DANE
INFORMACIÓN PARA TODOS

Predicción del IPM con imágenes satelitales

El futuro es de todos
Gobierno de Colombia

DANE
INFORMACIÓN PARA TODOS

Antecedentes

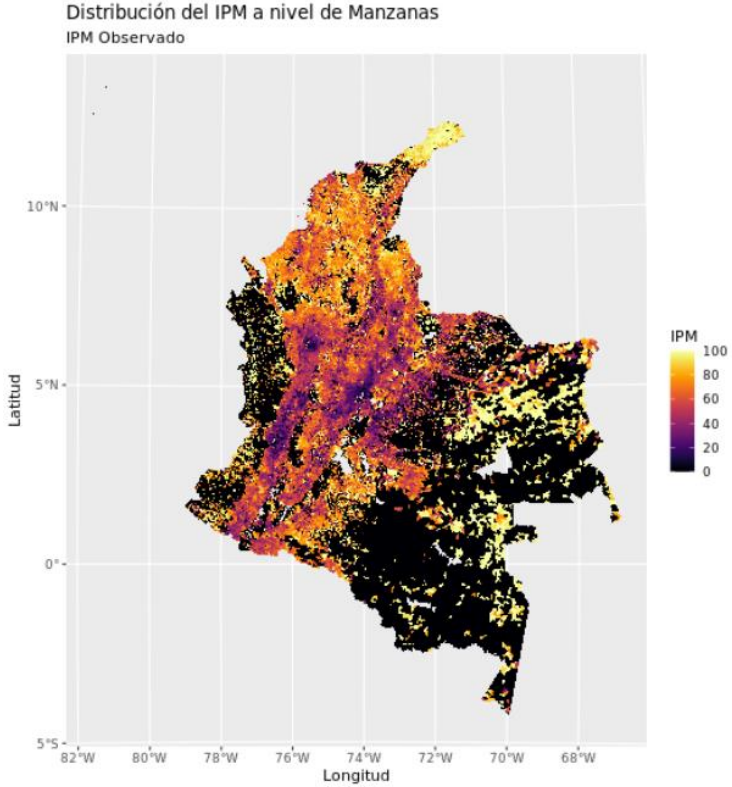
El futuro es de todos
Gobierno de Colombia

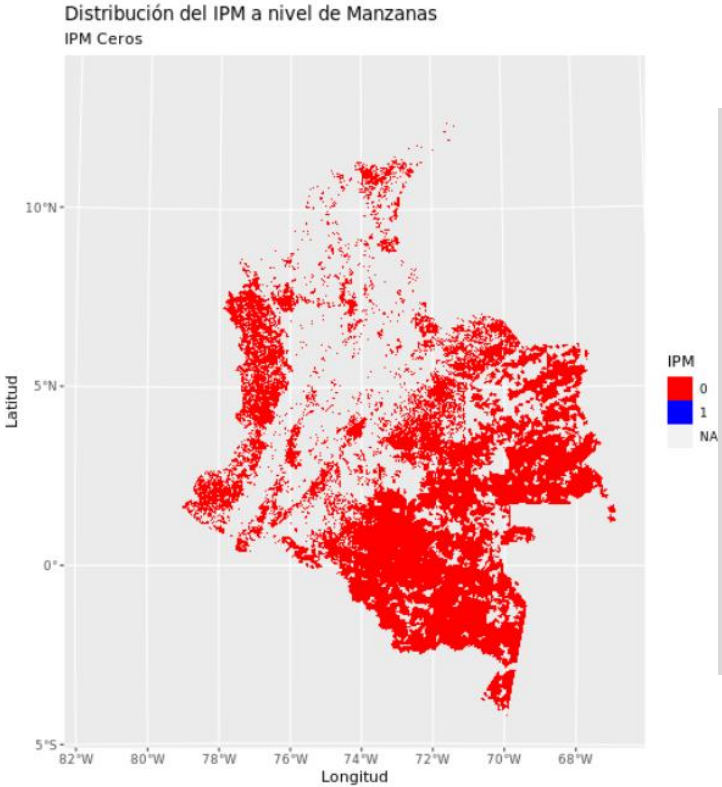
La medida del IPM de fuente Censal* se encuentra actualmente publicada y desagregada para municipios y a nivel de manzana para las cabeceras municipales.

No obstante, es de interés obtener esta medida, no solo para las cabeceras municipales sino también para las zonas rurales.

* Ver Boletín Técnico

IPM fuente censal a nivel manzana y sección rural





El IPM censal contiene un medición de 0 y de 1 en gran parte del territorio, especialmente en el área rural dispersa.

Por tanto, las expectativas consisten en determinar una medición para estas zonas para el año 2018, así como una completa medición para periodos intercensales.

Versión 1

Estimación IPM con fuente censal

INFORMACIÓN PARA TODOS

Reducción Dimensionalidad

Con el fin de reducir el número de covariables que entran al modelo, se realizaron componentes principales para los cuales se utilizaron las siguientes variables*:

- P_P18_2_PISO: Proporción de viviendas con material del piso en Baldosa, vinilo, tableta, ladrillo, lamina
- P51_09_EDUC: Proporción de personas con nivel educativo Secundaria: Sexto
- P51_04_EDUC: Proporción de personas con nivel educativo Primaria: Primero
- TPH: Tamaño promedio de Hogar
- P_P49_1_ALFAB: Proporción de personas que saben leer y escribir
- P51_13_EDUC: Proporción de personas con nivel educativo Primaria Secundaria Décimo
- P_P18_4_PISO: Proporción de viviendas con material del piso en Cemento y gravilla
- P_P34_2_EDAD: Proporción de personas de 10 -19 años de edad
- P51_02_EDUC: Proporción de personas con nivel educativo Jardín
- P51_07_EDUC: Proporción de personas con nivel educativo Primaria: Cuarto
- P_P19_ALC_1: Proporción de viviendas con Alcantarillado
- P51_10_EDUC': Proporción de personas con nivel educativo Secundaria: Séptimo
- P_P18_6_PISO: Proporción de viviendas con material del piso en Tierra, arena y barro

Componente Principal	Varianza Explicada (%)
1	74,3
2	17,7
3	5,5
4	1,2
5	0,7

Todas la variables se encuentran a nivel de manzana y sección rural del marco geoestadístico nacional (MGN). Se tomaron los primeros 5 componentes principales que explican 99,4% de la varianza de dichas variables.

* Variables con importancia superior al 1% ajustando los modelos con todos los indicadores censales a nivel de manzana y sector rural

Ajuste de los modelos (Estimación IPM)

De esta forma, se modeló la transformación logit del IPM como variable dependiente y como variables independientes se tomaron los primeros cinco componentes principales, la regionalización de los departamentos 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 así 303 497 registros para el conjunto de entrenamiento y 75 810 para el de prueba.

Se ajustaron dos modelos, Gradient Boosted Tree Regression (GBTR) y Random Forest (RF), para los cuales se tienen los resultados presentados en el cuadro.

Medida de Rendimiento	GBTR	RF
Coeficiente de determinación (R^2)	0,6789	0,6621
Raíz del error cuadrático medio ($RMSE$)	0,7818	0,8095

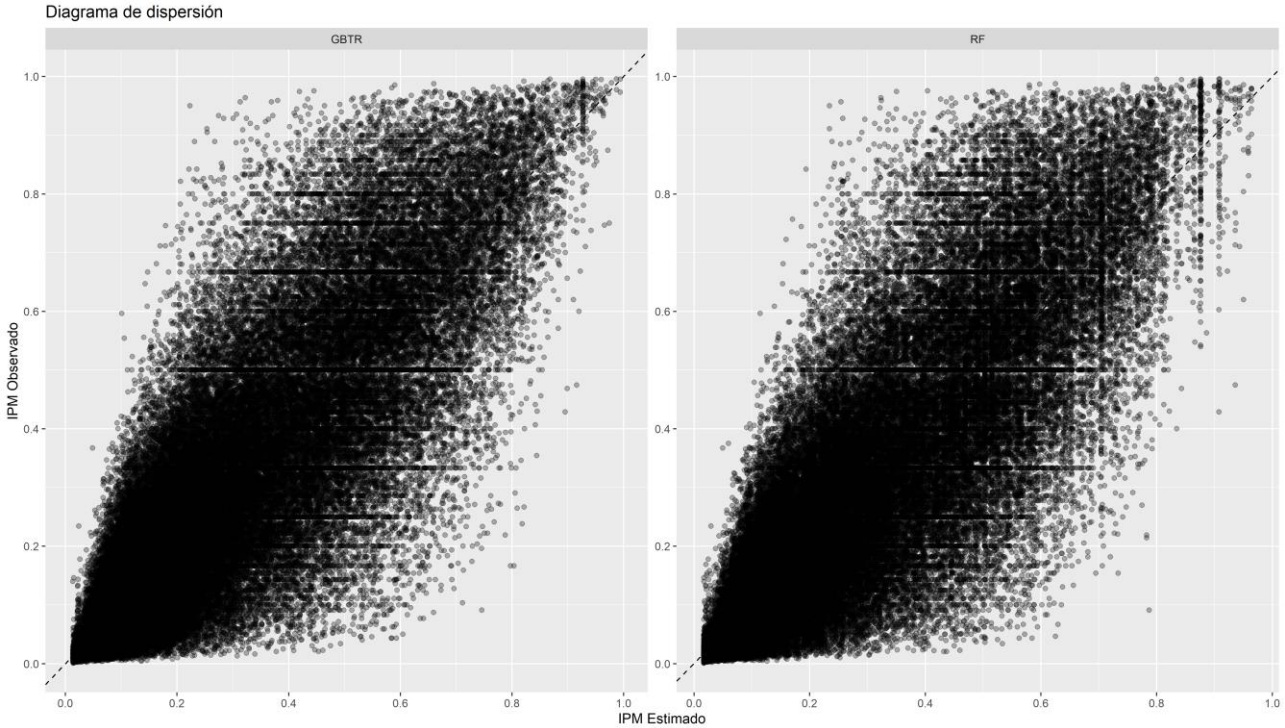
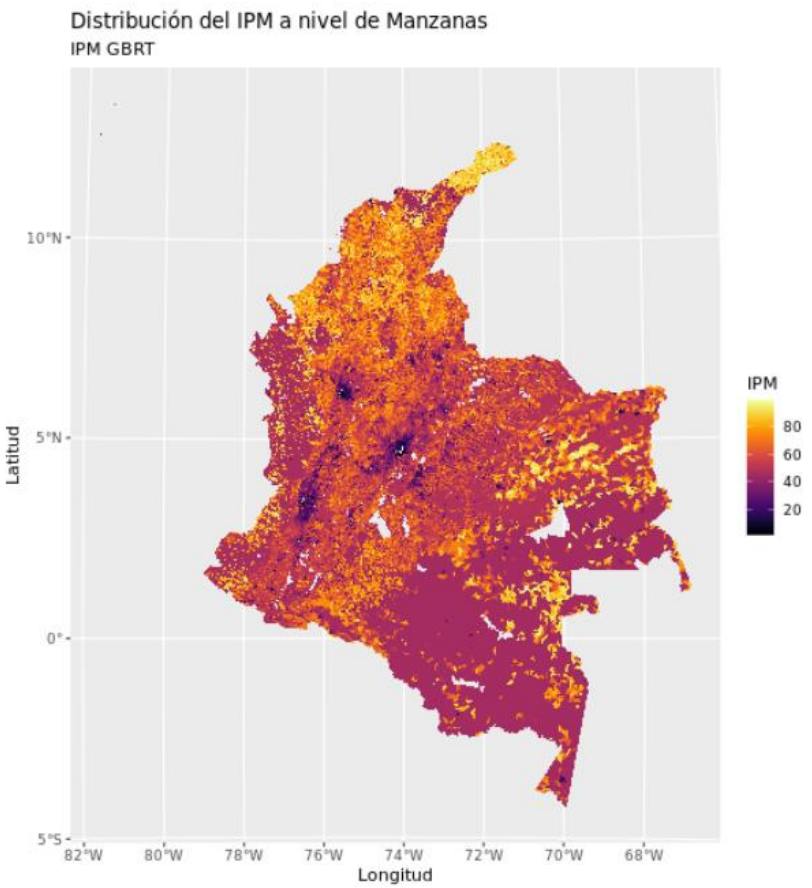
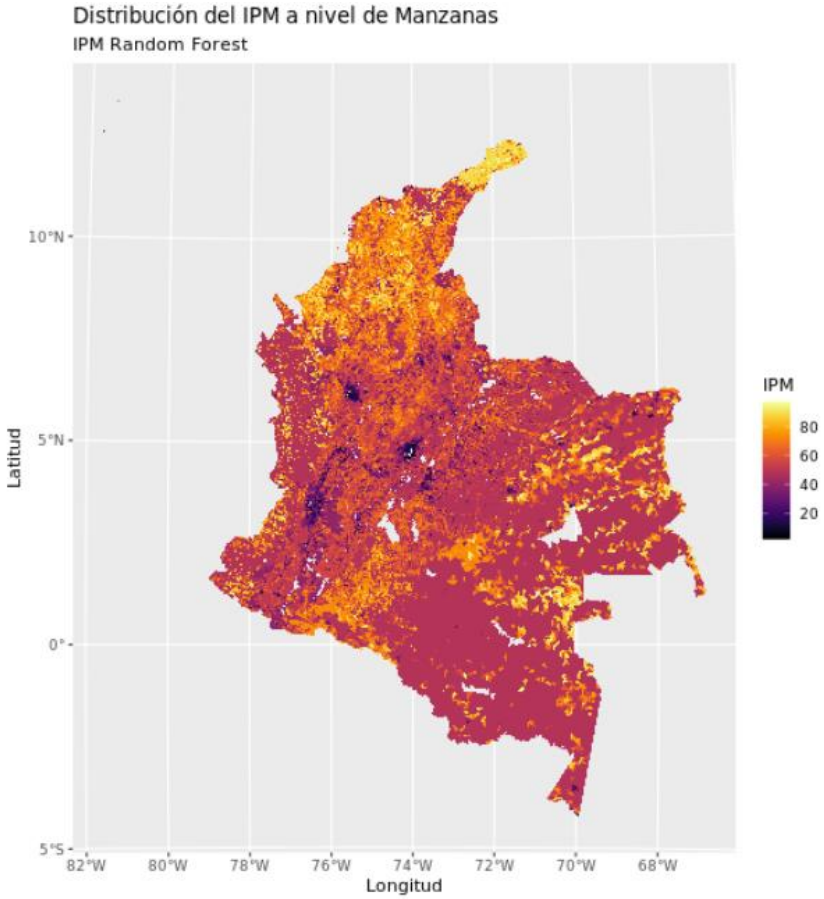
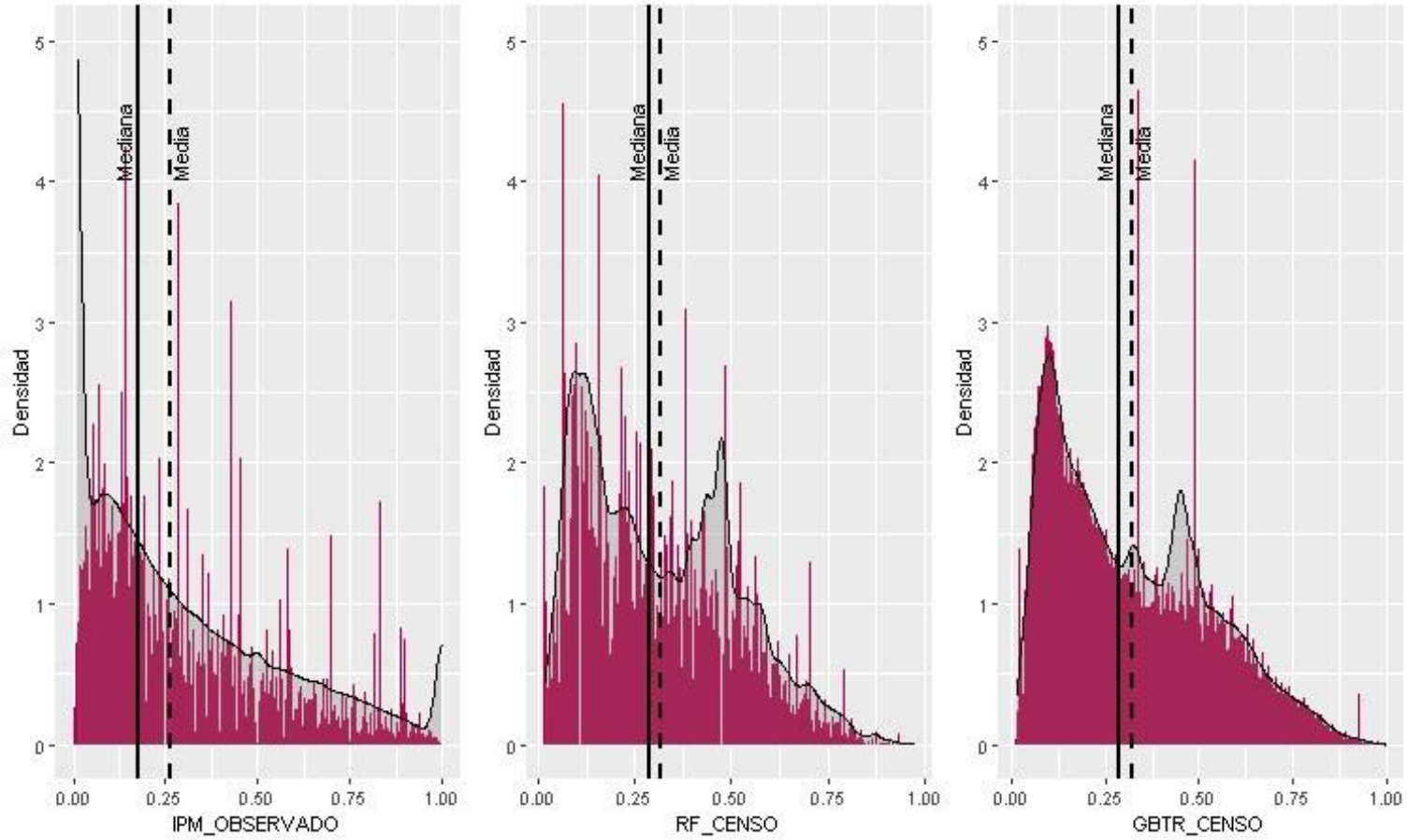


Figura 2: Diagrama de dispersión del IPM observado y estimado para los dos modelos ajustados


Distribución del IPM a nivel de Manzanas
IPM GBRT
10°N
5°N
0°
5°S
Latitud
82°W
80°W
78°W
76°W
74°W
72°W
70°W
68°W
Longitud
IPM
80
60
40
20


Distribución del IPM a nivel de Manzanas
IPM Random Forest
10°N
5°N
0°
5°S
Latitud
82°W
80°W
78°W
76°W
74°W
72°W
70°W
68°W
Longitud
IPM
80
60
40
20



La distribución de las predicciones muestra un comportamiento apuntado para valores ceercanos a 0.5, lo cual es un indicio de la poca información que aportan los indicadores censales en las áreas rurales.

Versión 2

Estimación IPM con fuente censal

Reducción Dimensionalidad

Con el fin de reducir el número de covariables que entran al modelo, se realizaron componentes principales (excluyendo las observaciones del IPM en 0) para los cuales se utilizaron las mismas variables de la versión 1.

Se tomaron los primeros 5 componentes principales que explican 76,12% de la varianza de dichas variables.

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

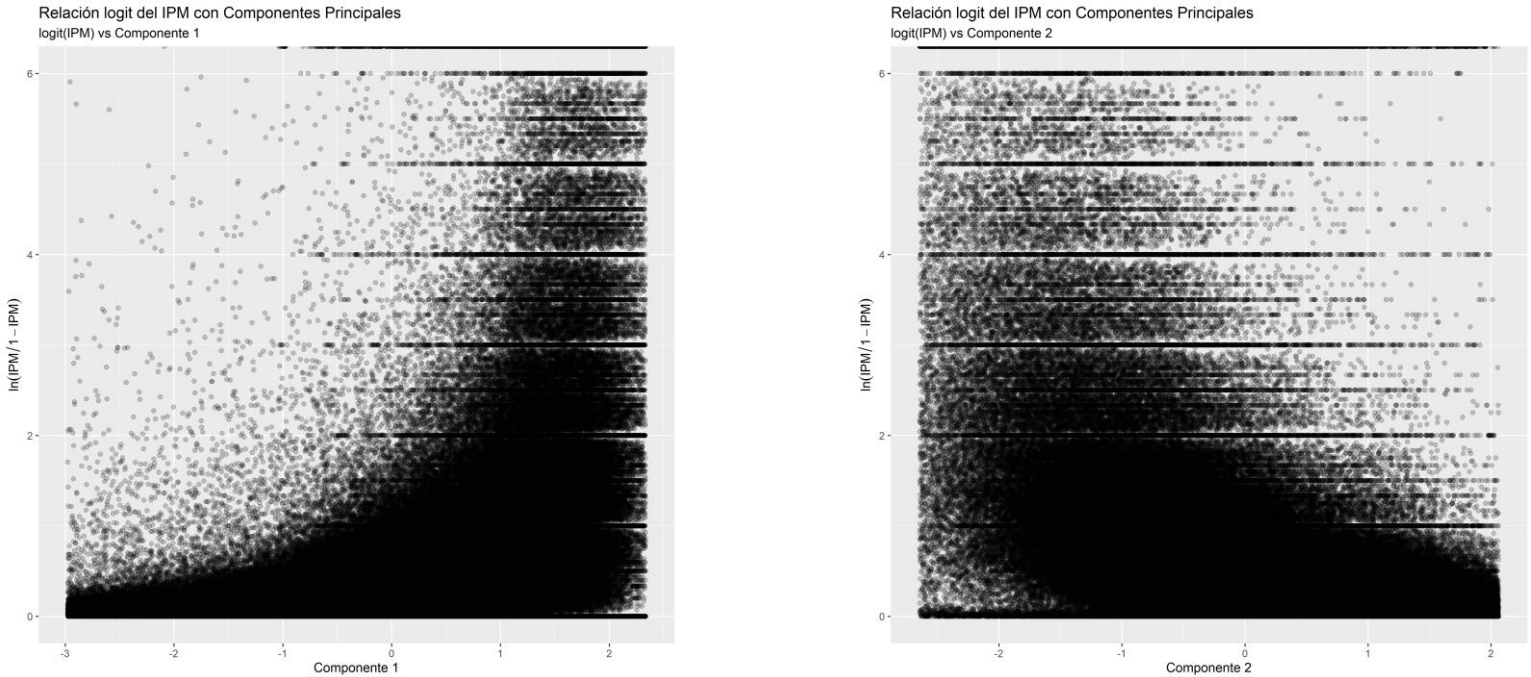


Figura 1: Diagramas de dispersión de la transformación logit del IPM con los dos primeros componentes principales

Ajuste de los modelos (Estimación IPM)

De esta forma, se modeló la transformación logit del IPM como variable dependiente y como variables independientes se tomaron los primeros cinco componentes principales, la regionalización de los departamentos 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 0 y donde no existía población efectivamente censada, obteniendo así 316 377 registros para el conjunto de entrenamiento y 78 794 para el de prueba donde el conjunto total de datos tiene 513 421 registros, es decir que se excluyeron 118 250 manzanas.

Se ajustaron dos modelos, Gradient Boosted Tree Regression (GBTR) y Random Forest (RF), para los cuales se tienen los resultados presentados en el cuadro.

Medida de Rendimiento	GBTR	RF
Coeficiente de determinación (R^2)	0,6537	0,6281
Raíz del error cuadrático medio ($RMSE$)	1,1898	1,233

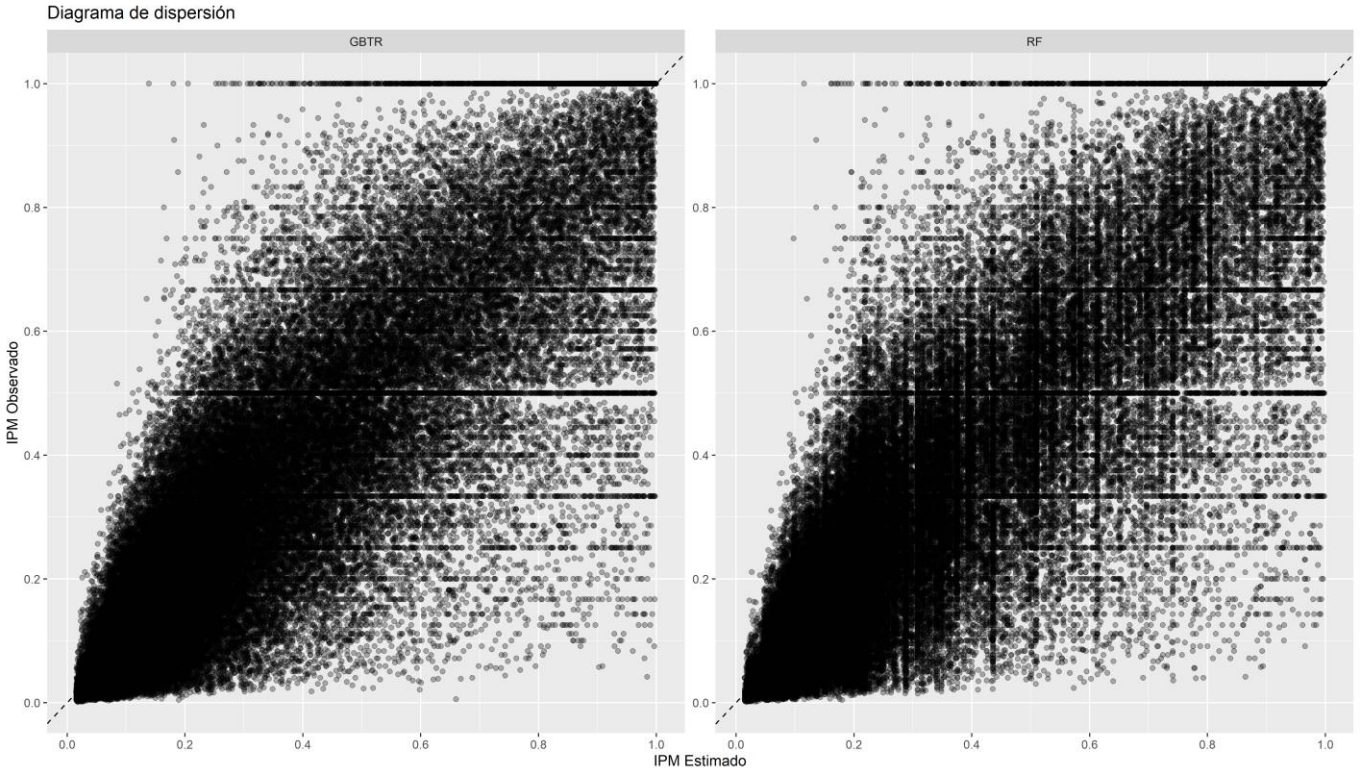
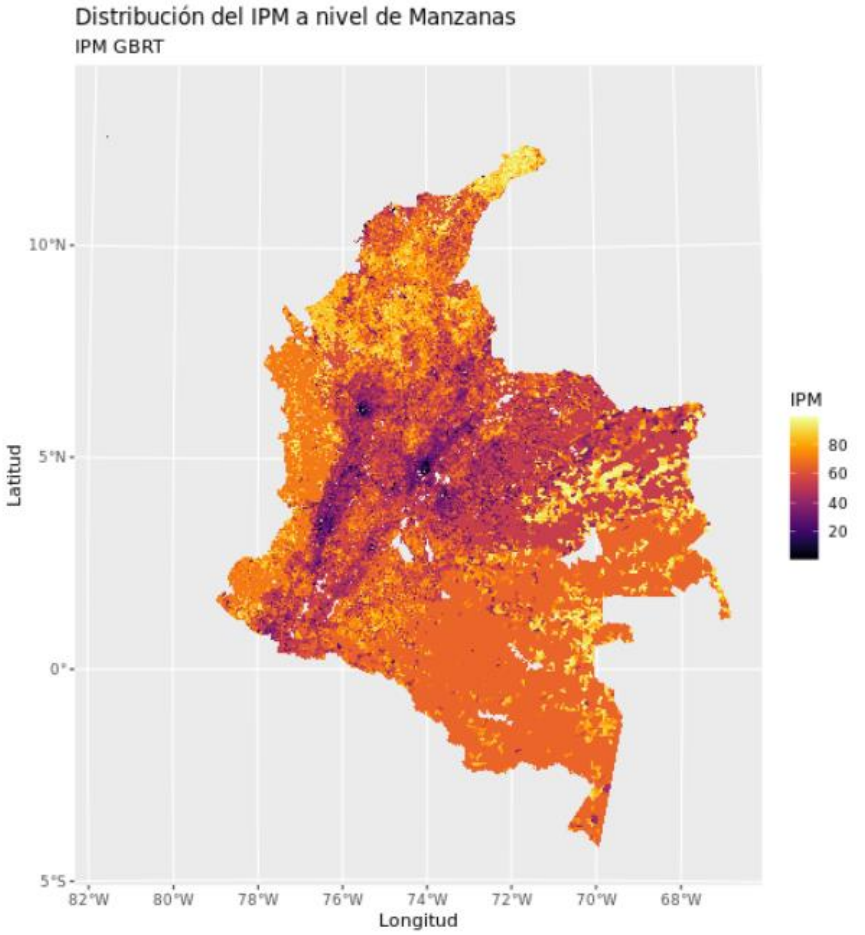
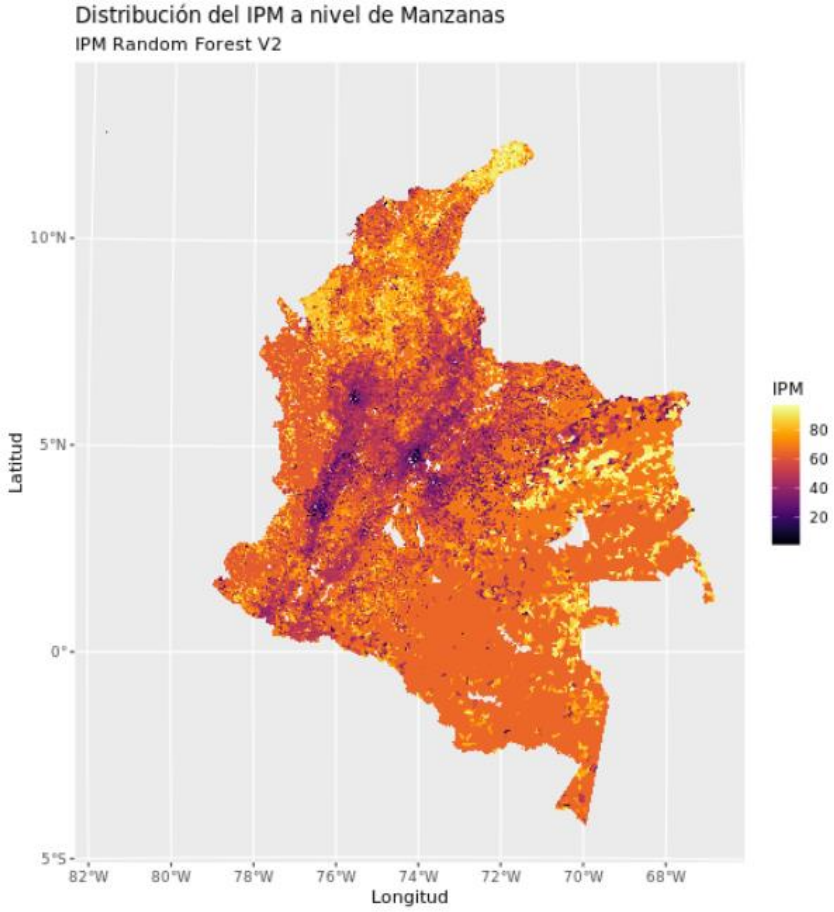


Figura 3: Diagrama de dispersión de valores predichos y observados en el conjunto de prueba para los dos modelos ajustados.


Distribución del IPM a nivel de Manzanas
IPM GBRT
10°N
5°N
0°
5°S
82°W
80°W
78°W
76°W
74°W
72°W
70°W
68°W
Latitud
Longitud
IPM
80
60
40
20


Distribución del IPM a nivel de Manzanas
IPM Random Forest V2
10°N
5°N
0°
5°S
82°W
80°W
78°W
76°W
74°W
72°W
70°W
68°W
Latitud
Longitud
IPM
80
60
40
20

Versión 3

Estimación IPM con Imágenes satelitales

Se aplicaron los modelos de redes neuronales convolucionales compartidos por **PARÍS 21**, de donde se obtienen 512 covariados ubicados en **77 979** centroides, extraídos de las imágenes satelitales 2018.

Posteriormente, se aplica ACP para reducir dimensionalidad, se obtienen las 5 primeras componente principales que explican el 99.9% de la variabilidad del conjunto de datos.

Por último, se implementa el método de interpolación determinístico por vecinos naturales, para ubicar las 5 componentes principales de las 77 979 imágenes dentro del las más de 500 000 zonas del MGN.

Componente Principal	Varianza Explicada (%)
1	80.3400
2	0.1955
3	0.0010
4	0.0001
5	0.0000

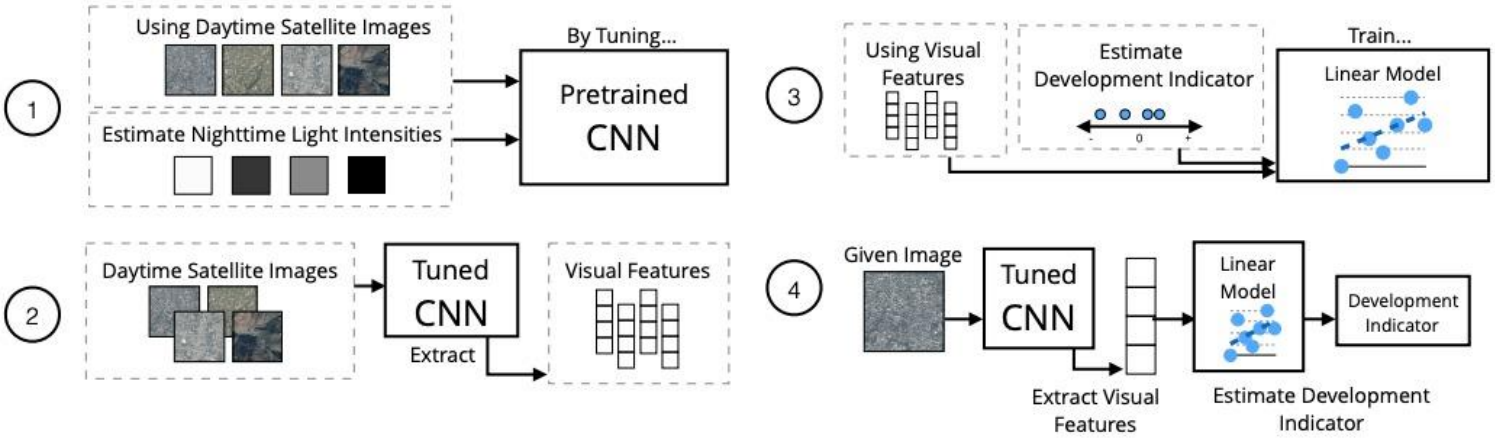


Figure 1: The transfer learning process used to predict development indicators from daytime satellite images. (1) A pre-trained CNN is tuned to predict nighttime light intensity; (2) High-level visual features are extracted from the top layers of the tuned CNN; (3) A linear model is trained to estimate a development indicator using ridge regression. (4) Given an arbitrary image, we can predict the development indicator by feeding the extracted visual features into the trained linear model. We tune a separate CNN for each country, and train a linear model to predict each development indicator for each country.

Ajuste de los modelos (Estimación IPM)

Se toma una configuración del conjunto de entrenamiento y prueba de 80-20. Se incluyen las observaciones de 0 y 1

Método 1

Estimación features de imágenes satelitáles

Medidas Rendimiento / Estimación	R^2		$RMSE$	
	GBTR	RF	GBTR	RF
P1	0.042	0.034	466.167	467.518
P2	0.159	0.157	213.006	213.252
P3	0.173	0.141	111.786	113.929
P4	0.024	-0.063	57.834	60.359
P5	-0.206	-0.026	47.467	43.793

Estimación Logit-IPM

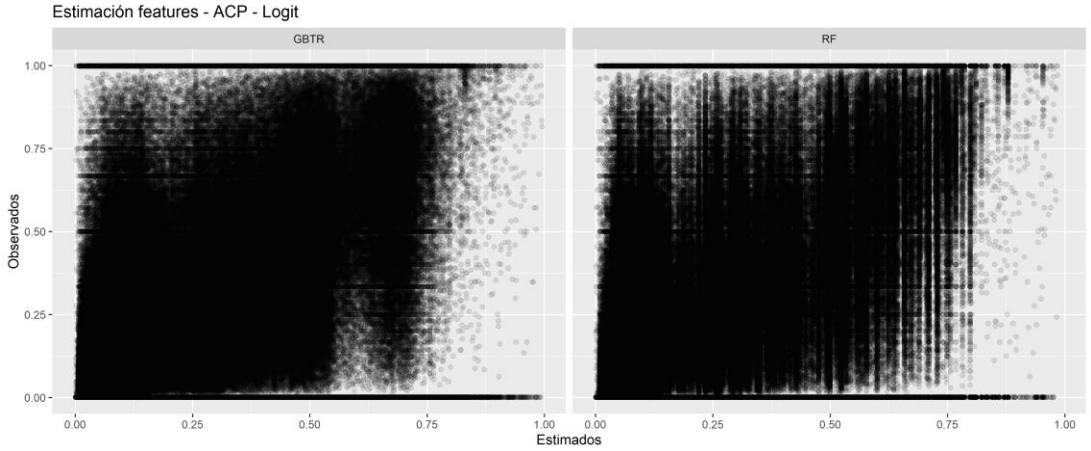
R^2		$RMSE$	
GBTR	RF	GBTR	RF
0.002	-0.007	3.332	3.347

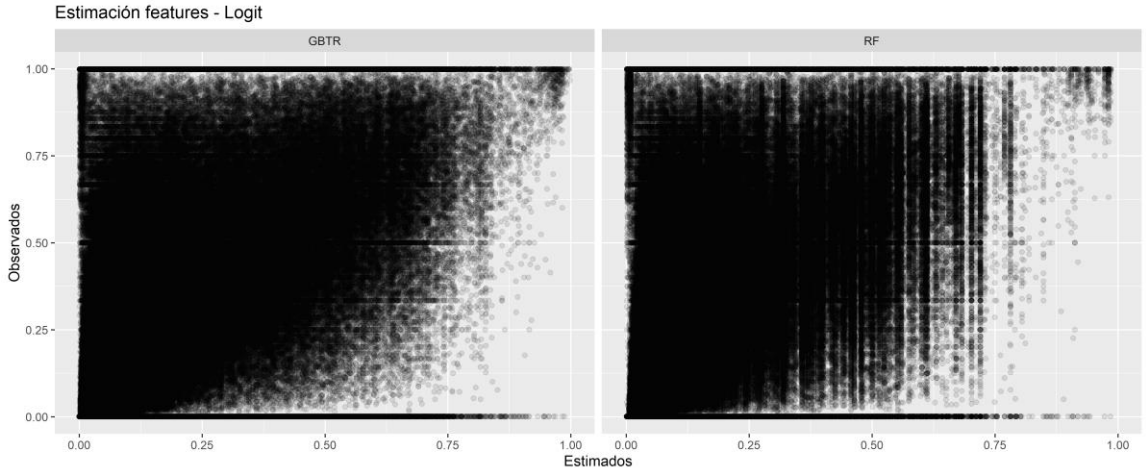
Método 2

Estimación features de imágenes satelitáles

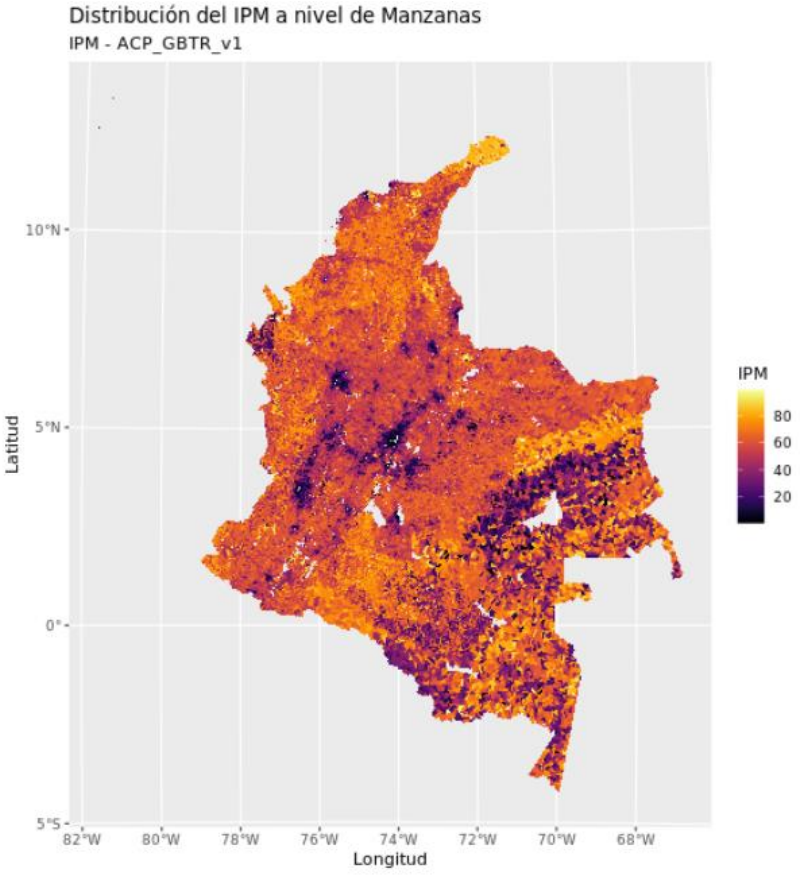
Estimación Logit-IPM

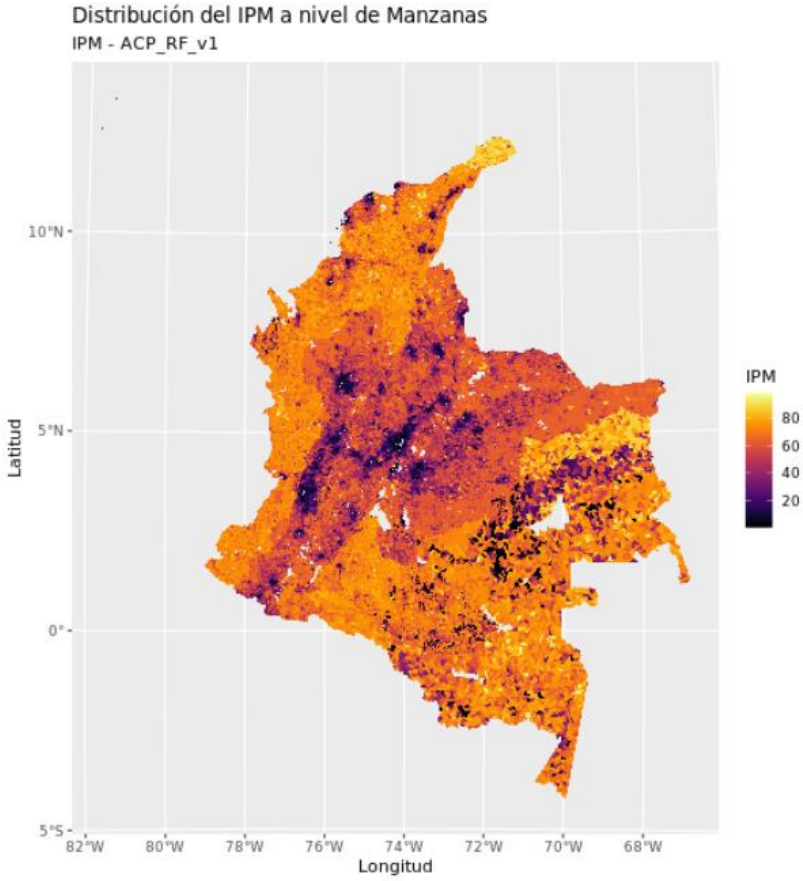
R^2		$RMSE$	
GBTR	RF	GBTR	RF
0.203	0.171	2.984	3.042


Estimación features - ACP - Logit
GBTR
RF
Observados
Estimados
0.00
0.25
0.50
0.75
1.00

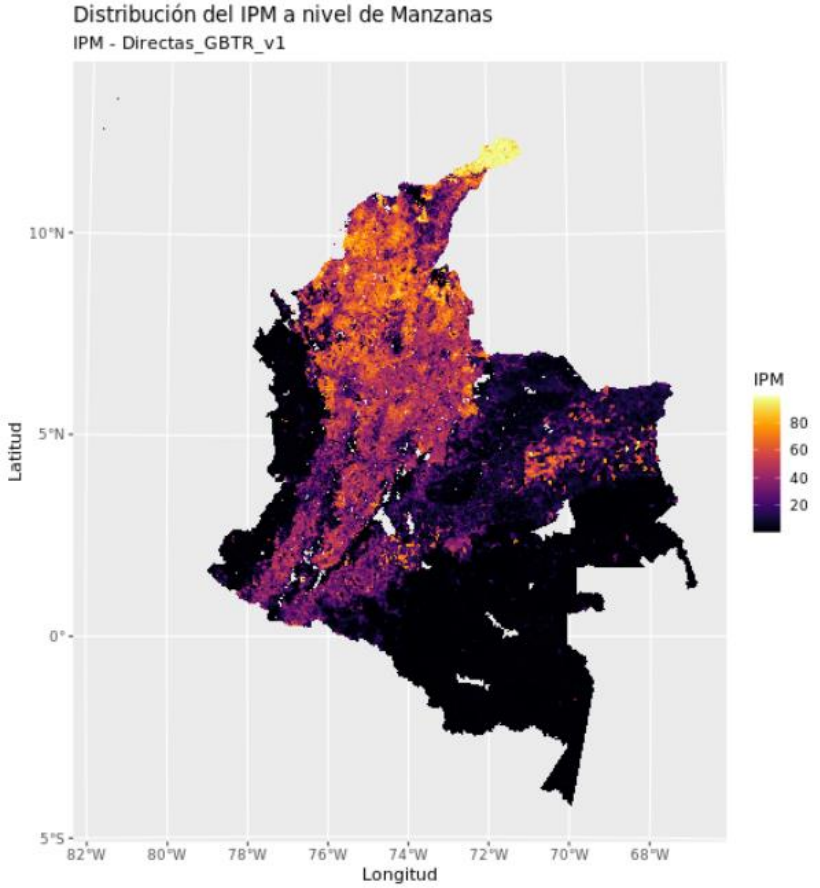

Estimación features - Logit
GBTR
RF
Observados
Estimados
0.00
0.25
0.50
0.75
1.00

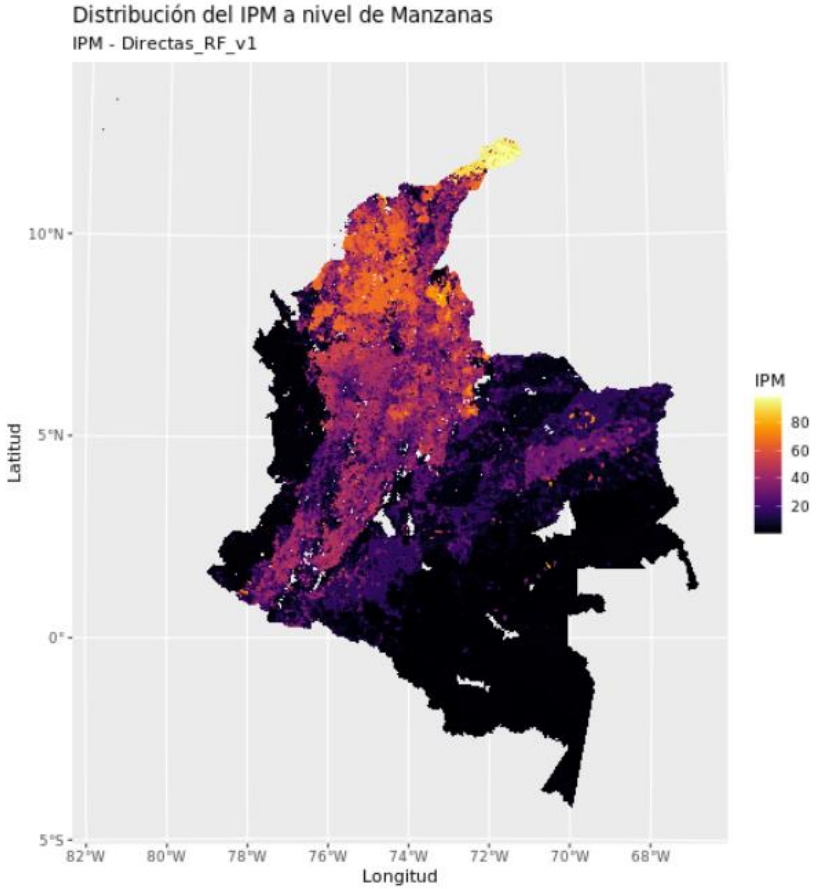
Predicciones Método 1





Predicciones Método 2

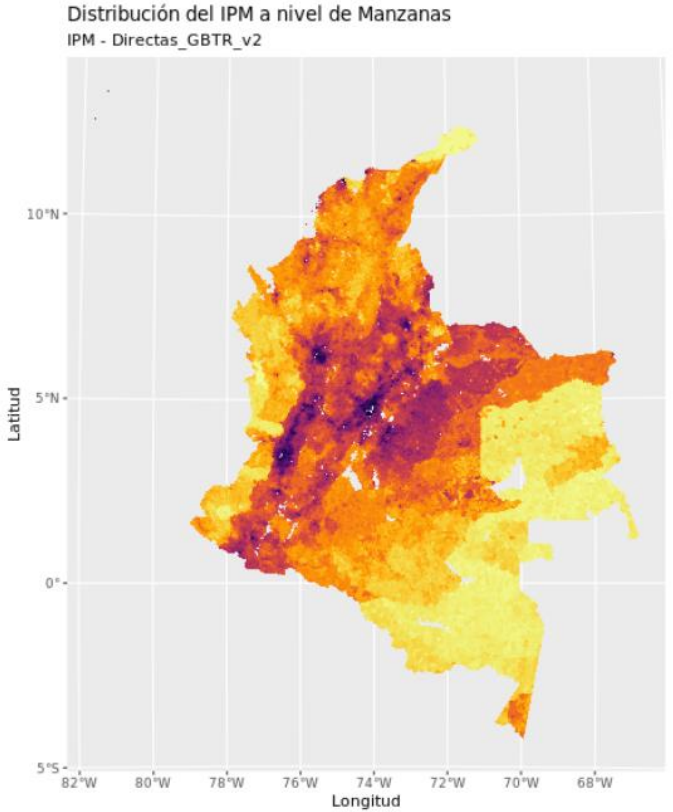


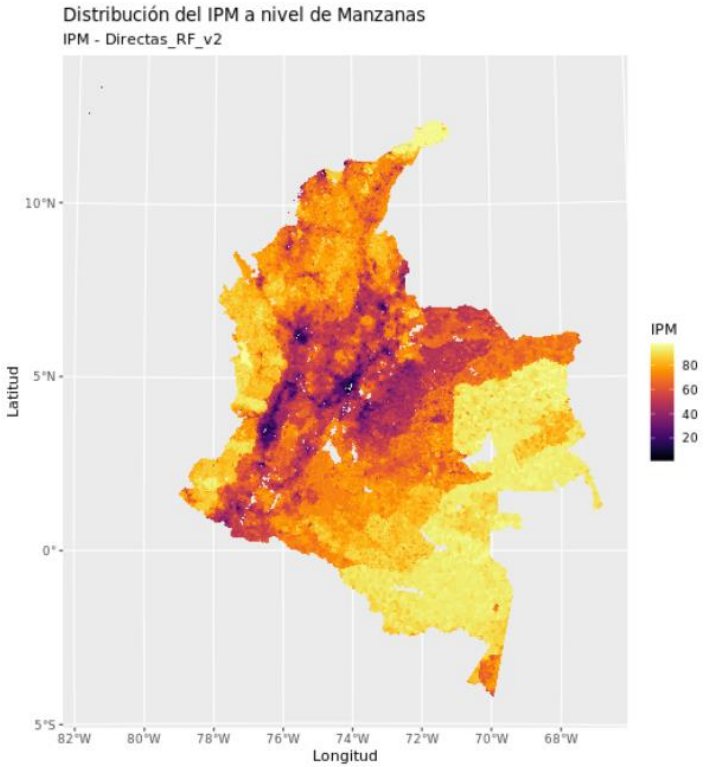


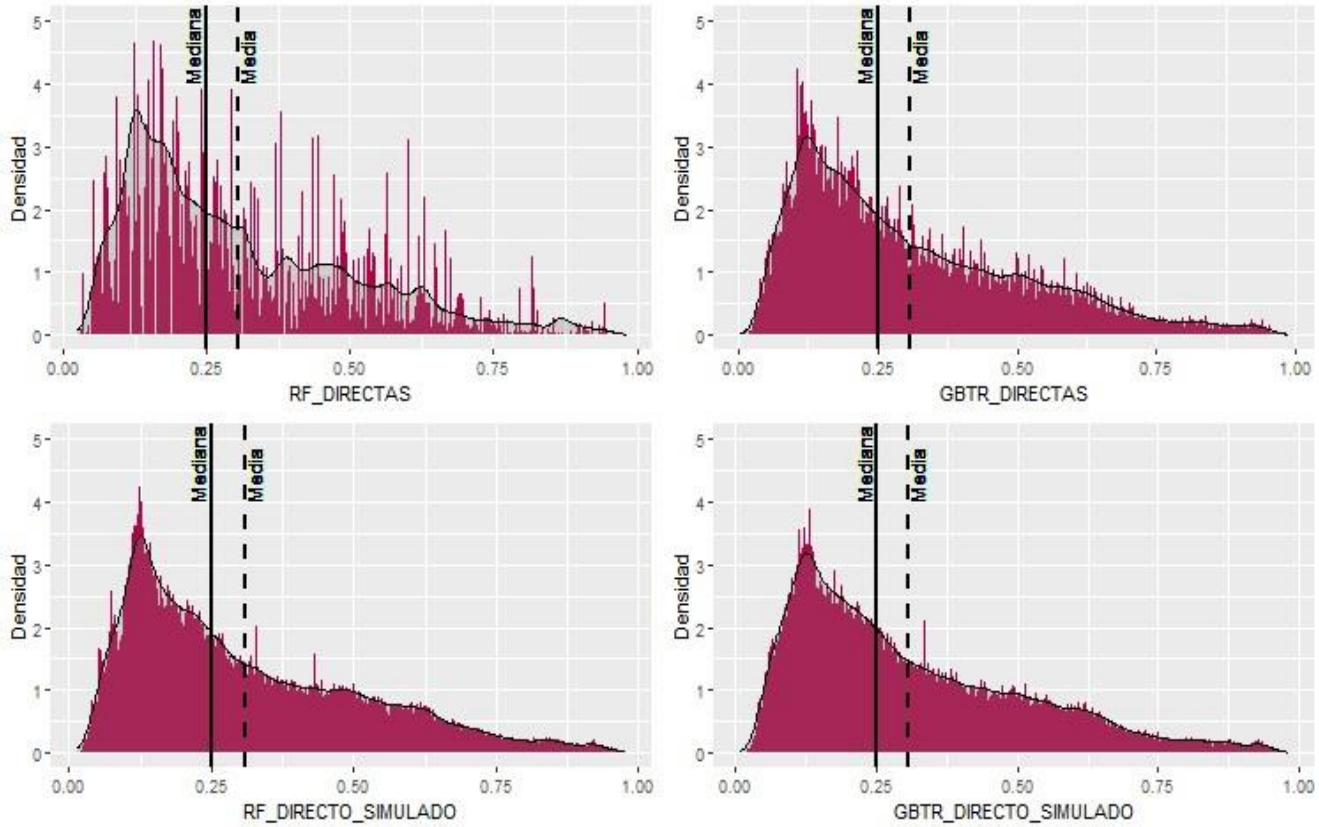
Método 3

Se toma una configuración del conjunto de entrenamiento y prueba de 80-20 en un ciclo de Montecarlo de 100 iteraciones. Se excluyen las observaciones de 0 y 1.

Medida de Rendimiento	GBTR	RF
R^2	0,5289	0,5757
RSME	0,9555	0,9067

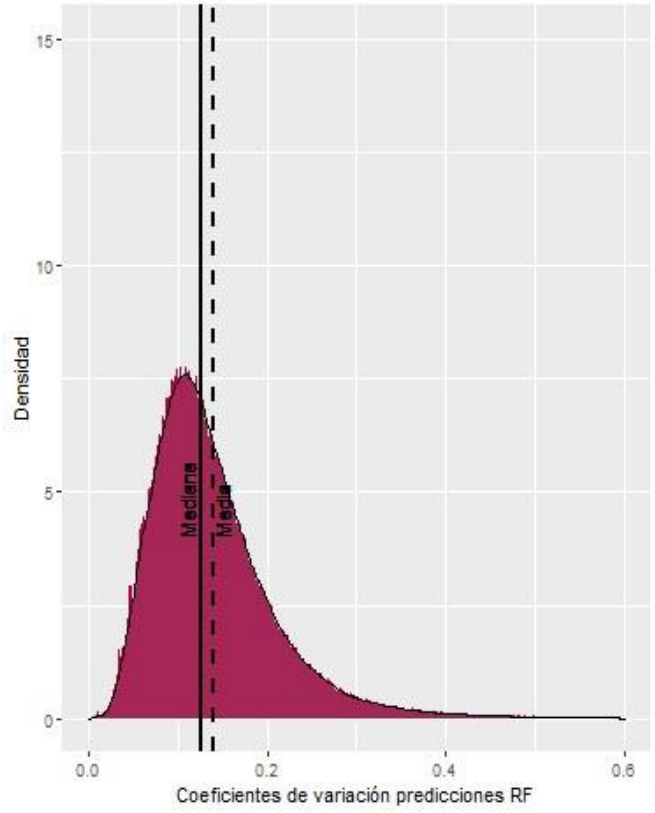


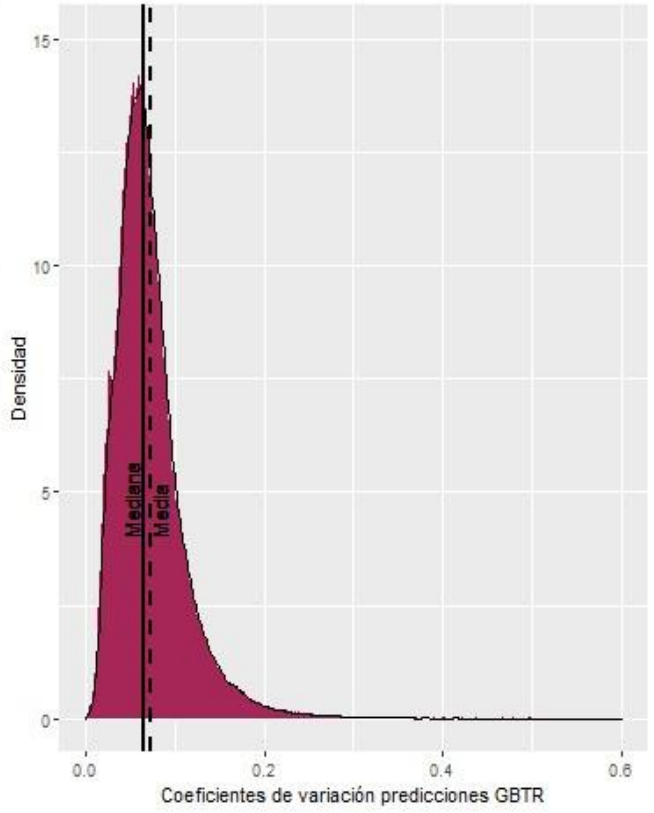


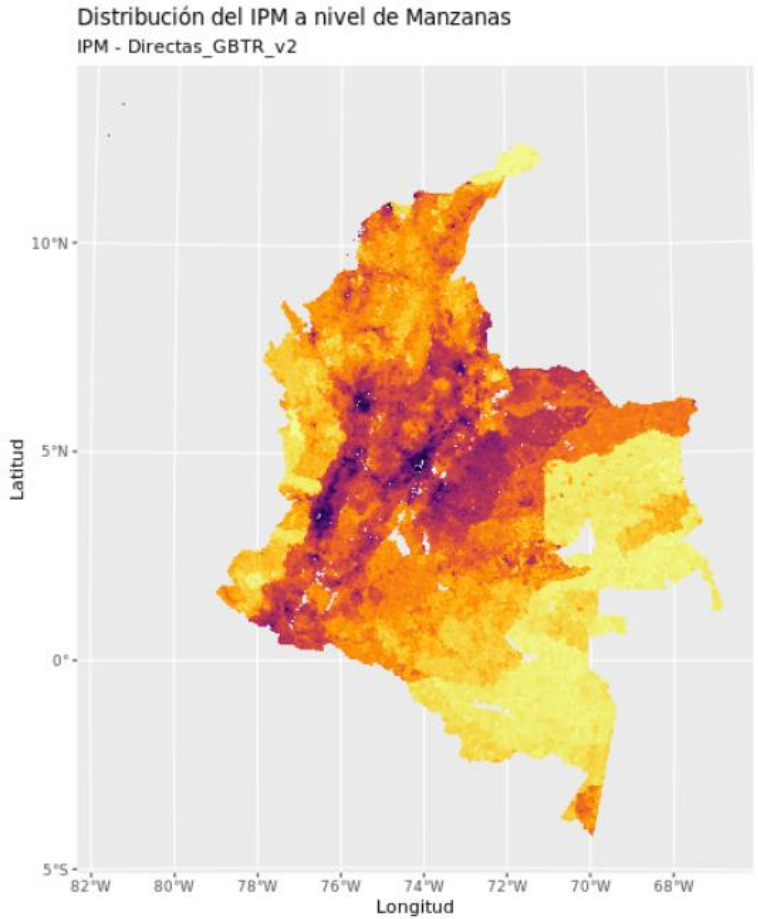


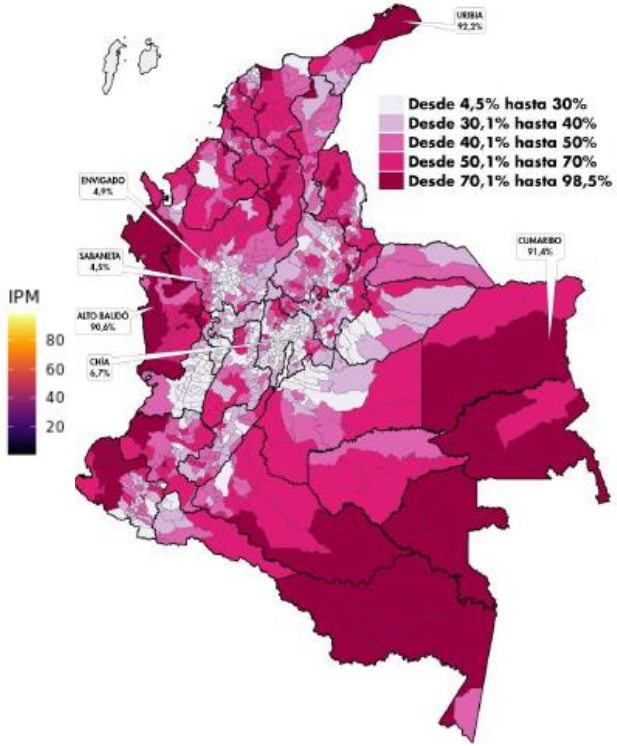
Aunque las métricas de desempeño del modelo GBTR son inferiores al modelo de RF, la predicciones del GBTR tienen menor volatilidad y describen un comportamiento con un número de menor de máximos locales.

Por tanto es considerada la mejor estimación.


15
10
5
0
Densidad
Mediana
Media
0.0
0.2
0.4
0.6
Coeficientes de variación predicciones RF


15
10
5
0
Densidad
Mediana
Media
0.0
0.2
0.4
0.6
Coeficientes de variación predicciones GBTR





Al comparar las estimaciones censales a nivel municipal con los resultados del modelo, se puede observar un comportamiento similar.

Es decir que los covariados extraídos de las imágenes satelitales describen el IPM del forma similar, ya que consiguen explicar el 52.9% de la variabilidad del indicador.

Conclusiones

- 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 presente 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, podría reflejar el mismo comportamiento.

- 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.

- En este ejercicio de pobreza, todas las fuentes de información con excepción del CNPV son públicas, 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.