mora, J. J., y Muro, J. (2017). Dynamic effects of the minimum wage on informality in colombia. *Labour*, *31*(1), 59–72.

Petersen, L., Minkkinen, P., y Esbensen, K. H. (2005). Representative sampling for reliable data analysis: Theory of sampling. *Chemometrics and Intelligent Laboratory Systems*, *77*(1), 261-277. Descargado de https://www.sciencedirect.com/science/article/pii/S0169743904002448 (FESTSCHRIFT HONOURING PROFESSOR D.L. MASSART) doi: https://doi.org/10.1016/j.chemolab.2004.09.013

Rao, J. N., y Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Ruffer, T., y Knight, J. (2007). Informal sector labor markets in developing countries.

University of Oxford, *44*.

Schmid, T., Bruckschen, F., Salvati, N., y Zbiranski, T. (2017). Constructing socio-demographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in senegal. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(4), 1163–1190.

Tamayo, J. (2008). La tasa natural de desempleo en colombia y sus determinantes. *Borradores de Economía*, *491*, 1–31.

7. Apéndice

7.A. Definiciones de informalidad

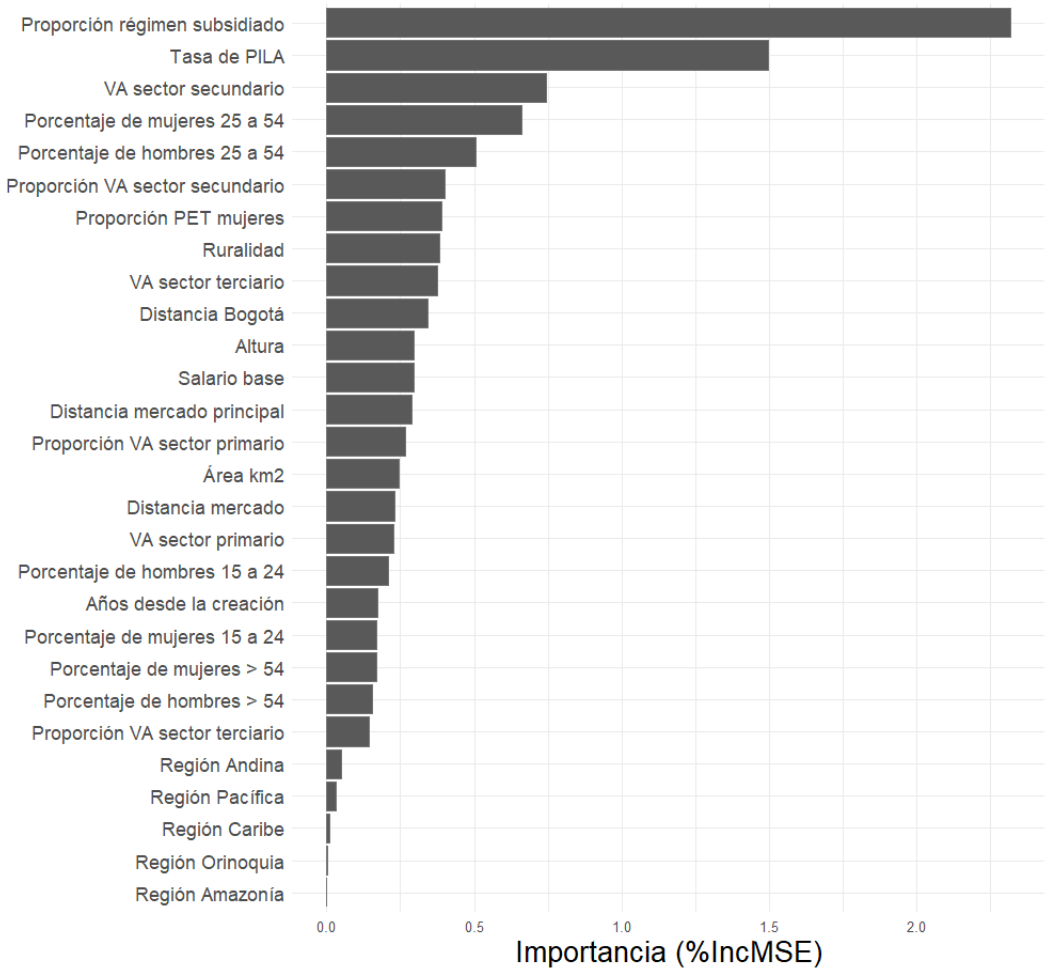
Cuadro 5: Definiciones de informalidad en la literatura

Fuente	Definición
DANE (2009)	Personas que cumplan con los siguientes criterios: 1. Los empleados que laboran en establecimientos que ocupen hasta cinco personas, incluyendo al patrono y/o socio; 2. Los trabajadores familiares sin remuneración; 3. Los trabajadores sin remuneración en empresas de otros hogares; 4. Los empleados domésticos; 5. Los jornaleros o peones; 6. Los trabajadores por cuenta propia que laboran en establecimientos hasta cinco personas, excepto los independientes profesionales; 8. Se excluyen los obreros o empleados del gobierno.
DANE (2018)	Asalariados que no cuentan con cotizaciones de salud ni a pensión. Igualmente todos los trabajadores familiares sin remuneración, así como los trabajadores por cuenta propia
Celín Camargo y cols. (2023)	**Enfoque legalista**: trabajadores con falta de protección laboral o beneficios legales. **Enfoque productivo**: trabajadores con baja productividad, depende de las características de la empresa, como el tamaño.

Fuente	Definición
Guataquí y cols. (2011)	**Fuerte**: asalariados y trabajadores domésticos que no cumplen con las siguientes características: Pertenecen al régimen contributivo o especial de salud como cotizantes, están cotizando a un fondo de pensiones o están pensionados, tienen contrato escrito de trabajo, ganan más del 95 % del salario mínimo por hora. También los independientes que cumplen las primeras dos. **Débil**: trabajadores asalariados, domésticos o independientes que no están afiliados como cotizante al sistema de seguridad social en salud ya sea bajo el régimen contributivo o subsidiado, o estar afiliado al régimen subsidiado de salud en régimen especial.
Bernal (2009)	Utiliza 27 definiciones. Las primeras 23 incluyen distintos niveles de cobertura o afiliación a prestaciones sociales y salud, la 24 y 25 son dependientes del tamaño de la empresa, y la 26 y 27 son referentes a tener contratos legales.
Flórez (2002)	Enfoque estructuralista: depende de las características de la empresa como el tamaño. El enfoque institucionalista: depende del cumplimiento de la normativa en los trabajadores. El enfoque basado en nuevas formas de organización del trabajo: depende de los ingresos y la calidad del trabajo.

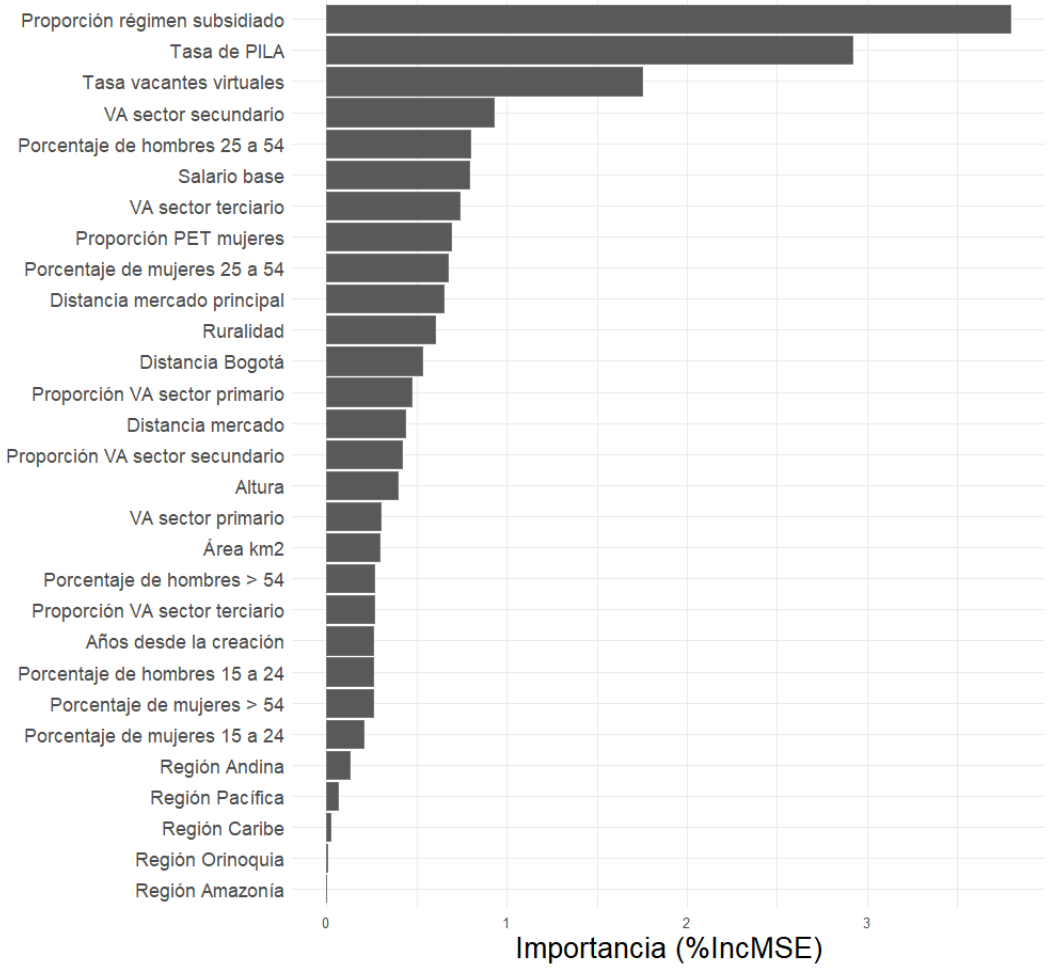
7.B. Importancia de variables

Figura 10: Importancia de las variables auxiliares para la estimación por MERF, 2011



Nota: importancia de las variables estimadas a través del número de veces que se generan particiones en la variables auxiliares (x) en la predicción de la informalidad

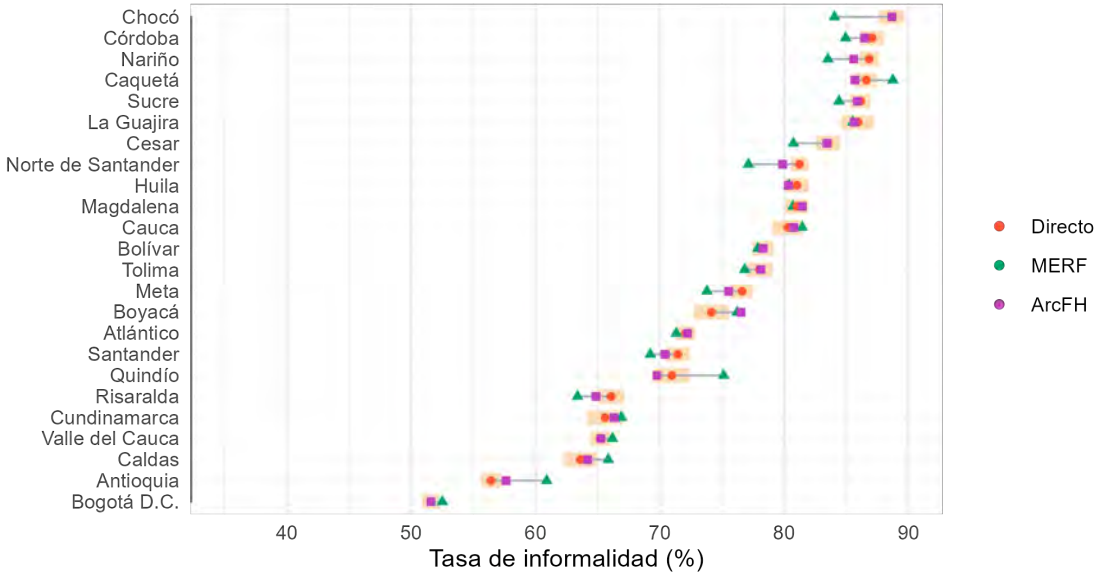
Figura 11: Importancia de las variables auxiliares para la estimación por MERF, 2016



Nota: importancia de las variables estimadas a través del número de veces que se generan particiones en la variables auxiliares (x) en la predicción de la informalidad

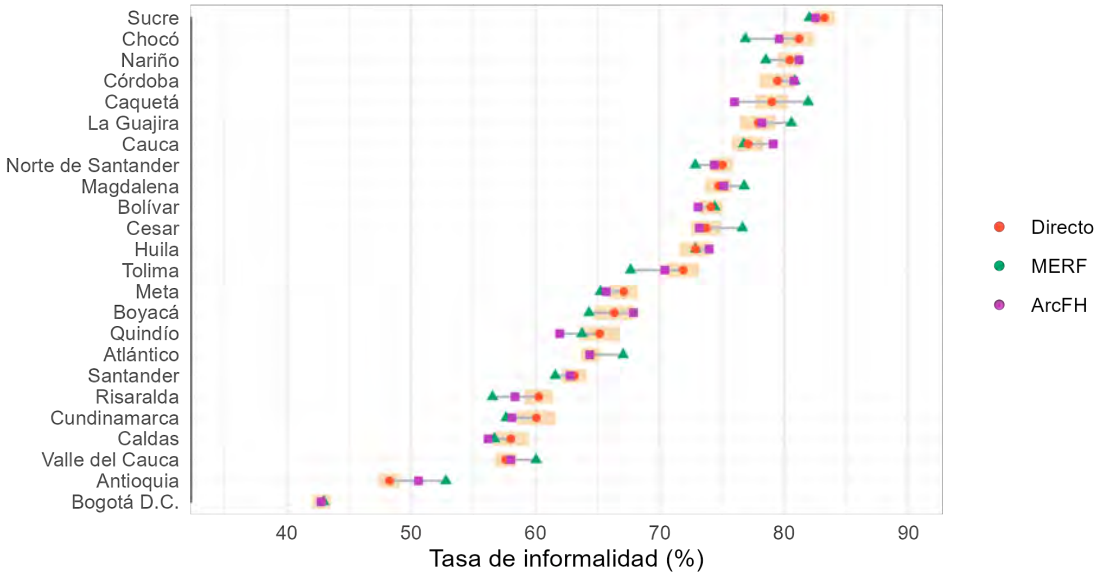
7.C. Comparación entre estimaciones departamentales

Figura 12: Estimaciones departamentales, 2011



Notas: la información departamental correspondiente a los modelos SAE y MERF representan un estimado ponderado de los valores municipales. Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE.

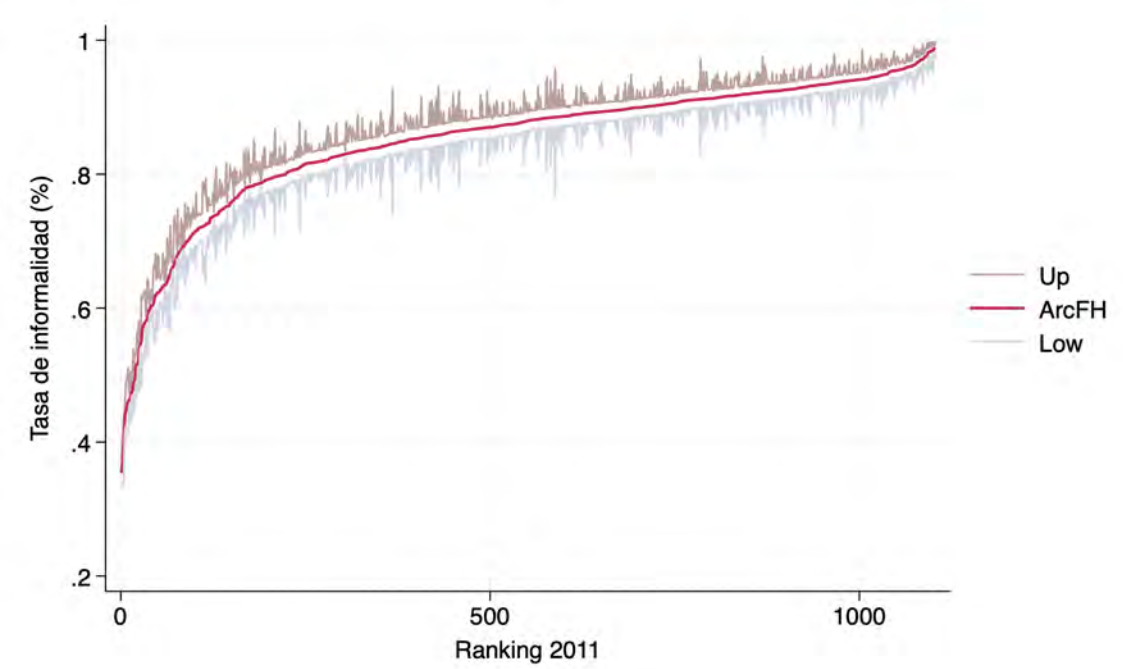
Figura 13: Estimaciones departamentales, 2016



Notas: la información departamental correspondiente a los modelos SAE y MERF representan un estimado ponderado de los valores municipales. Fuentes: Departamento Administrativo Nacional de Estadística (DANE), Planilla Integrada de Liquidación de Aportes (PILA), Servicio Público de Empleo (SPE), Panel municipal CEDE.

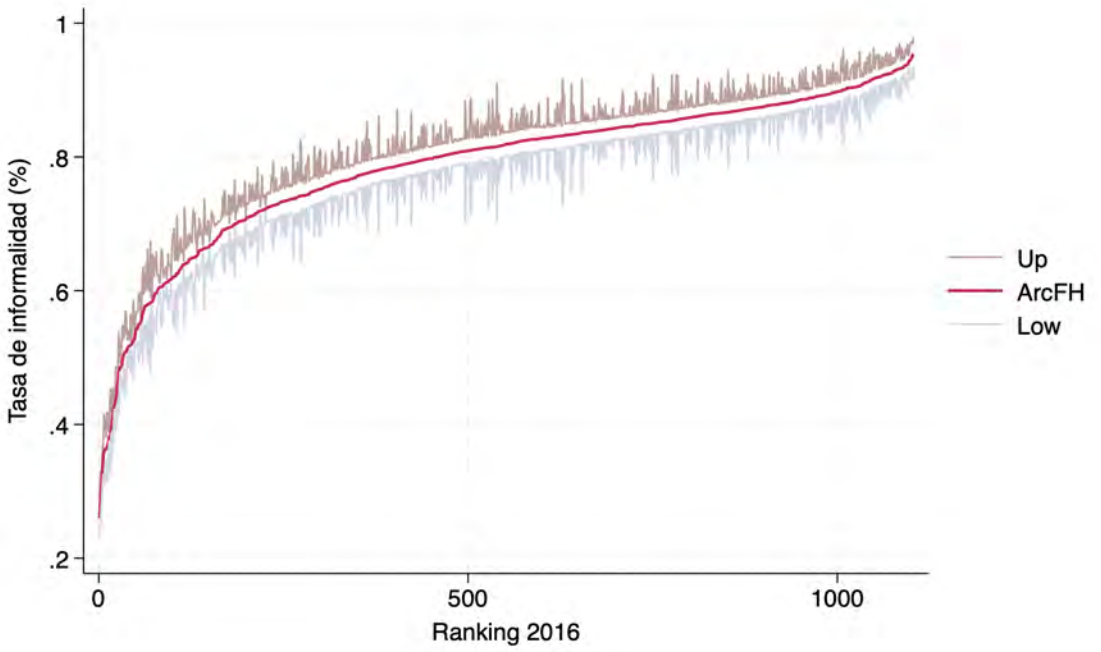
7.D. Estimadores FH transformados con intervalos de confianza basados en bootstrap

Figura 14: Estimaciones municipales, 2011



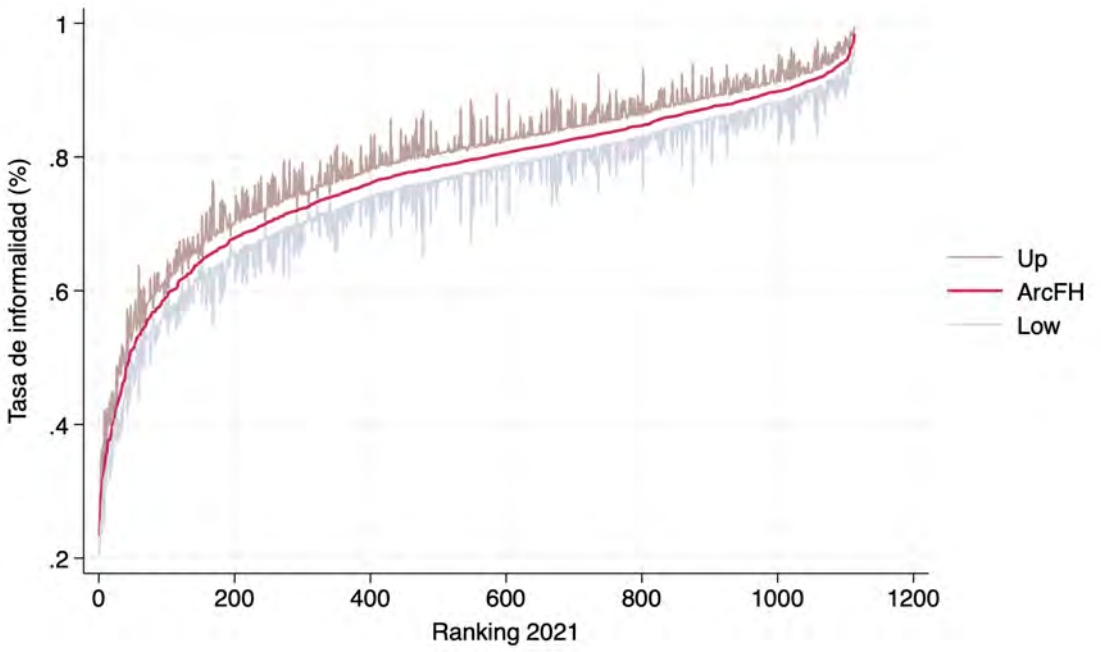
Notas: (1) Estimaciones basadas en FH transformado. (2) los intervalos de confianza resultan de 500 réplicas de Bootstrap.

Figura 15: Estimaciones municipales, 2016



Notas: (1) Estimaciones basadas en FH transformado. (2) los intervalos de confianza resultan de 500 réplicas de Bootstrap.

Figura 16: Estimaciones municipales, 2021



Notas: (1) Estimaciones basadas en FH transformado. (2) los intervalos de confianza resultan de 500 réplicas de Bootstrap.

7.E. Municipios en clústeres fijos en 2005, 2011, 2016 y 2021

Cuadro 6: Clústeres de municipios constantes (I)

Código mpio	Municipio	Departamento	Clúster
25769	SUBACHOQUE	CUNDINAMARCA	LL
25772	SUESCA	CUNDINAMARCA	LL
25785	TABIO	CUNDINAMARCA	LL
25793	TAUSA	CUNDINAMARCA	LL
25799	TENJO	CUNDINAMARCA	LL
25817	TOCANCIPÁ	CUNDINAMARCA	LL
25898	ZIPACÓN	CUNDINAMARCA	LL
25899	ZIPAQUIRÁ	CUNDINAMARCA	LL
63690	SALENTO	QUINDÍO	LL
66001	PEREIRA	RISARALDA	LL
66682	SANTA ROSA DE CABAL	RISARALDA	LL
68276	FLORIDABLANCA	SANTANDER	LL
68307	GIRÓN	SANTANDER	LL
76001	CALI	VALLE DEL CAUCA	LL
76130	CANDELARIA	VALLE DEL CAUCA	LL
76306	GINEBRA	VALLE DEL CAUCA	LL
76520	PALMIRA	VALLE DEL CAUCA	LL
76563	PRADERA	VALLE DEL CAUCA	LL
76869	VIJES	VALLE DEL CAUCA	LL

Clústeres de municipios constantes (II)

Código mpio	Municipio	Departamento	Clúster
5088	BELLO	ANTIOQUIA	LL
5129	CALDAS	ANTIOQUIA	LL
5148	EL CARMEN DE VIBORAL	ANTIOQUIA	LL
5212	COPACABANA	ANTIOQUIA	LL
5266	ENVIGADO	ANTIOQUIA	LL
5282	FREDONIA	ANTIOQUIA	LL
5308	GIRARDOTA	ANTIOQUIA	LL
5318	GUARNE	ANTIOQUIA	LL
5360	ITAGÜÍ	ANTIOQUIA	LL
5376	LA CEJA	ANTIOQUIA	LL
5380	LA ESTRELLA	ANTIOQUIA	LL
5440	MARINILLA	ANTIOQUIA	LL
5541	PEÑOL	ANTIOQUIA	LL
5607	RETIRO	ANTIOQUIA	LL
5615	RIONEGRO	ANTIOQUIA	LL
5631	SABANETA	ANTIOQUIA	LL
5656	SAN JERÓNIMO	ANTIOQUIA	LL
5664	SAN PEDRO DE LOS MILAGROS	ANTIOQUIA	LL
5679	SANTA BÁRBARA	ANTIOQUIA	LL
5697	EL SANTUARIO	ANTIOQUIA	LL
11001	BOGOTÁ, D.C.	BOGOTÁ, D.C.	LL
15806	TIBASOSA	BOYACÁ	LL
15820	TÓPAGA	BOYACÁ	LL
17174	CHINCHINÁ	CALDAS	LL
19573	PUERTO TEJADA	CAUCA	LL
19845	VILLA RICA	CAUCA	LL
25099	BOJACÁ	CUNDINAMARCA	LL
25126	CAJICÁ	CUNDINAMARCA	LL
25175	CHÍA	CUNDINAMARCA	LL
25200	COGUA	CUNDINAMARCA	LL
25214	COTA	CUNDINAMARCA	LL
25224	CUCUNUBÁ	CUNDINAMARCA	LL
25260	EL ROSAL	CUNDINAMARCA	LL
25269	FACATATIVÁ	CUNDINAMARCA	LL
25286	FUNZA	CUNDINAMARCA	LL
25295	GACHANCIPÁ	CUNDINAMARCA	LL
25322	GUASCA	CUNDINAMARCA	LL
25326	GUATAVITA	CUNDINAMARCA	LL
25377	LA CALERA	CUNDINAMARCA	LL
25430	MADRID	CUNDINAMARCA	LL
25473	MOSQUERA	CUNDINAMARCA	LL
25486	NEMOCÓN	CUNDINAMARCA	LL
25645	SAN ANTONIO DEL TEQUENDAMA	CUNDINAMARCA	LL
25658	SAN FRANCISCO	CUNDINAMARCA	LL
25754	SOACHA	CUNDINAMARCA	LL
25758	SOPÓ	CUNDINAMARCA	LL

Clústeres de municipios constantes (III)

Código mpio	Municipio	Departamento	Clúster
5001	MEDELLÍN	ANTIOQUIA	LL
5030	AMAGÁ	ANTIOQUIA	LL
5036	ANGELÓPOLIS	ANTIOQUIA	LL
25535	PASCA	CUNDINAMARCA	HL
73616	RIOBLANCO	TOLIMA	HL
19001	POPAYÁN	CAUCA	LH
23001	MONTERÍA	CÓRDOBA	LH
52001	PASTO	NARIÑO	LH
52019	ALBÁN	NARIÑO	HH
52036	ANCUYA	NARIÑO	HH
52227	CUMBAL	NARIÑO	HH
52250	EL CHARCO	NARIÑO	HH
52250	EL CHARCO	NARIÑO	HH
52254	EL PEÑOL	NARIÑO	HH
52256	EL ROSARIO	NARIÑO	HH
52260	EL TAMBO	NARIÑO	HH
52317	GUACHUCAL	NARIÑO	HH
52320	GUAITARILLA	NARIÑO	HH
52352	ILES	NARIÑO	HH
52354	IMUÉS	NARIÑO	HH
52356	IPIALES	NARIÑO	HH
52385	LA LLANADA	NARIÑO	HH
52399	LA UNIÓN	NARIÑO	HH
52405	LEIVA	NARIÑO	HH
52411	LINARES	NARIÑO	HH
52435	MALLAMA	NARIÑO	HH
52540	POLICARPA	NARIÑO	HH
52573	PUERRES	NARIÑO	HH
52585	PUPIALES	NARIÑO	HH
52612	RICAURTE	NARIÑO	HH
52678	SAMANIEGO	NARIÑO	HH
52683	SANDONÁ	NARIÑO	HH
52687	SAN LORENZO	NARIÑO	HH
52694	SAN PEDRO DE CARTAGO	NARIÑO	HH
52720	SAPUYES	NARIÑO	HH
52786	TAMINANGO	NARIÑO	HH
52838	TÚQUERRES	NARIÑO	HH
52885	YACUANQUER	NARIÑO	HH
70124	CAIMITO	SUCRE	HH
70204	COLOSÓ	SUCRE	HH
70508	OVEJAS	SUCRE	HH
70678	SAN BENITO ABAD	SUCRE	HH
70717	SAN PEDRO	SUCRE	HH
70771	SUCRE	SUCRE	HH

Clústeres de municipios constantes (IV)

Código mpio	Municipio	Departamento	Clúster
13006	ACHÍ	BOLÍVAR	HH
13030	ALTOS DEL ROSARIO	BOLÍVAR	HH
13062	ARROYOHONDO	BOLÍVAR	HH
13074	BARRANCO DE LOBA	BOLÍVAR	HH
13212	CÓRDOBA	BOLÍVAR	HH
13244	EL CARMEN DE BOLÍVAR	BOLÍVAR	HH
13248	EL GUAMO	BOLÍVAR	HH
13300	HATILLO DE LOBA	BOLÍVAR	HH
13430	MAGANGUÉ	BOLÍVAR	HH
13442	MARÍA LA BAJA	BOLÍVAR	HH
13458	MONTECRISTO	BOLÍVAR	HH
13468	SANTA CRUZ DE MOMPOX	BOLÍVAR	HH
13549	PINILLOS	BOLÍVAR	HH
13600	RÍO VIEJO	BOLÍVAR	HH
13650	SAN FERNANDO	BOLÍVAR	HH
13657	SAN JUAN NEPOMUCENO	BOLÍVAR	HH
13667	SAN MARTÍN DE LOBA	BOLÍVAR	HH
13810	TIQUISIO	BOLÍVAR	HH
18610	SAN JOSÉ DEL FRAGUA	CAQUETÁ	HH
19075	BALBOA	CAUCA	HH
19100	BOLÍVAR	CAUCA	HH
19355	INZÁ	CAUCA	HH
19397	LA VEGA	CAUCA	HH
19450	MERCADERES	CAUCA	HH
19693	SAN SEBASTIÁN	CAUCA	HH
19785	SUCRE	CAUCA	HH
20175	CHIMICHAGUA	CESAR	HH
23182	CHINÚ	CÓRDOBA	HH
23417	LORICA	CÓRDOBA	HH
23660	SAHAGÚN	CÓRDOBA	HH
41359	ISNOS	HUILA	HH
41378	LA ARGENTINA	HUILA	HH
41530	PALESTINA	HUILA	HH
41660	SALADOBLANCO	HUILA	HH
41668	SAN AGUSTÍN	HUILA	HH
47161	CERRO DE SAN ANTONIO	MAGDALENA	HH
47170	CHIVOLO	MAGDALENA	HH
47245	EL BANCO	MAGDALENA	HH
47541	PEDRAZA	MAGDALENA	HH
47555	PLATO	MAGDALENA	HH
47707	SANTA ANA	MAGDALENA	HH
47720	SANTA BÁRBARA DE PINTO	MAGDALENA	HH
47960	ZAPAYÁN	MAGDALENA	HH