Departamento Nacional de Planeación

www.dnp.gov.co

Estimación de ingresos con **machine learning para el territorio rural** colombiano

Dirección de Desarrollo Digital
Dirección de Desarrollo Rural Sostenible

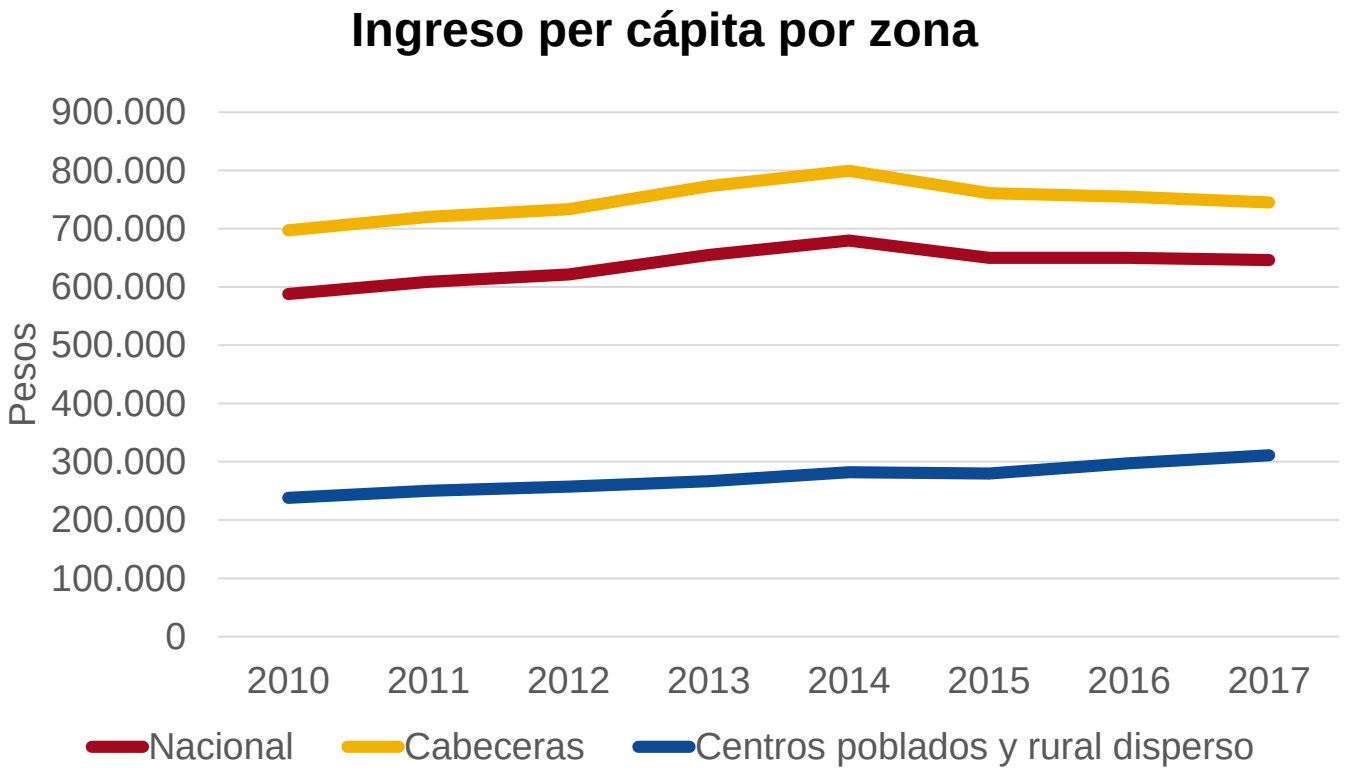
Mayo, 2018
dnp.gov.co

DNP Departamento Nacional de Planeación

GOBIERNO DE COLOMBIA

Resumen del proyecto

Objetivo: Conocer cómo los entornos territoriales afectan la generación de ingresos de los pequeños productores agropecuarios



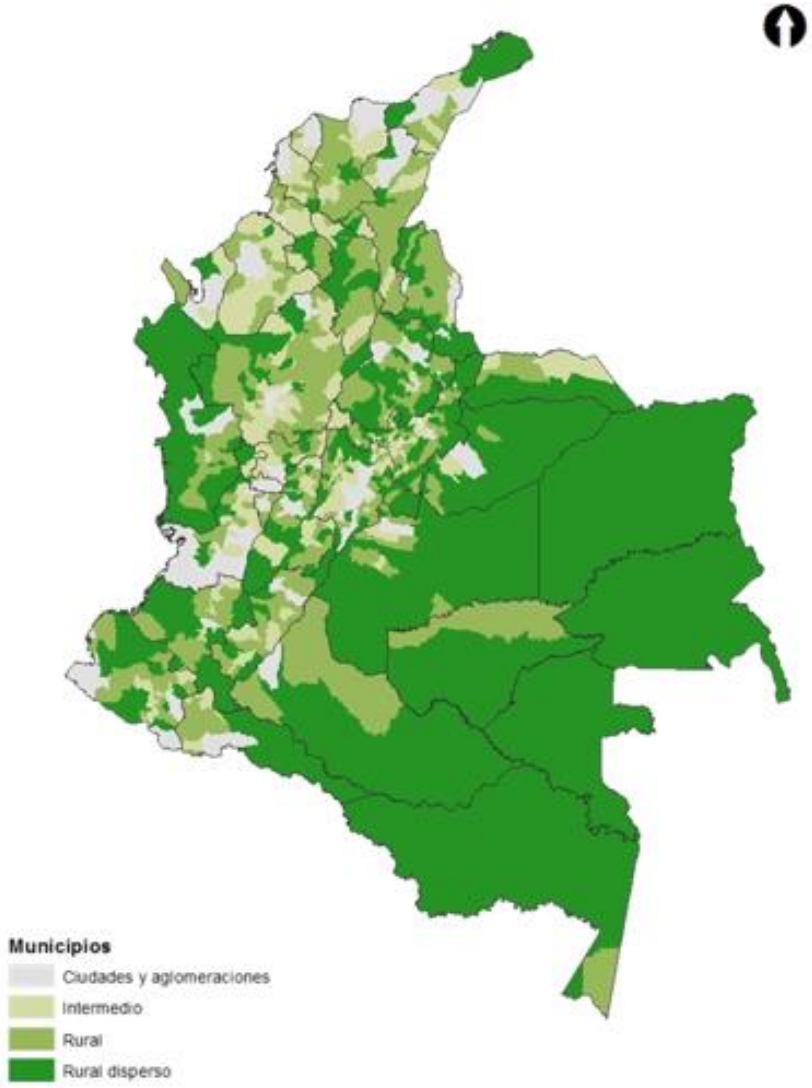
La Gran Encuesta Integrada de hogares permite calcular el **ingreso promedio nacional (cabecera, y centros poblados y rural disperso), y departamental.**

Para la política de generación de Ingresos Rurales se requiere identificar la capacidad de las zonas rurales

Con la aproximación de *Machine Learning* se busca disponer de un **primer mapa del territorio nacional** que visualice la **distribución de ingresos** del territorio rural

Fuente: DANE (2017)

Categorización municipal



Categorías definidas:

Ciudades y aglomeraciones: Cabeceras municipales de gran tamaño – **Principales centros urbanos del país**

Intermedio: Municipios con cabeceras de tamaño medio o alta densidad poblacional – **Nodos subregionales**

Rural: Municipios con cabeceras de tamaño pequeño o densidades poblacionales intermedias

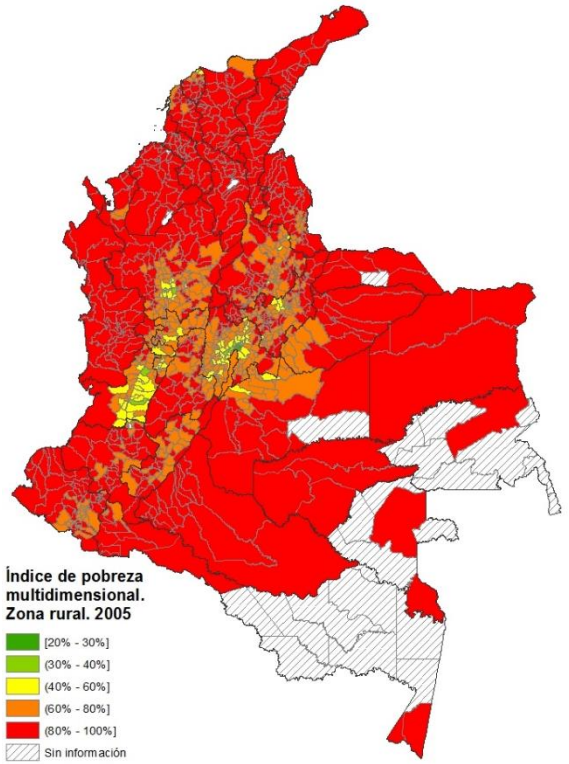
Rural disperso: Municipios con cabeceras de tamaño pequeño y bajas densidades poblacionales

Fuente: DDRS-DNP a partir de DNP (2015)

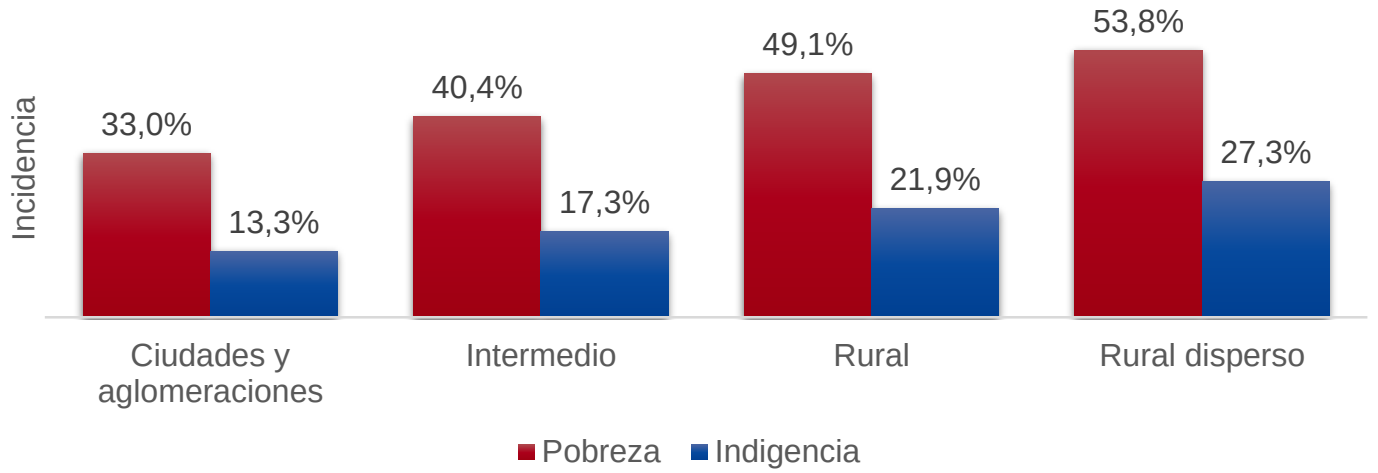
Medición de la pobreza según el IPM

Las zonas rurales que se encuentran distantes a los principales centros urbanos presentan mayores niveles de pobreza

Índice de Pobreza Multidimensional. Zona rural



Incidencia de pobreza monetaria y pobreza extrema monetaria



Conociendo la distribución de ingresos se puede contribuir en el **diseño de política pública que involucre los factores territoriales en la capacidad de generación de ingresos** de la población rural

Fuente: DDRS-DNP a partir de CNPV (2005) y GEIH

No existe mucha información rural en el país

El III CNA cuenta con datos georreferenciados e información social y productiva de los habitantes rurales dispersos. La GEIH es representativa para la zona rural (nacional) tiene información económica y social.

III CNA (2014)
5,1 millones de personas – 45 variables
Información georreferencia de cada vivienda rural ocupada

GEIH (2010 – 2017)
654 mil personas en zonas rurales– 216 variables
Información económica y de ingreso mensuales

12 variables en común

Viviendas Hogares Personas

Utilizando la información en común se puede **encontrar patrones** que permitan **estimar los ingresos** de los hogares rurales dispersos

Identificación de variables explicativas

Utilizando *Stepwise* se busca el modelo más simple que explique suficientemente bien la variable ingresos

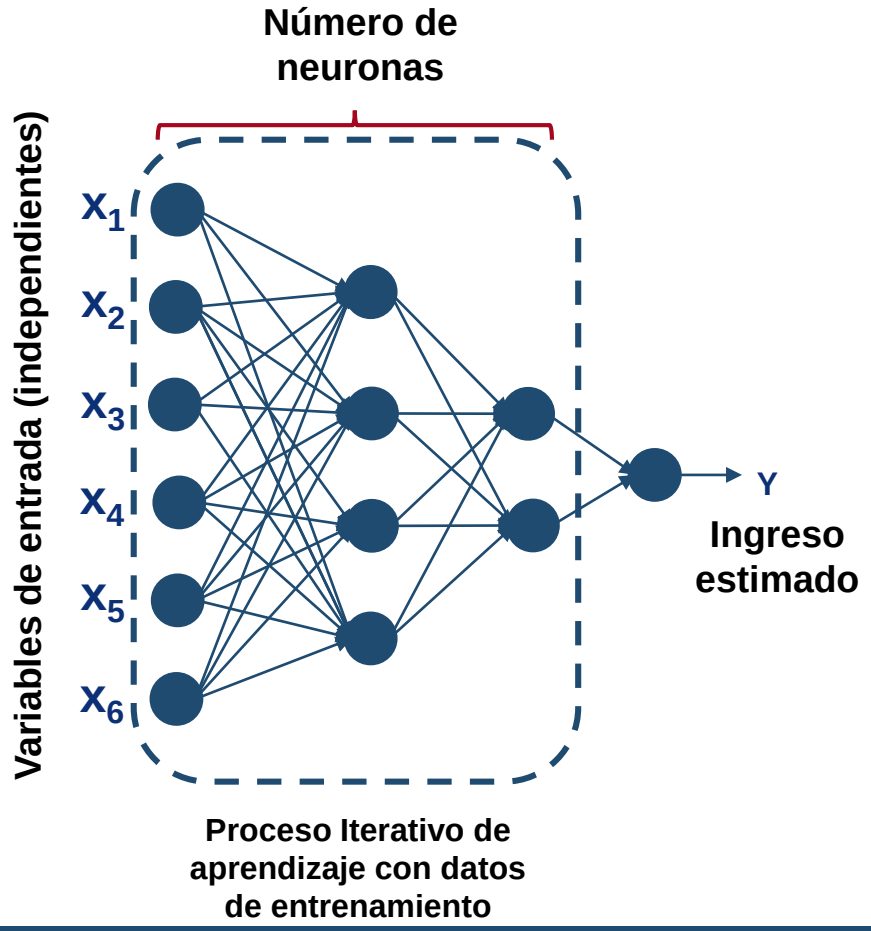
Resultados del modelo *stepwise* MCO

Variables	Ingreso	Variables	Ingreso
Madera	-0.002***	Edad	-0.000***
Adobe	-0.003***	Alfabetismo	0.000***
Bareque	-0.003***	Asistencia Esc.	-0.001***
Madera Burda	-0.001***	Preescolar	0.002***
Guadua	-0.002***	Primaria	0.001***
Caña	-0.002***	Secundaria	0.002***
Zinc	-0.001***	Superior	0.018***
Sin Paredes	-0.004***	Media	0.004***
Cemento	0.001***	No Informar	-0.002***
Madera Burda	0.001***	Contributivo	0.014***
Baldosin	0.008***	Especial	-0.011***
Marmol	0.056***	Subsidiado	-0.007***
Madera Pulida	0.017***	No sabe	0.021***
Alfombra	0.079***	Observations	73,311,708
Energía	0.002***	R-squared	0.208
Alcantarillado	0.001***	Standard errors in parentheses	
Acueducto	0.000***	*** p<0.01, ** p<0.05, * p<0.1	
Hombre	0.001***		

Las 12 variables son significativas para el ingreso

Modelo de redes neuronales

La red neuronal es un gran conjunto de unidades neuronales simples (neuronas artificiales) conectadas entre sí para modelar relaciones complejas entre las variables de entrada y de salida



Conjunto de entrenamiento (70%)

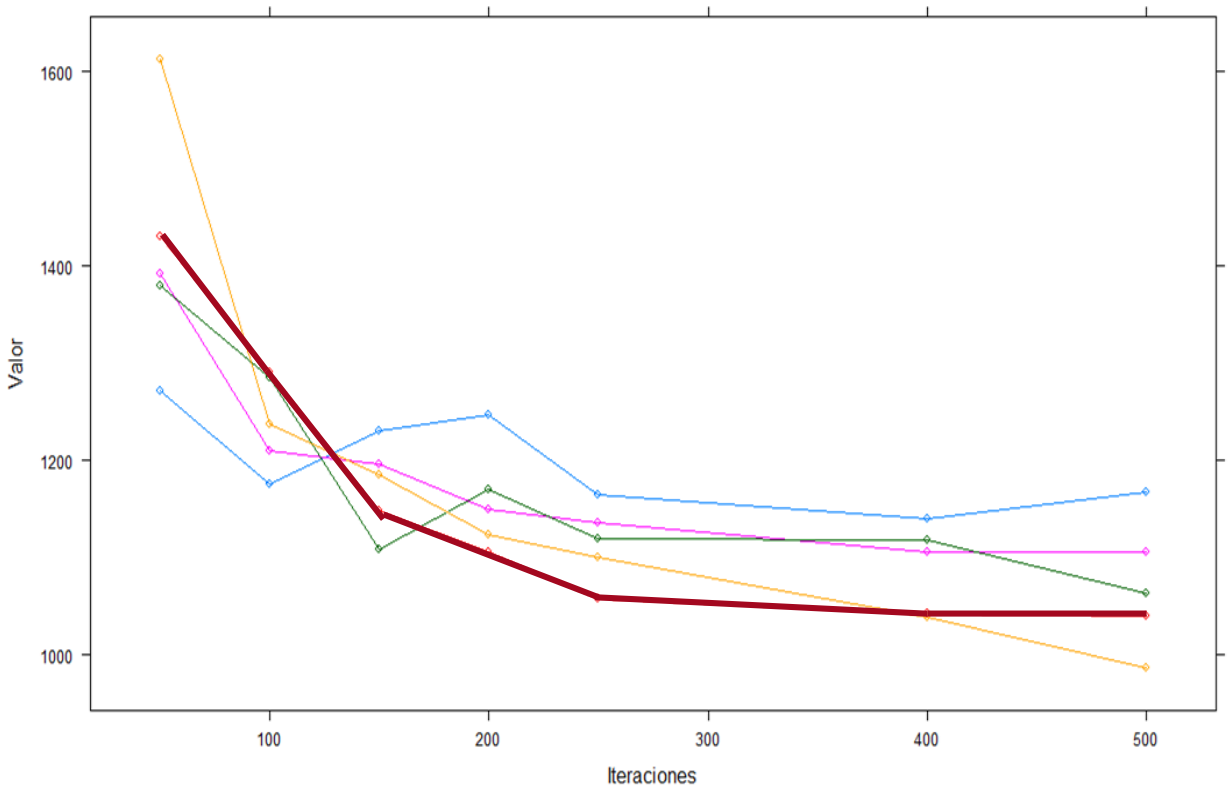
Conjunto de prueba (30%)

Es necesario encontrar un **número de neuronas y de iteraciones óptimo** para conseguir proyecciones adecuadas de la variable dependiente

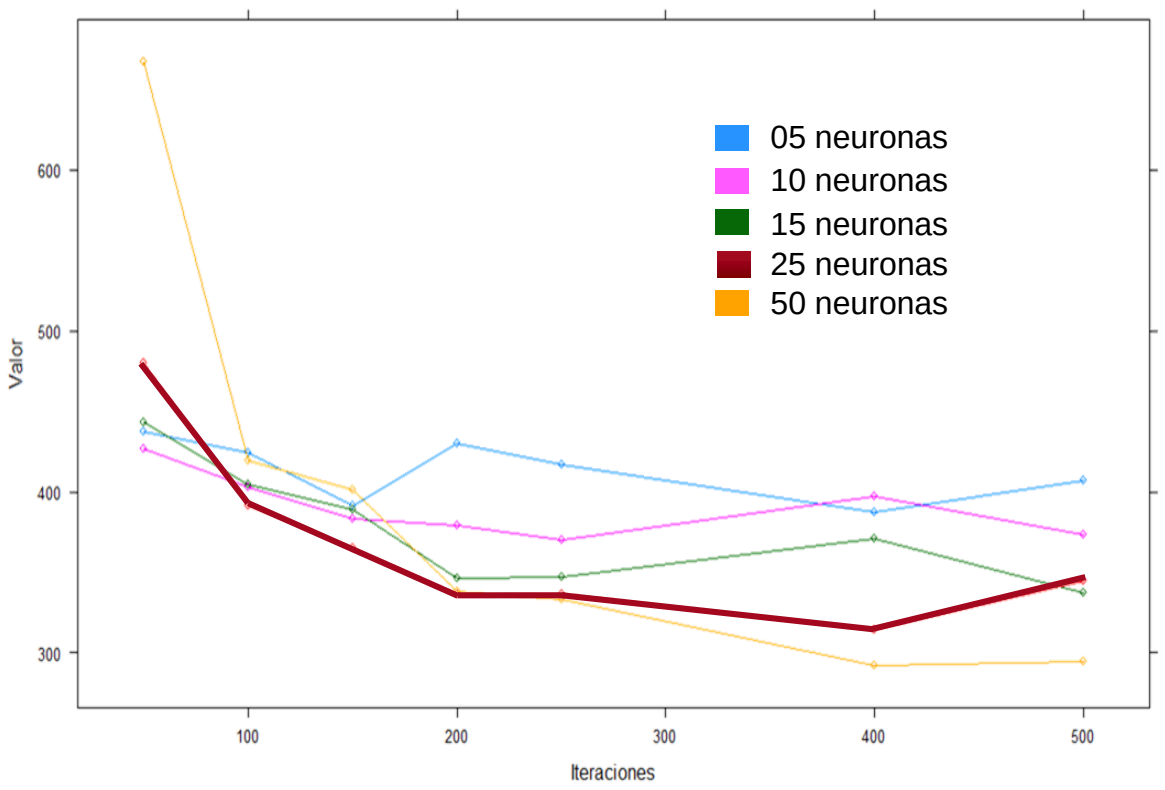
Resultados de los errores en la red neuronal

La combinación escogida para aplicar a la red fueron 25 neuronas y 250 iteraciones sobre el set de entrenamiento

Resultados sobre el conjunto de entrenamiento



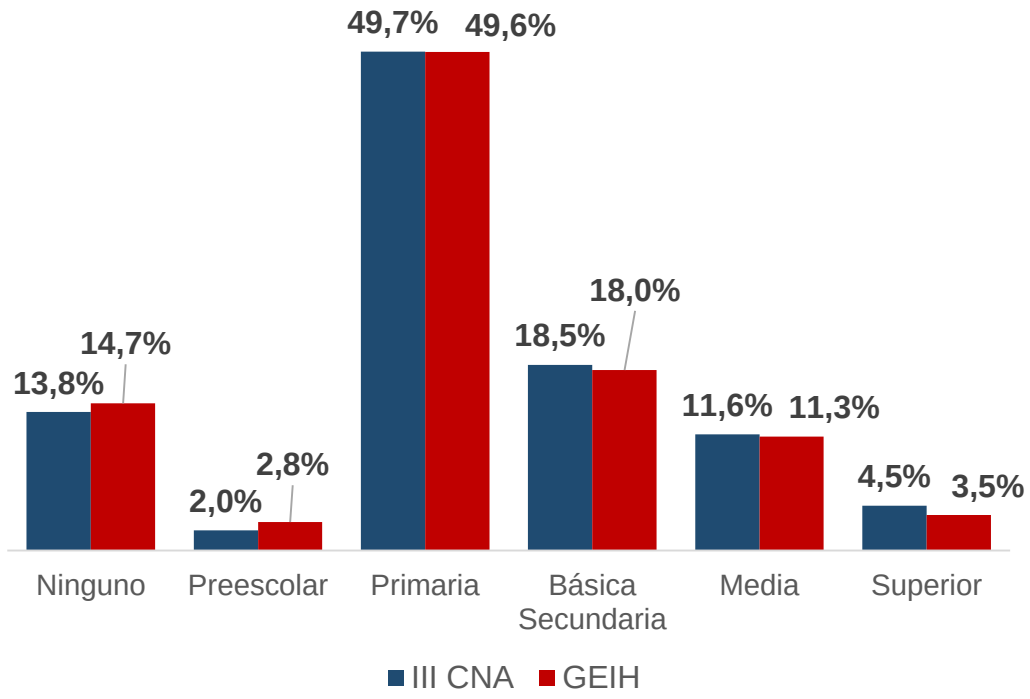
Resultados sobre el conjunto de prueba



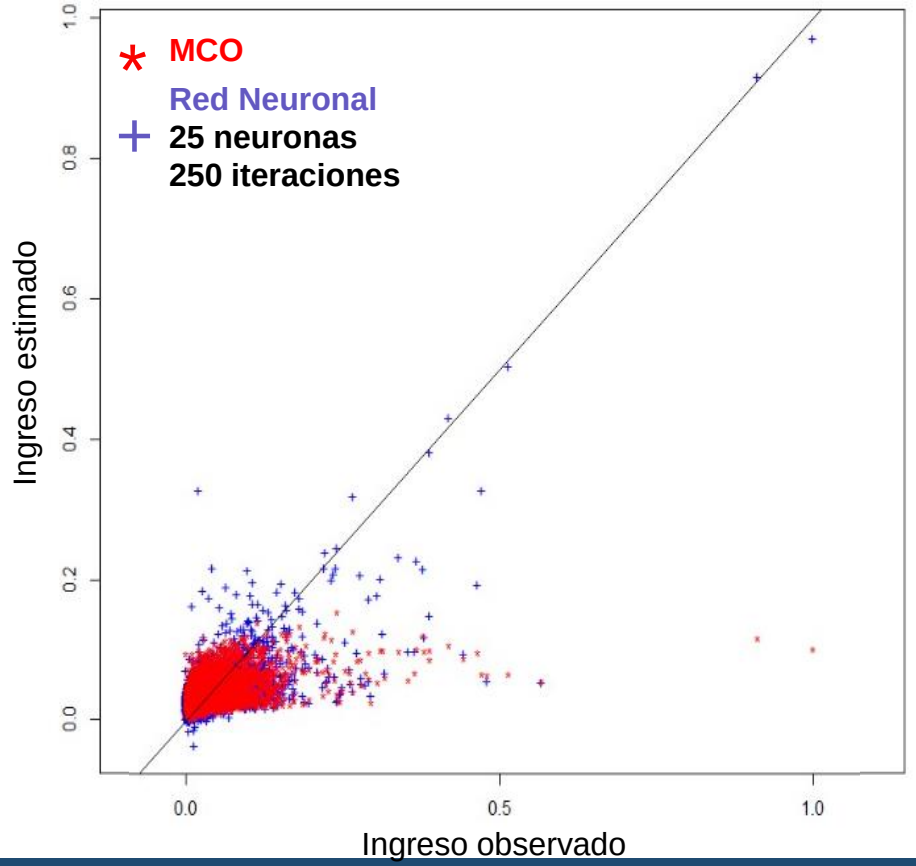
Validación del comportamiento entre los datos

El comportamiento de las variables es altamente similar entre ambas bases de datos, lo cual permite utilizar la información de la GEIH como proxy para la aplicación de la red neuronal en el III CNA

Máximo nivel educativo

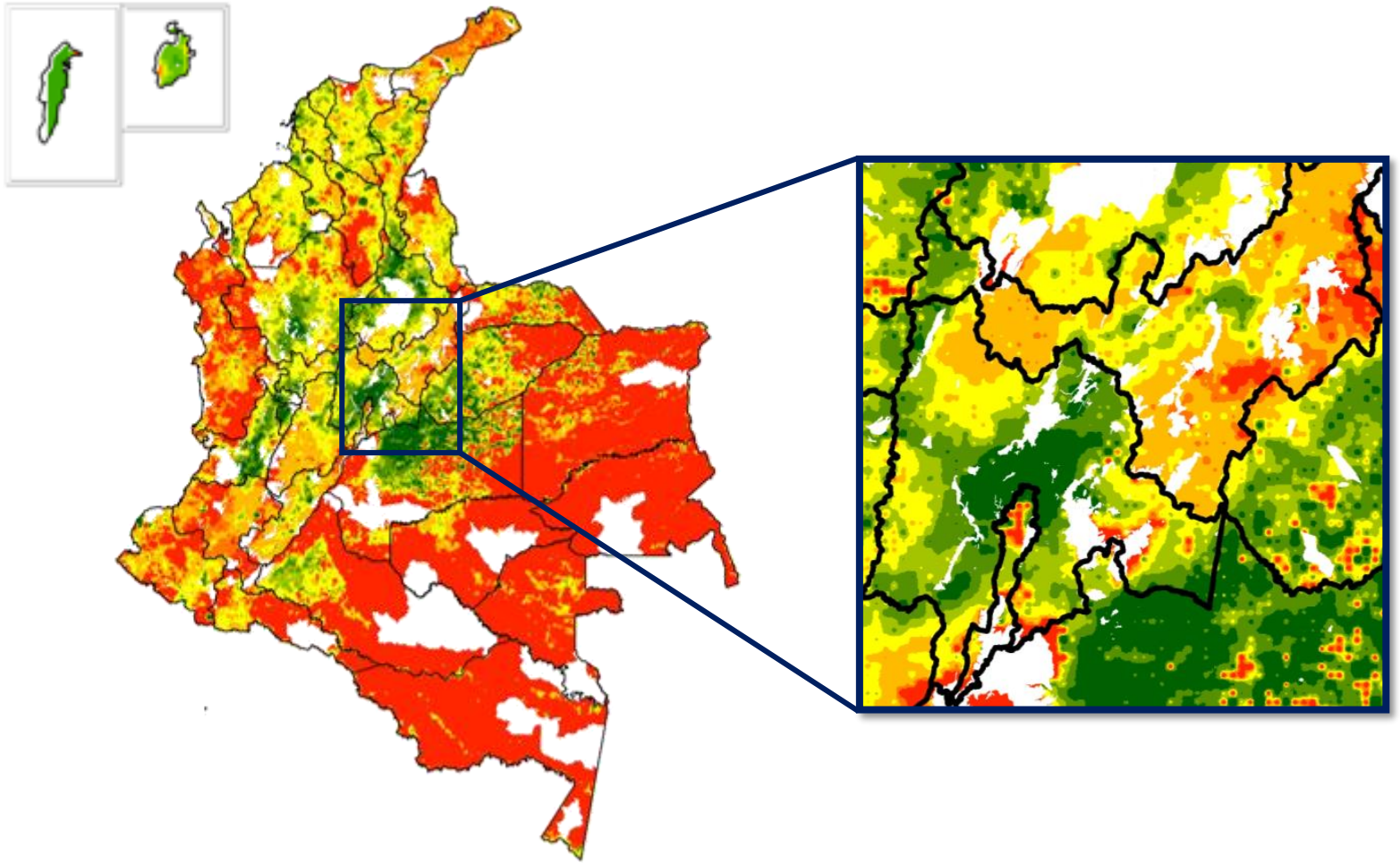


Validación entre modelos sobre el conjunto de prueba de la GEIH



Proyección final del mapa

Proyección de ingresos para 1.495.843 hogares rurales, se identifica altos ingresos en zonas cercanas a ciudades o cabeceras municipales



Leyenda

Zona reserva

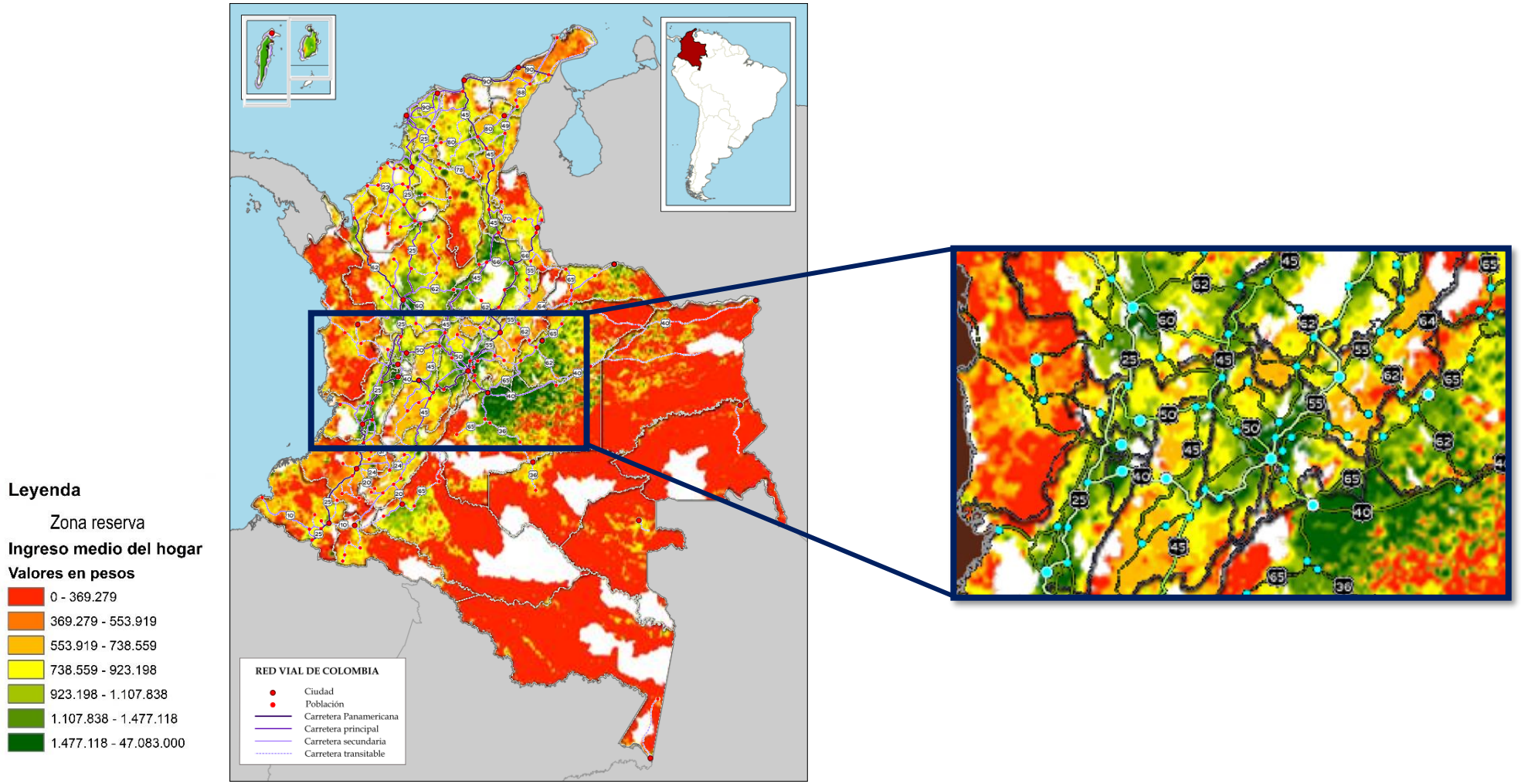
Ingreso medio del hogar

Valores en pesos

- 0 - 369.279
- 369.279 - 553.919
- 553.919 - 738.559
- 738.559 - 923.198
- 923.198 - 1.107.838
- 1.107.838 - 1.477.118
- 1.477.118 - 47.083.000

Cruce con vías primarias

Se evidencia altos ingresos en lugares donde pasa la maya vial primaria colombiana



Departamento Nacional de Planeación

www.dnp.gov.co