DANE
INFORMACIÓN PARA TODOS

DIRECCIÓN
DE CENSOS Y DEMOGRAFÍA

ÍNDICE DE POBREZA MULTIDIMENSIONAL

PREDICCIÓN DEL IPM CENSAL USANDO APRENDIZAJE DE MÁQUINAS E IMÁGENES SATELITALES

FEBRERO DE 2020

El futuro es de todos
Gobierno de Colombia

CONTENIDO

Lista de tablas

Lista de gráficos

Lista de figuras

1. Introducción

El Departamento Administrativo Nacional de Estadística – DANE tiene como misión fundamental garantizar la disponibilidad y la calidad de la información estadística estratégica para el desarrollo económico y político del país, en el marco de sus objetivos misionales, así como garantizar la producción, disponibilidad y calidad de la información estadística estratégica; y dirigir, planear, ejecutar, coordinar, regular y evaluar la producción y difusión de información oficial básica.

En ese sentido, el DANE tiene como objetivo mejorar la disponibilidad de estadísticas relevantes con niveles de desagregación más detallados (relacionado con grupos poblacionales, o con dominios geográficos); así como de producir estadísticas relevantes con una mayor frecuencia. Un medio para lograrlo es a partir de la integración de las diferentes fuentes de información tradicionales (como encuestas y censos) con fuentes alternativas (como imágenes satelitales, registros administrativos, técnicas de big data, entre otros).

En Colombia, actualmente se cuenta con dos tipos de mediciones oficiales de la pobreza, estas medidas son la pobreza monetaria y la pobreza multidimensional. Dichos indicadores, cuentan con metodologías oficiales establecidas en el CONPES 150 de 2012. En particular, en el desarrollo de este trabajo solo será de interés el indicador del índice de pobreza multidimensional (IPM) que se obtuvo del Censo nacional de población y vivienda 2018 (CNPV 2018) y sus niveles de desagregación geográfico.

El enfoque empleado en la construcción del IPM censal está basado en la metodología de Alkire y Foster (Alkire, 2011), en la cual, se construye la medida basándose en cinco dimensiones (condiciones educativas del hogar, condiciones de la niñez y juventud, salud, trabajo, acceso a servicios públicos domiciliarios y condiciones de la vivienda) y 15 indicadores[1]. Sin embargo, los resultados de dicha implementación no son comparables ni reemplazan las mediciones oficiales del IPM realizadas por el DANE con periodicidad anual. Por lo tanto, el objetivo principal de dicho ejercicio, es el aprovechamiento de la información del CNPV 2018 y su alto grado de desagregación.

El aprovechamiento de la información censal, no se limitó a las desagregaciones municipales por domino (cabeceras municipales y centros poblados- rural disperso), ya que el CNPV cuenta con niveles inferiores a estos dominios, como lo son las manzanas en los centros urbanos y los sectores rurales en rural

[1] Ver Nota metodología censal pobreza municipal 2018

disperso de acuerdo a lo establecido en el marco geoestadístico nacional (MGN)[2] , lo cual permitió obtener visualizaciones del comportamiento del IPM dentro de las 1122 cabeceras municipales del país[3].

Estas implementaciones permiten a los tomadores de decisiones a nivel local, la generación e implementación de política pública de forma focalizada. Sin embargo, dichas mediciones cuentan con limitaciones en niveles rurales, en los cuales el operativo censal presentó dificultades, ya sea porque la zona fue visitada de forma parcial o no fue posible acceder a dicho territorio, debido a las condiciones climáticas, geográficas o de orden público. Adicionalmente, ya que el insumo es la fuente censal, la periodicidad de las mediciones se limitan a los años en los que se realicen las operaciones censales, que, según recomendaciones de Naciones Unidas, se deben realizar con una periodicidad de 10 años.

El presente documento, describe un ejercicio que pretende mitigar las limitaciones presentadas en el cálculo del IPM censal, así como disponer de dichas mediciones a niveles bajos de granularidad en periodos intercensales, , mediante la implementación de fuentes no tradicionales de información, como es el caso de las imágenes satelitales.

1.1 Antecedentes

Uno de los temas prioritarios para el DANE, tanto por que se encuentra enmarcado en los compromisos globales de la Agenda 2030 para el Desarrollo Sostenible, como por su importancia para el proceso de elaboración de políticas públicas, es la medición periódica de pobreza, bien sea monetaria o multidimensional. Prueba de la importancia de esta medición para Colombia es que actualmente se producen dos medidas oficiales de pobreza: la pobreza monetaria y la pobreza multidimensional.

Si bien se considera que el DANE cuenta con avances significativos en lo que respecta a la medición de pobreza a nivel departamental y en las ciudades principales del país, así como en lo que refiere a la medición de la pobreza multidimensional de fuente censal para todos los municipios, se ha identificado la necesidad de contar con mediciones periódicas de pobreza para los municipios que no son representativos en las encuestas de hogares.

Como se mencionó anteriormente, una herramienta para lograr suplir la medición periódica de pobreza es el uso de nuevas fuentes alternativas de datos. En particular, el uso de imágenes satelitales,

[2] Ver Marco geoestadístico nacional y GUÍA DE DESCARGA Y VISUALIZACIÓN
[3] Ver Medida de pobreza multidimensional de fuente censal - DANE

combinado con técnicas de big data y estadística, han demostrado ser efectivas para la medición de la pobreza en altos niveles de desagregación.

Dichas implementaciones parten de identificar cuatro aspectos fundamentales:

1. Disponer la mejor medida de la pobreza disponible, ya se sean fuentes censales o de encuestas de hogares al menor nivel de desagregación posible.
2. Determinar los insumos requeridos para el cálculo de la medida de pobreza, dichos insumos van desde información oficial, como encuestas de hogares, imágenes satelitales, registros de telefonía móvil entre muchos otros.
3. Establecer el tipo de modelamiento que permita obtener mediciones a los mismos niveles de desagregación disponibles o inferiores.
4. Obtener la métrica al nivel de desagregación deseado garantizando un nivel de desempeño óptimo.

En trabajos como (Steele, 2017) o (Heitmann, 2019), se implementan enfoques para le medición de la pobreza en Bangladesh, Ghana y Uganda, en los que las estadísticas oficiales producto de encuestas o mediciones como el PPI[4] son usadas como aproximaciones de la variable dependiente. A su vez, los insumos van desde encuestas de hogares, RS (RS remote sensing por sus siglas en inglés) y CMR (Call Detail Records); a escalas espaciales que van desde las divisiones político administrativas de cada país hasta polígonos de Voronoi, construidos con la ubicación y cobertura de las antenas de telefonía celular. Por último, el uso de modelos predictivos como BGM[5] o el más usado Random Forest (RF)[6]

2. Implementación

2.1 Contexto

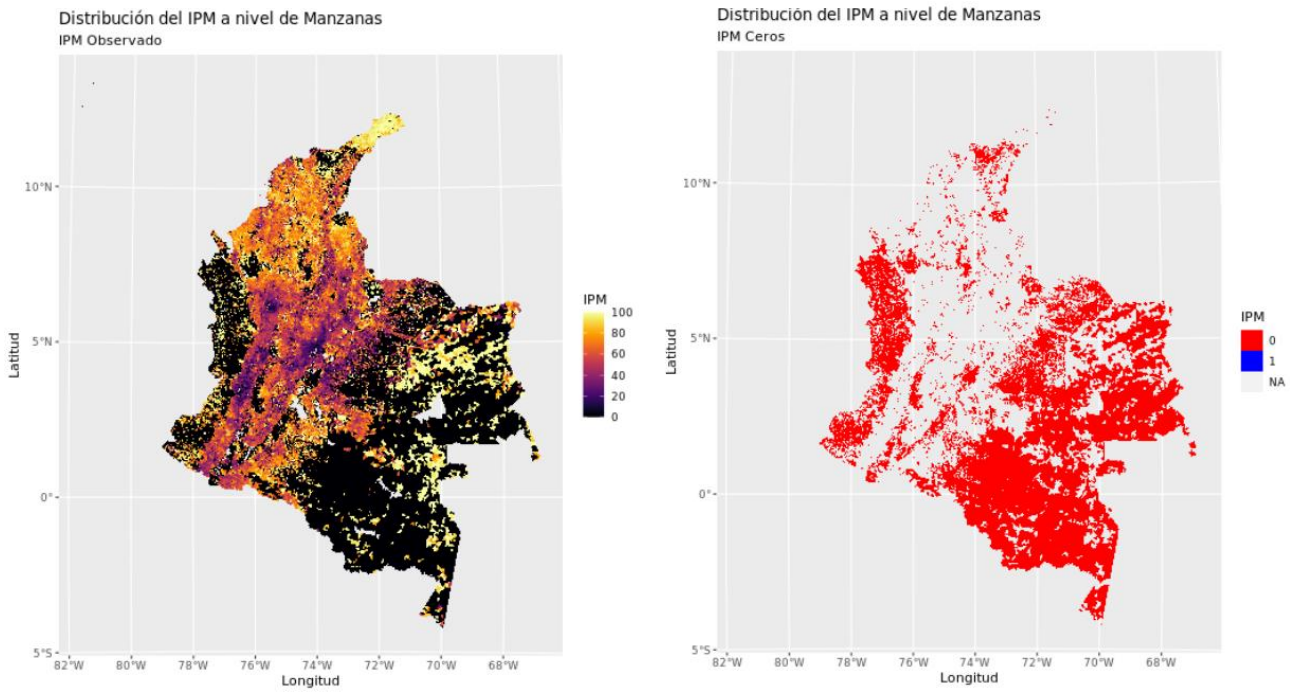
La medida de IPM a nivel del MGN presenta varios desafíos, en particular con las observaciones extremas del indicador, que resultan ser las manzanas o secciones rurales con valores del indicador en 0 o 1. Generalmente, estas medidas se presentan en áreas con poca información censal, en zonas urbanas y

[4] Ver Poverty Probability Index
[5] Ver Hierarchical Bayesian geostatistical models
[6] Ver (Breiman, 2001)

rurales de difícil acceso, dicha dificultad se presenta mayoritariamente en zonas rurales en las cuales el proceso de recolección de información censal fue realizado por rutas o convocatoria[7].

EL gráfico 1 muestra cómo se distribuye geográficamente el IPM censal, así como las áreas en las cuales se tienen extremos del indicador. Por tanto, las expectativas consisten en determinar una medición del IPM en estas zonas para el año 2018, así como una completa medición para periodos intercensales.

Gráfico 1. IPM censal 2018 a nivel manzana y sección rural



Fuente: DANE. Dirección de censos y demografía

2.2 Insumos

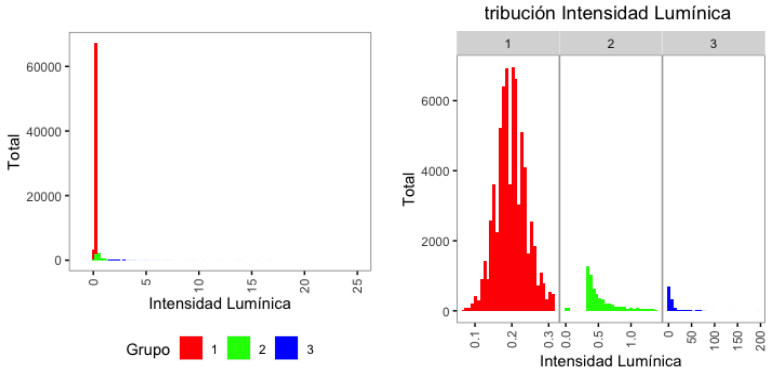
- Como variable dependiente, se usará la estimación del IPM censal a nivel de manzana y sección rural[8], es decir, se cuenta con una medida del IPM en cada una de las 513 421 manzanas o secciones rurales del MGN . Esta medición se tomará como la mejor medida oficial disponible, los modelos que serán implementados tendrán como resultado una medida del IPM a los mismos niveles de desagregación.

[7] Ver Ficha Metodológica Censo Nacional de Población y Vivienda 2018
[8] Resultados cabeceras municipales visor

- Indicadores censales: se cuenta con 54 indicadores censales para todas las manzanas o secciones rurales del MGN, la información disponible se encuentra a nivel de persona (proporción de alfabetismo, pertenencia étnica, rangos etarios, etc), hogar (tamaño promedio de hogar) y vivienda (proporción de viviendas por acceso a servicios públicos, materiales de pisos y techos, etc).
- MGN, se usan los niveles geográficos de manzana censal y sección rural, dicha información está disponible en MGN.
- Se descargaron 77 979 imágenes satelitales que cubren la totalidad del territorio colombiano para el año 2018. La fuente de dicha información proviene del satélite *SENTINEL-2*[9]. Las imágenes para este ejercicio se obtuvieron con ayuda de la plataforma de *Google, Google Earth Engine* (GEE), siguiendo los pasos a continuación: 1) Crear una cuadrícula regular sobre el territorio colombiano, donde cada punto representa un centroide y la distancia de cada punto al otro es de 3840 m^2. 2) GEE utiliza cada centroide para generar una imagen centrada en ese punto, el área de cada imagen se obtiene a partir del número de pixeles y el área en metros de cada pixel, en este caso se tienen 384*384 pixeles y cada pixel cubre un área de 100m^2 para un total de 14 74 Km^2. 3) Cada imagen es el resultado de concatenar varios trozos de diferentes imágenes, esto es, GEE toma imágenes cada mes y las combina para generar una sola imagen. Adicionalmente, para mejorar la calidad de la mismas, se decidió definir el porcentaje de nubosidad en 60%, es decir, se eliminan las imágenes que estén mayormente cubiertas por nubes.
- Para cada centroide se descargó la información de la intensidad lumínica nocturna con la ayuda de GEE y "*Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB)*[10]". Una de las características más importante es la detección de la luz eléctrica presente en la superficie de la tierra, la mayoría proveniente de asentamientos humanos. El proceso para obtener la información consistió en: 1) Guardar el valor mediano de intensidad lumínica de cada píxel de todas las imágenes disponibles en cada centroide, 2) sumar todos los valores medianos de todos los pixeles en cada imagen y 3) promediar estos valores para obtener un valor de intensidad lumínica nocturna para cada centroide. Estos valores se agruparon en tres categorías correspondientes a Baja =1, Media = 2 y Alta = 3 con la ayuda de un modelo de Mixtura Gaussiana (GMM) presentado en (Fraley, 2002); cada una con una frecuencia de 70 757, 5 942 y 1 280 imágenes respectivamente. Este método permite crear grupos diferenciados con distribución normal univariada con media y varianza diferente como se muestra en el gráfico 2.

[9] Ver THE EUROPEAN SPACE AGENCY
[10] Ver Visible Infrared Imaging Radiometer Suite (VIIRS)

Gráfico 2. Distribución de intensidad lumínica por grupo



Fuente: DANE. Dirección de censos y demografía

3. Modelamiento

3.1 Versión 1

Debido al número de variables censales que se procesaron a nivel de manzana (52 variables), se ajustó un modelo *Gradient Boosting Tree Regression* (GBTR[11]) para identificar las variables más relacionadas con el IPM, con el fin de eliminar aquellas que no le aportan varianza al modelo. De esta forma, se seleccionaron las variables que alcanzaran una importancia superior al 1% seleccionando las siguientes:

- Proporción de viviendas con material de piso en baldosa, vinilo, tableta, ladrillo o lamina.
- Proporción de viviendas con material de piso en tierra, arena o barro.
- Proporción de viviendas con material de piso cemento o gravilla.
- Proporción de viviendas con alcantarillado.
- Proporción de personas con máximo nivel educativo: jardín.
- Proporción de personas con máximo nivel educativo: primero de primaria.
- Proporción de personas con máximo nivel educativo: cuarto de primaria.
- Proporción de personas con máximo nivel educativo: quinto de secundaria.
- Proporción de personas con máximo nivel educativo: primero de secundaria.
- Proporción de personas con máximo nivel educativo: segundo de secundaria.
- Proporción de personas que saben leer y escribir.
- Proporción de personas con 10 a 19 años de edad.
- Tamaño promedio del hogar.

[11] Ver (Friedman, 2002)

Así mismo, adicional a estas variables, se incluyó la clase geográfica (cabecera=1, centros poblados = 2 y rural disperso=3) y la región[12] a la que pertenece la manzana.

Una vez establecidas las covariables más importantes para estimar el IPM, como un primer enfoque se usaron los primeros cinco componentes principales obtenidos a partir de las variables censales anteriormente descritas eliminando los registros que tenían IPM observado igual a 0 o 1, ya que la inclusión de estos valores genera que las estimaciones de los modelos bajen su precisión en términos del coeficiente de determinación (R^2) y la raíz del error cuadrático medio (RMSE). La tabla 1 muestra el porcentaje de variabilidad explicada por cada uno de las componentes resultantes, en el que el porcentaje total de varibilidad explicado por estas componentes es del 99,4%.

Tabla 1. Porcentaje de variabilidad explicada por componente principal de las variables con importancia superior al 1%

Componente Principal	Varianza Explicada (%)
1	74.3
2	17.7
3	5.5
4	1.2
5	0.7

Fuente: DANE. Dirección de censos y demografía

De esta forma, se modeló la transformación logit[13] del IPM como variable dependiente y como variables independientes se tomaron las componentes principales seleccionadas, la regionalización de los municipios y la clase geográfica. Adicionalmente, para la partición de los datos en el conjunto de prueba y entrenamiento de los modelos se eliminaron los registros que tenían IPM igual a 1 o 0, obteniendo de forma aleatoria 303 497 registros para el conjunto de entrenamiento y 75 810 para el de prueba. Se ajustaron dos modelos[14] , GBTR y RF, para los cuales se tienen los resultados[15] presentados en la tabla 2.

[12] Ver anexo 1
[13] $logit(IPM) = ln\,(IPM/(1 - IPM))$
[14] Hiperparametros ajustados mediante validación cruzada
[15] Métricas de desempeño calculadas sobre el conjunto de prueba

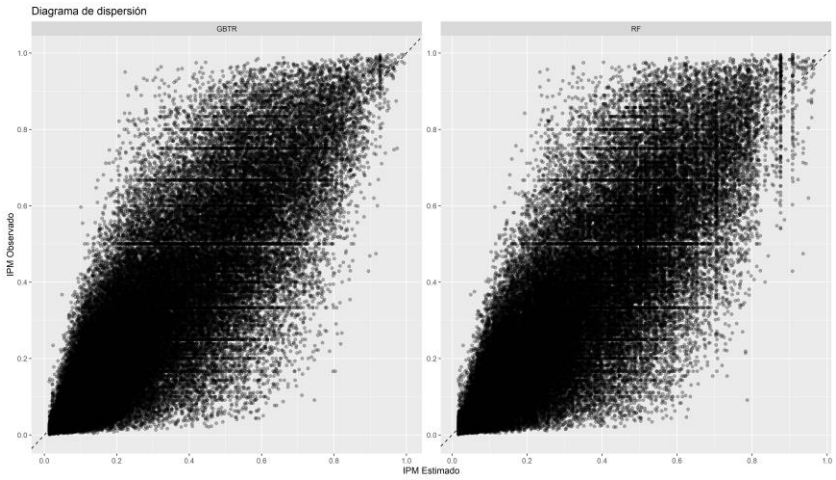
Tabla 2. Métricas desempeño modelos GBTR y RF usando covariados censales

Medida de Rendimiento	GBTR	RF
R^2	0,6789	0,6621
RMSE	0,7818	0,8095

Fuente: DANE. Dirección de censos y demografía

La varianza explicada por los modelos GBTR y RF es del 67.89% y 66.21% respectivamente, lo cual permite evidenciar un comportamiento de ojiva en los diagramas de dispersión entre el IPM observado versus las predicciones de los modelos, así como niveles de variabilidad considerablemente bajos. El gráfico 3 muestra los comportamientos descritos.

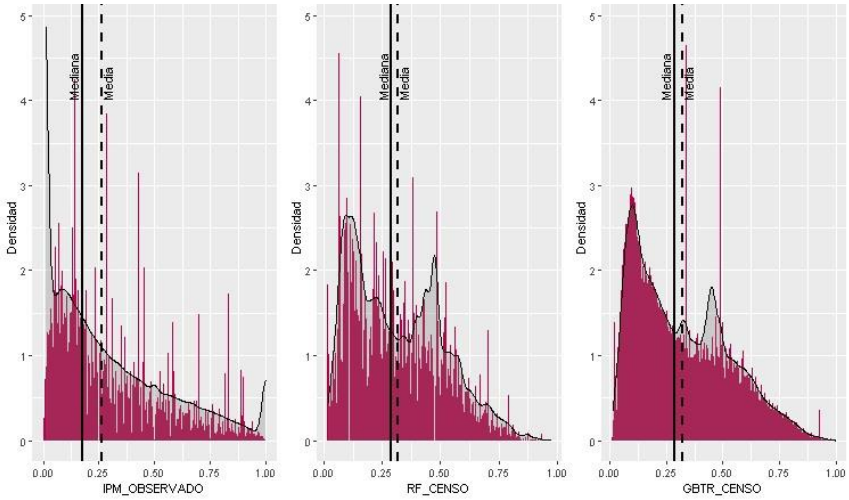
Gráfico 3. Diagramas de densidad IPM observado vs Predicciones del IPM- Distribución geográfica de las predicciones



Fuente: DANE. Dirección de censos y demografía

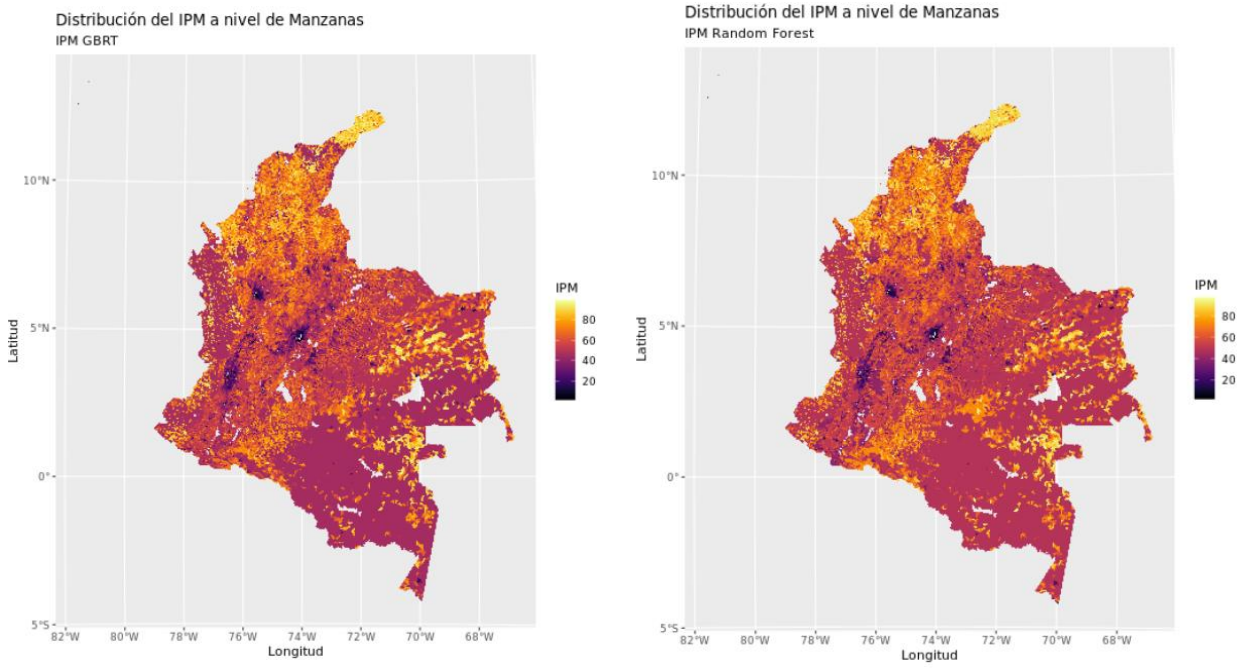
Las predicciones generadas por los anteriores ajustes, reproducen satisfactoriamente el comportamiento del IPM para las zonas de mayor densidad poblacional, como lo son Bogotá, Medellín y Cali. Sin embargo, la distribución de las predicciones, que se observa en el gráfico 4, muestra un comportamiento apuntado para valores del IPM cercanos a 0.50, lo cual es un indicio de la poca información que aportan los indicadores censales en las áreas rurales. El gráfico 5 muetra la distribución geografica de la predicciones, las cuales se comportan de forma uniforme en gran parte del andén Pacifico y los departamentos de la Orinoquia y Amazonia.

Gráfico 4. Diagramas de densidad IPM observado vs Predicciones del IPM



Fuente: DANE. Dirección de censos y demografía

Gráfico 5. Distribución geográfica de las predicciones versión 1



Fuente: DANE. Dirección de censos y demografía

3.2 Versión 2

Para este escenario se consideran dentro del proceso de reducción de dimensionalidad y posterior partición del conjunto en entrenamiento y prueba, a todas las observaciones del IPM en 1 y las observaciones en 0 para aquellos territorios en los que existe población, esto con el fin de aportar información al modelo y conseguir predicciones que rompan con el esquema de distribución uniforme presentado en la anterior implementación.

De acuerdo con estas consideraciones, se tienen 316 377 registros para el conjunto de entrenamiento y 78 794 para el de prueba. La tabla 3 presenta el porcentaje de variabilidad explicado por cada una de las 5 componentes principales seleccionadas, está selección consigue explicar el 76,11% de variabilidad.

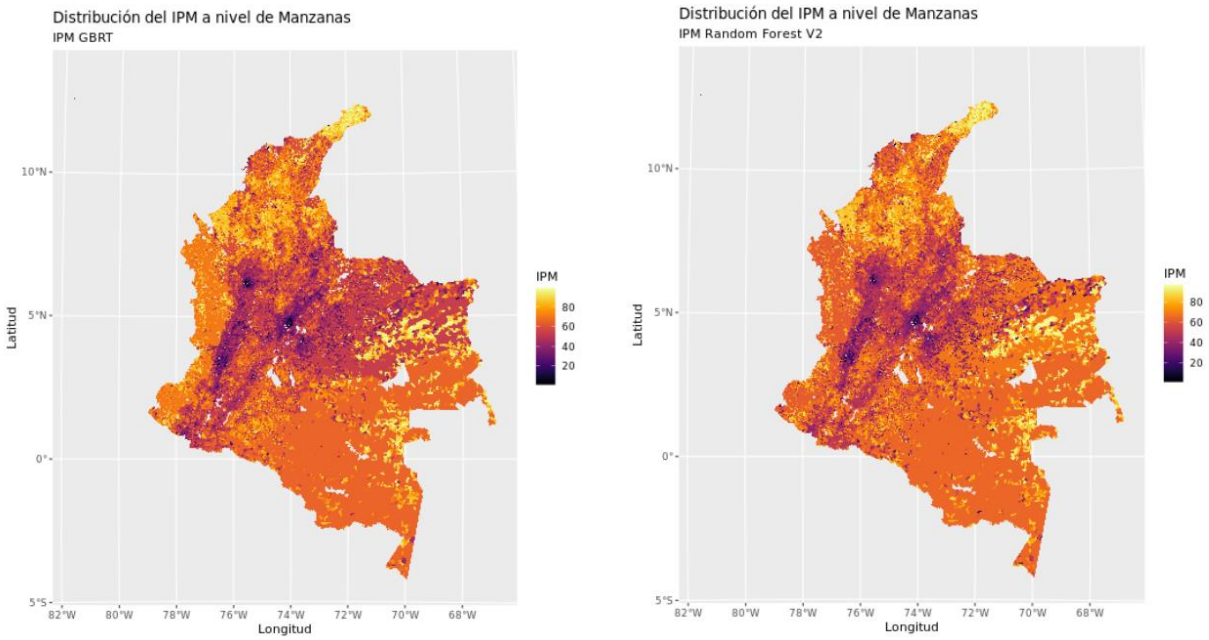
Tabla 3. Porcentaje de variabilidad explicada excluyendo valores del IMP en 0 para áreas sin población.

Componente Principal	Varianza Explicada (%)
1	28,78
2	19,46
3	10,87
4	9,03
5	7,98

Fuente: DANE. Dirección de censos y demografía

Los resultados de dicha implementación alcanzan un coeficiente de determinación de 65.37% y 62.81% para los para los modelos GBRT y RF respectivamente; y métricas de RMSE de 1,1898 y 1,233. Estas pérdidas porcentuales en los R^2 y aumentos de los RMSE son explicados por la perdida en variabilidad explicada en el análisis de componentes principales, sin embargo, los resultados son alentadores, debido a que, así exista perdida de información, los modelos consiguen tener desempeños aceptables. La distribución geográfica que muestra el gráfico 6, cambia los niveles de la predicción, es decir aumenta las zonas con niveles altos del IPM, debido a la inoculación a observaciones del IPM en 1. No obstante, este cambio de nivel sigue presentando un comportamiento uniforme en las mismas zonas que la implementación anterior.

Gráfico 6. Distribución geográfica de las predicciones modelo GBTR y RF versión 2



Fuente: DANE. Dirección de censos y demografía

3.3 Versión 3

A diferencia de los indicadores censales, las imágenes satelitales necesitan procesamientos de mayor complejidad y costo computacional, por tal motivo, este ejercicio consideró utilizar un modelo de redes neuronales convolucionales, con el fin de clasificar y extraer indicadores que puedan ser integrados al MGN para luego modelar y predecir el comportamiento del IPM. El objetivo de la red es clasificar las imágenes de acuerdo con los tres niveles creados anteriormente de intensidad lumínica, es decir: Baja, Media y Alta. El proceso inicia con la ingesta de las imágenes en la capa de entrada, luego, la imagen pasa por cada una de las capas ocultas hasta que en la última capa se genera un vector de dimensión 3x1 que contiene las probabilidades de que la imagen pertenezca a cada una de las categorías, la predicción final será la categoría que tenga la probabilidad más alta.

Algunos ejemplos de las imágenes que ingresan al proceso de clasificación y extracción se muestran en la figura 1.

Figura 1. : Imágenes extraídas de satélite tomadas en Colombia en año 2018 por SENTINEL-2. De izquierda a derecha la clasificación para cada imagen es: Alta, Media y Baja respectivamente

Fuente: DANE. https://sentinel.esa.int/web/sentinel/home

Para llevar a cabo el entrenamiento de esta red se tuvieron varios factores importantes como lo son:

- Para optimizar los tiempos de procesamiento, se decidió reutilizar los pesos *(w_i)* de otro modelo de redes convoluciones denominado desarrollo *ResNet34*[16], el cual fue entrenado con un conjunto más grande de imágenes.
- Inicialmente se corrieron 10 *epochs*[17], es decir, se corre el modelo 10 veces sobre los mismos datos y en cada *epoch* se van cambiando los pesos iniciales *(w_i)*.
- Los porcentajes de separación del conjunto de entrenamiento y prueba son 80% y 20% respectivamente, sin embargo, se configuró el modelo de tal manera que en cada *epoch* se conservara un 20% de la muestra de entrenamiento para validar los resultados.
- Las métricas que se consideraron en este ejercicio de clasificación es el *CrossEntropyLoss*[18] y el *error_rate*[19], la idea es conservar los parámetros del modelo que en alguno de los 10 *epoch* dé los menores resultados de estas dos métricas.
- Para aumentar la variabilidad en el modelo se decidió considerar trasformaciones de las imágenes de tal manera que no se altere la respectiva clasificación o etiqueta de cada una, en este caso se realizaron giros sobre el eje vertical y horizontal (efecto espejo), cambios en el contraste de la imagen y efectos de dihedral, la cual es una combinación de un giro de 90 grados y un giro sobre el eje x o y.
- Para encontrar el parámetro óptimo de *learning_rate* se seleccionó el *epoch* que genera los mínimos valores de *CrossEntropyLoss* y el *error_rate*, se congelan los valores de las últimas 2 capas de la red y se le dan valores al *learning_rate*, la idea es encontrar el valor del *learning_rate*

[16] Ver (He, 2016)
[17] Ver What is an Epoch?
[18] Ver CROSSENTROPYLOSS
[19] Ver Accuracy in Machine Learning

que menor *error_rate* genere y con esos valores se ejecutan 5 *epochs* adicionales para asegurarnos que los resultados que se generen sean óptimos.

Una vez la red está entrenada y las métricas de desempeño son adecuadas, se procede a extraer las características visuales (*visual features*) de las imágenes, para ello se usa la técnica "*transfer learning*", la cual permite cortar la red en la capa n-1 y obtener los covariados, que para este caso, resultan en una matriz que contiene una fila por cada imagen que se usó para entrenar el modelo, y 512 columnas las cuales contienen toda la información que la red consigue aprender desde las capas iniciales, identificando formas sencillas como círculos, bordes horizontales y verticales hasta las últimas capas de la red que se encargan de identificar objetos más complejos como los techos de las casas, ríos, carreteras y bosques.

Dicha matriz de covariados, tiene la finalidad de ser integrada al MGN y ser usada como las variables independientes de un modelo de predicción del IPM, en reemplazo de los indicadores censales, con la ventaja de contener información de todo el territorio nacional, en particular, las áreas con poca o nula información censal. Sin embargo, el proceso de integrar esta fuente de información al MGN no puede ser realizado de forma directa, en primer lugar, debido al elevado número de variables y, en segundo lugar, los 512 covariados se encuentran localizados en los centroides de cada una de las 77 979 imágenes.

Como primera propuesta, se reduce la dimensionalidad de la matriz empleando análisis de componentes principales, consiguiendo explicar el 99.9% de la variabilidad con las primeras 5 componentes principales. Posteriormente, se aplica un proceso de interpolación que permita tomar la información de las 5 componentes y ubicarlos en los centroides del MGN, para ello se tienen las siguientes consideraciones:

- Debido a la distribución regular y densidad de los puntos de "*muestra*", se opta por emplear el método de interpolación determinístico por vecinos naturales (*natural neighbors*), para la generación de la superficie de continuidad base. Este método es útil para datos de entrada densos y espaciados regularmente, y se basa en hallar un subconjunto de muestras más cercano a un punto central para aplicar ponderaciones a partir de áreas proporcionales. Esto hace que el método sea local, buscando asegurar que la interpolación esté dentro del rango de las muestras utilizadas.
- Luego de generar las superficies base de interpolación para cada variable, se procede a realizar la estimación de valores sobre los centroides de las manzanas/secciones rurales. Para ello, se opta por una siguiente interpolación bilineal, con el fin de garantizar una mayor "suavización" y transición de los valores estimados, evitando cambios abruptos. En este sentido, el valor de cada

punto se determina a partir del promedio de la distancia ponderada de las celdas más cercanas de la superficie base.

- Debido a los denominados efectos de borde propios de los métodos de interpolación, donde algunas zonas en los extremos de la región no logran ser interpoladas para la estimación de valores por la escasez de muestras, se aplicó el valor del punto más cercano. Estos casos se identificarán con valores repetidos de estimación, pero corresponden a menos del 0,2% de los centroides, tanto urbanos como rurales.

Finalmente, se usan dos enfoques para estimar el IPM, enfoque 1: incluyendo los valores de 0 y 1 del IPM, se estima directamente el logit del IPM utilizando los componentes principales obtenidos a través de las imágenes satelitales y la interpolación de los mismos a nivel de manzana. Enfoque 2: excluyendo los valores de 0 y 1 del PM, se implementa una simulación de Montecarlo con 100[20] iteraciones del enfoque 1, seleccionando en cada ciclo una configuración de entrenamiento (80% de los datos) y prueba (20% de los datos), para luego tomar el promedio de las predicciones en cada área, la fortaleza de este segundo enfoque radica en la obtención de la distribución empírica de las predicciones y por ende permite identificar la variabilidad de las mismas. En la tabla 4, se muestran las métricas de desempeño de los modelos ajustados por cada enfoque, evidenciando un mejor desempeño del enfoque 2 ya que consigue explicar un 32.59% y 40.47% más de variabilidad del IPM que los modelos de enfoque 1, así como una reducción considerable en términos de RSME de las predicciones.

Tabla 4. : Métricas de desempeño modelos GBTR y RF por enfoque 1 (E1) y enfoque 2 (E2)

Modelo	R^2	RSME
GBTR E1	0.203	2.984
RF E1	0.171	3.042
GBTR E2	0.5289	0.9555
RF E2	0.5757	0.9067

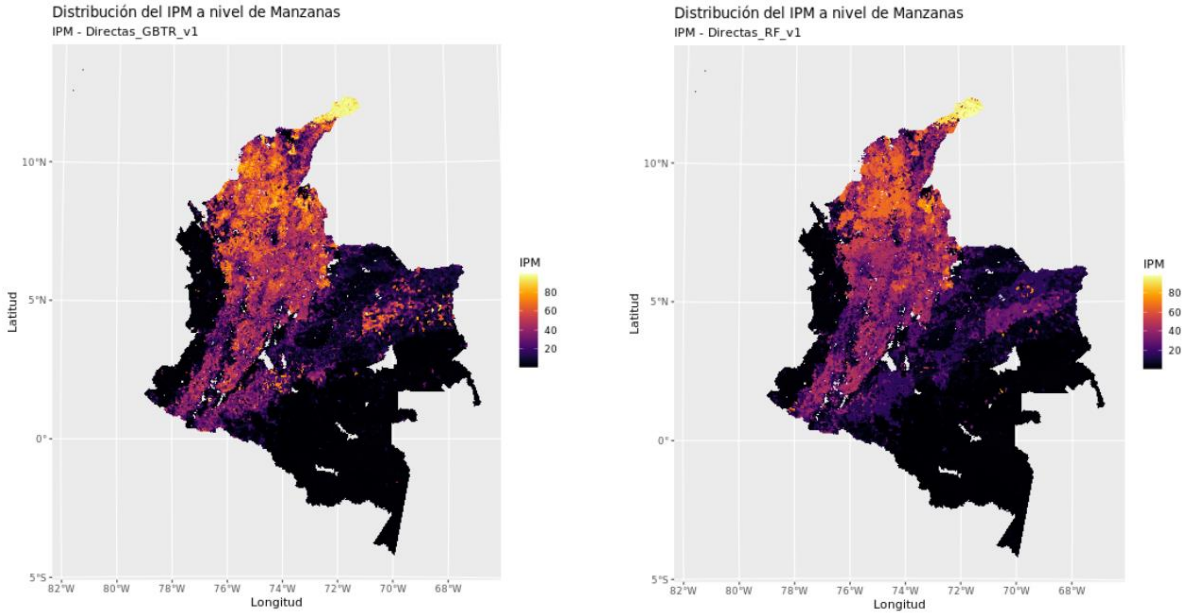
Fuente: DANE. Dirección de censos y demografía

El grafico 7 muestra el comportamiento de las predicciones de los modelos GBTR y RF bajo el enfoque 1, en donde se refleja el bajo rendimiento de las métricas de desempeño alcanzadas bajo este escenario, en particular en las áreas rurales, mostrando un patrón uniforme y considerablemente bajo en los niveles de IPM, debido principalmente a alto porcentaje de áreas con valor 0 en su IPM observado. En contraste, la distribución de las predicciones en el enfoque 2, que se observa en el gráfico 3.6, no presentan un patrón uniforme en los departamentos amazónicos, resaltado el comportamiento de Putumayo, Caquetá, Guaviare, Meta, Casanare y Arauca, como departamento de la Amazonia y Orinoquia en los

[20] La escogencia de este número de iteraciones se debe a la capacidad computacional disponible para ejecutar los modelos.

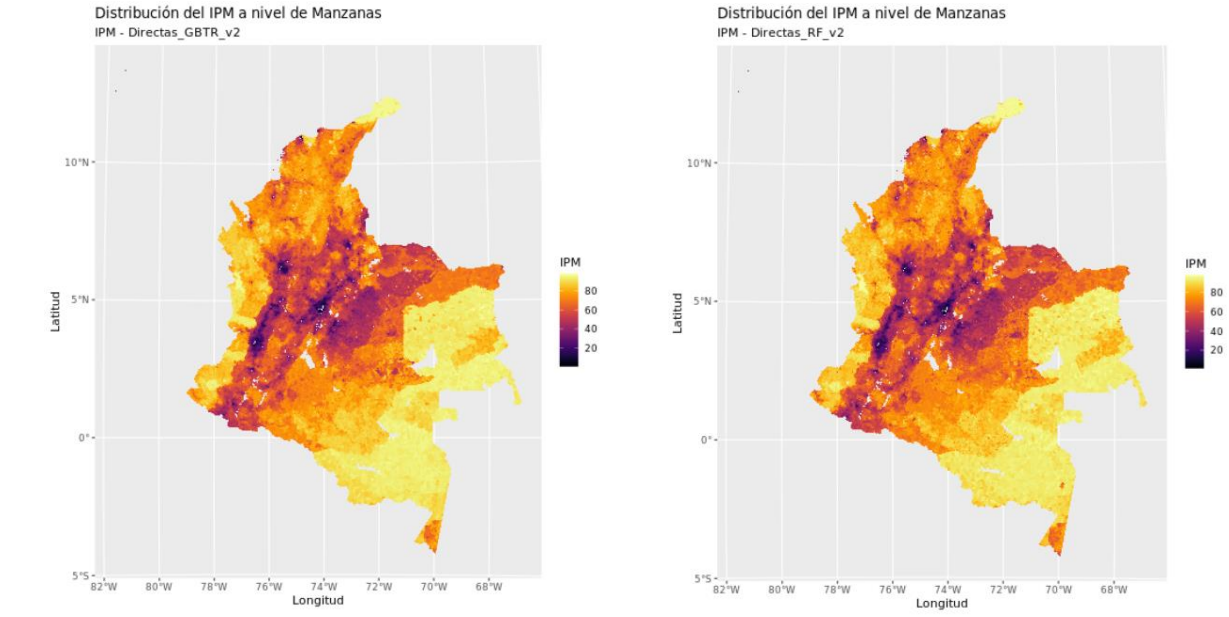
cuales se observan comportamientos diferenciales en los niveles del IPM, al igual que los departamentos del Anden Pacífico, en particular Chocó y Nariño.

Gráfico 7. Predicciones IPM enfoque 1



Fuente: DANE. Dirección de censos y demografía

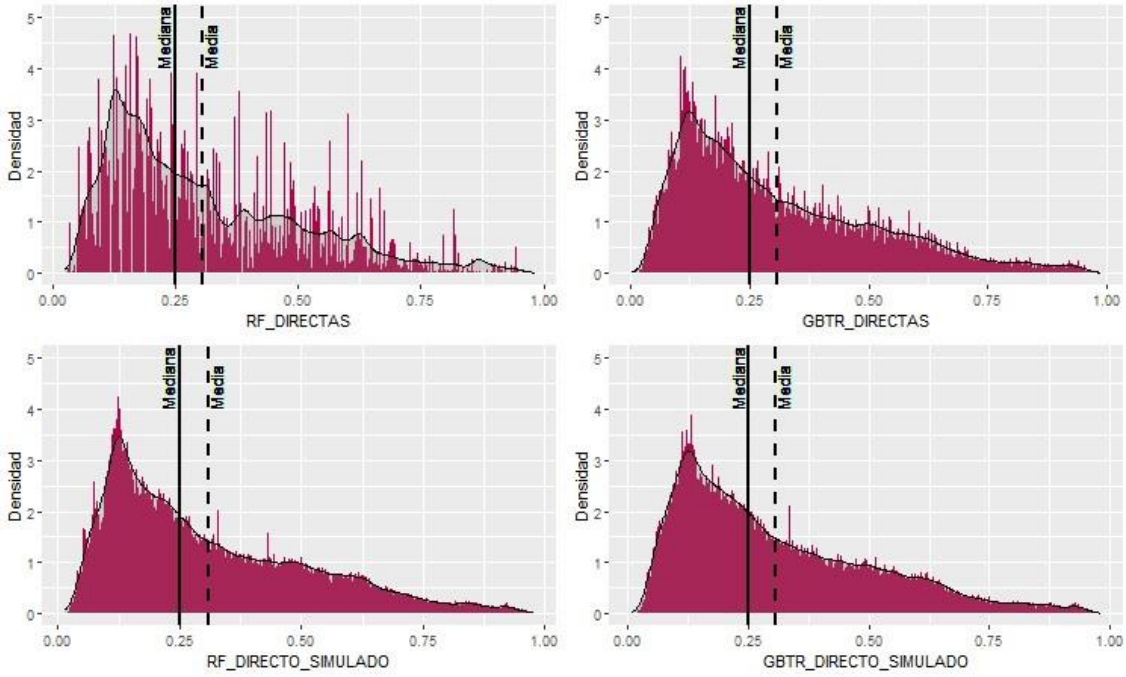
Gráfico 8. Predicciones IPM enfoque 2



Fuente: DANE. Dirección de censos y demografía

El gráfico 9 muestra las densidades generadas mediante el enfoque 1 (RF_DIRECTAS y GBTR_DIRECTAS) y en enfoque 2 (RF_DIRECTAS_SIMULADO y GBTR_DIRECTAS_SIMULADO), destacando que todos los modelos predicen de forma similar el comportamiento medio del IPM, no obstante, se observa una mayor heterogeneidad bajo el enfoque 1, debido principalmente a los mayores niveles en el RMSE que presenta dichas predicciones.

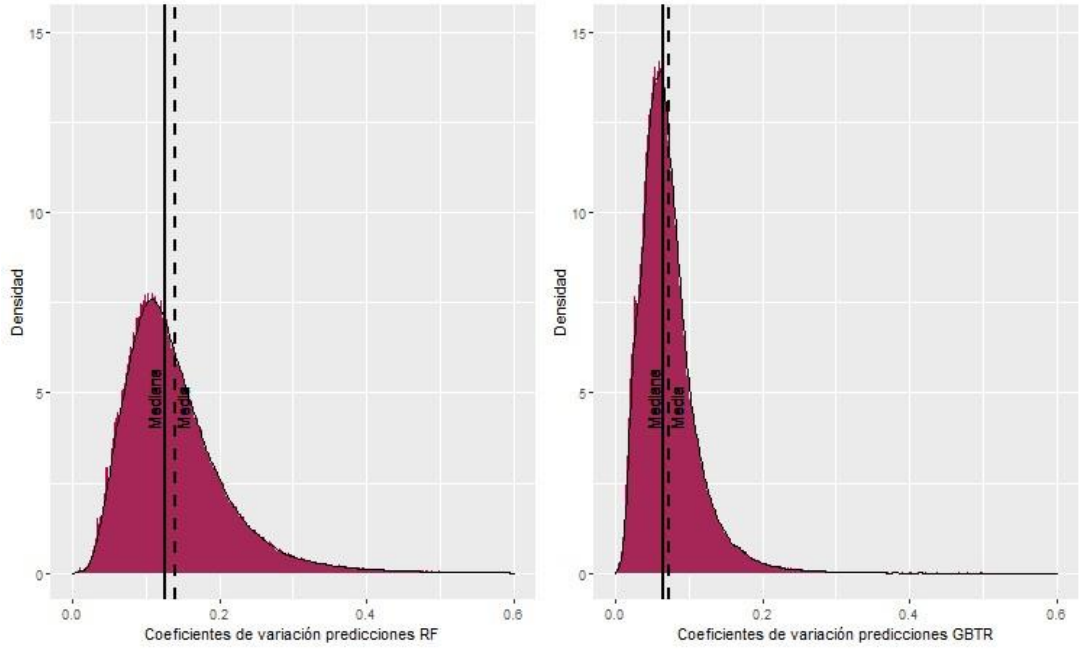
Gráfico 9. Predicciones RF_DIRECTAS y GBTR_DIRECTAS mediante enfoque 1 - Predicciones RF_DIRECTAS y GBTR_DIRECTAS mediante enfoque 2



Fuente: DANE. Dirección de censos y demografía

Como ya fue mencionado, una de las ventajas que presenta en modelmiento bajo el enfoque 2, es la capacidad de medir la varibilidad que estan presentando las predicciones mediante el calculo de la distribuciones empiricas, esto permite obtner un coeficiente de variación (cv) para cada area en el que se aplica el modelo. Aunque se tiene una gran limitacón en el número de iteraciones implementado, el gráfico 10 muestra la distribución de dichos metricas, evidenciando niveles de variabilidad considerablemente bajos para los dos modelos, no obstante, el modelo GBTR genera los mejores estimaciones, ya que sus estimaciones cuentan con menor varibilidad.

Gráfico 10. Distribuciones cv de las estimaciones de los modelos en el enfoque 2



Fuente: DANE. Dirección de censos y demografía

4. Conclusiones y recomendaciones

- Las métricas de desempeño presentadas por el modelo GBTR mediante simulación de Montecarlo son satisfactorias, ya que se consiguen explicar el 52.9% de la variabilidad del indicador IPM solo usando la reducción de dimensionalidad de 512 covariados extraídos de las imágenes satelitales. Adicionalmente, resulta ser la mejor estimación que se obtuvo ya que presenta menor variabilidad en sus predicciones.
- La propuesta implementada, describe de forma similar el comportamiento del IPM censal a nivel municipal, por tanto, se espera que una medida resumen de las predicciones a nivel de manzana y sección rural dentro del municipio, permita determinar el indicador de IPM a nivel municipal en cualquier para cualquier año del que se tenga informacional satelital.

- Se consideran dos escenarios para mejorar las métricas de desempeño, el primero consiste en determinar un subconjunto de los 512 covariados que logre aumentar el desempeño, y el segundo, aumentar el número de iteraciones del ciclo de Montecarlo, implementando un diseño muestral para la selección de del conjunto de entrenamiento y prueba.
- Al momento de iniciar el proyecto se tenían muchas expectativas sobre los resultados que se podían obtener y sobre todo, la extensión de esta misma metodología para la medición de otros indicadores con la mima importancia que la pobreza como lo son a) Indicadores de educación, b) Indicadores de salud; sin embargo, a lo largo del camino se hizo evidente que esta metodología sólo aplica para medir aspectos relacionados con pobreza, debido principalmente, a la estrecha relación que existe entre el material de los techos de las casas, el material de las carreteras, la presencia de bosques y de cuerpos de agua con los niveles de pobreza de una comunidad.
- En este ejercicio de pobreza, todas las fuentes de información son públicas[21], en este orden de ideas, cualquier persona puede descargar los mismos insumos y obtener resultados similares; sin embargo, un obstáculo es la calidad de las imágenes, que influye significativamente en las estimaciones finales, es decir, factores como los porcentajes de nubosidad y la cantidad de metros por pixel determinan la nitidez y claridad de las imágenes.
- El compromiso de obtener estimaciones de pobreza a niveles geográficos bajos desató nuevos retos que se convirtieron en aportes a los métodos conocidos hasta el momento; para superar estos, fue necesario unir los esfuerzos de varias oficinas del DANE y de sus expertos temáticos.

5. Visualización

Los resultados para algunas de las versiones, pueden ser vistos de manera focalizada en la plataforma https://visores.dane.gov.co/visor-nuevo-ipm/, que fue construido por la Dirección de Geoestadística y se alimenta de los resultados de este ejercicio.

[21] La información del CNPV no es de dominio publico

6. Agradecimientos

Este trabajo fue posible gracias el apoyo metrológico recibido en los convenios que el DANE mantiene con Paris 21 y D4N. Así, como el apoyo de Min Tic, quien suministró la infraestructura tecnología en la que se implementaron los modelos predictivos.

A los expertos de las direcciones técnicas DIG, DCD y por último a la dirección general quien incentivo la implementación de estadísticas experimentales y el uso de fuentes no tradicionales de información.

7. Bibliografía

- Alkire, S. &. (2011). Counting and multidimensional poverty measurement. Journal of public economics, 95(7-8), 476-487.
- Fraley, C. &. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association, 97(458), 611-631.
- Friedman, J. H. (2002). Stochastic gradient boosting. . Computational statistics & data analysis, 38(4), 367-378.
- Heitmann, S. &. (2019). Poverty Estimation with Satellite Imagery at Neighborhood Levels: Results and Lessons for Financial Inclusion from Ghana and Uganda. The World Bank, No. 137300, pp. 1-29.
- Steele, J. E. (2017). Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface, 14(127), 20160690.

8. Anexo 1

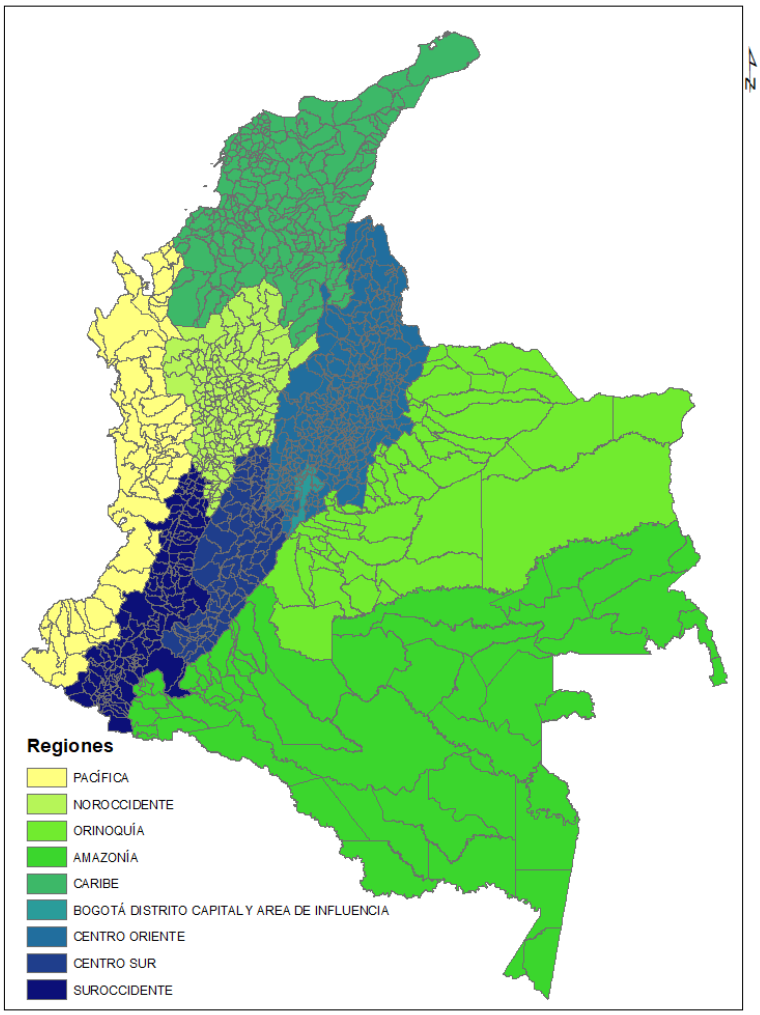
La regionalización empleada, organiza los municipios del país alrededor de regiones que incluyen municipios con patrones socioculturales y territoriales compartidos.

Etiqueta	Región	Descripción
1	PACÍFICA	La región se organiza a través de municipios que pertenecen al Andén Pacífico y que incluyen municipios de los departamentos de Chocó, Antioquia, Risaralda, Valle del Cauca, Cauca y Nariño. Es una región con alta prevalencia de población afro, entre ella, población que ha ocupado históricamente zonas de las cuencas del pacífico y que a través de prácticas tradicionales y organizativas se organiza a través de consejos comunitarios en territorios colectivos de comunidades negras.
2	NOROCCIDENTE	Región que incluye municipios que pertenecen a la zona influencia de la colonización antioqueña. Parte de su población afro ha estado en la zona desde la época de la colonia y otra ha migrado de la región pacífica.
3	ORINOQUÍA	La región articula los departamentos que se ubican en los llanos orientales de Colombia. Adicionalmente cuenta con baja prevalencia de población Afro que generalmente ha llegado a las zonas por procesos migratorios recientes.
4	AMAZONÍA	Departamentos localizados en la selva húmeda tropical de la cuenca del río Amazonas, con baja prevalencia de la población afro que llegó a las zonas a través de procesos migratorios recientes.
5	CARIBE	Departamentos que comprenden a los departamentos del norte del país con influencia del Caribe, con presencia significativa de población Afro, debido al tráfico trasatlántico de población africana esclavizada en épocas coloniales a través del puerto de Cartagena.
6	DISTRITO CAPITAL Y ÁREA DE INFLUENCIA	Comprende a Bogotá Distrito Capital, y a los municipios que son su zona de influencia por constituirse en municipios dormitorio de la capital. Cuenta con una población afro significativa principalmente en Bogotá y Soacha, aunque cuenten con porcentajes de participación relativamente bajos por al alto volumen de población que no se reconoce dentro de un grupo étnico.
7	CENTRO ORIENTE	Departamentos que cubren la zona centro-norte de la cordillera oriental, con procesos prolongados de mestizaje entre población indígena y de descendencia española, y baja prevalencia de población Afro, cuya presencia obedece a procesos migratorios recientes.
8	CENTRO SUR	Departamentos que corresponden al denominado Tolima grande ubicado sobre las cordillera central y oriental, y el valle del río Magdalena, con procesos prolongados de mestizaje entre indígenas y descendientes de españoles, y baja prevalencia de población afro.

9	SUROCCIDENTE	Municipios del suroccidente del país ubicados en zona andina y el valle del río Cauca, con alta presencia de población indígena y afro.

El gráfico 11 muestra la conformación de las de las regiones descritas

Gráfico 11. Distribuciones cv de las estimaciones de los modelos del enfoque 2



Fuente: DANE. Dirección de censos y demografía