DIRECCIÓN DE DESARROLLO DIGITAL

Proyecto: Big data para la distribución de ingresos rurales

Unidad de Científicos de Datos

DNP Departamento Nacional de Planeación

GOBIERNO DE COLOMBIA

Unidad de Científicos de Datos – ucd@dnp.gov.co

Big data para la distribución de ingresos rurales

Tabla de contenido

Introducción

La Gran Encuesta Integrada de Hogares (GEIH) es una encuesta realizada por el DANE para recoger información para el cálculo de indicadores del mercado laboral. Adicionalmente, permite observar estructura de ingresos y gastos y de las condiciones de vida de los hogares. A través de la encuesta es posible obtener información sobre sexo, edad, estado civil y nivel educativo de las personas en los hogares.

Esta encuesta es aplicada tanto en zonas rurales como urbanas en los departamentos de Colombia exceptuando los antiguos territorios nacionales (Casanare, Putumayo, Arauca, Guaviare, San Andrés, Amazonas, Vichada, Vaupés y Guainía) por lo que su representatividad es solo a nivel nacional y no a nivel territorial. Adicionalmente, con la información suministrada con esta encuesta, el DANE calcula indicadores económicos de interés tales como el índice de desempleo, tasas de ocupación, informalidad, entre otros.

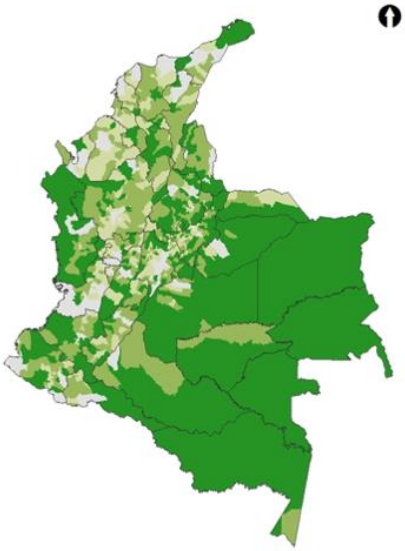
La información más reciente de la zona rural reportada es el tercer Censo Nacional Agropecuario (CNA), esta base de datos proporciona información estadística, georreferenciada o de ubicación satelital y actualizada del sector agropecuario del país. Según el DANE este censo cuenta con una cobertura operativa del 98.9%, cubriendo los 1.101 municipios del país, el archipiélago de San Andrés, Providencia y Santa Catalina, 32 departamentos, 20 áreas no municipalizadas, 773 resguardos indígenas, 181 tierras de comunidades negras y 56 parques nacionales naturales.

Sin embargo, el CNA no proporciona información económica de los hogares y viviendas rurales, por lo que Colombia no conoce cómo se distribuyen los ingresos en estas zonas y mediante el uso de *machine learning* se busca disponer de un primer mapa del territorio nacional que visualice la distribución de ingresos del territorio rural.

En este documento se relaciona la metodología utilizada, los resultados de las pruebas estadísticas realizadas y el resultado final cuya salida es el mapa de calor deseado que representa la distribución de ingresos en el territorio rural colombiano.

Motivación

Colombia cuenta con xx% de territorio definido como rural y rural disperso y que según la información económica que provee la GEIH, aquellas zonas rurales que se encuentran distantes a los principales centros urbanos presentan mayores niveles de pobreza.



Categorías de ruralidad[1]:

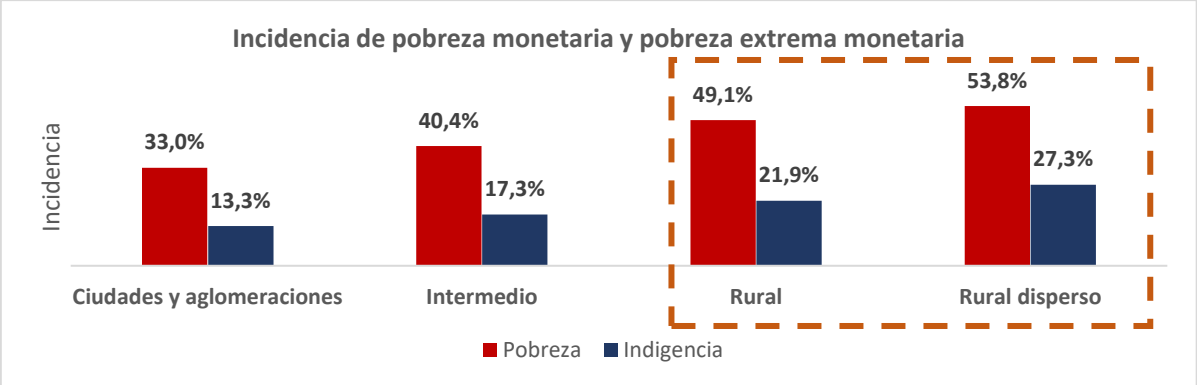
Ciudades y aglomeraciones: Cabeceras municipales de gran tamaño – **Principales centros urbanos del país**

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

Bajo estas categorías de ruralidad y la información provista por la GEIH permite conocer qué porcentaje de la población encuestada se encuentra en una situación de pobreza y pobreza extrema siendo para rural un 49,1% y 21,9% respectivamente, seguido de un 53,8% y 27,3% para rural disperso.



[1] (DNP, 2015)

Determinando la distribución de ingresos rurales del país, es posible conocer cómo los entornos territoriales afectan la generación de ingresos de los pequeños productores agropecuarios.

El uso de machine learning es considerado por Varian[2] como el desarrollo de sistemas informáticos de alto rendimiento que puedan proporcionar predicciones útiles sea para clasificación, o en este caso, para la proyección de un valor utilizando múltiples funciones de regresión.

A partir de los datos reportados en la GEIH, se pretende realizar una red neuronal que aprenda los patrones que explican el ingreso de los hogares rurales y con esta red predecir el ingreso para cada hogar reportado en el CNA.

Metodología

Información general de las fuentes de datos

Las bases de datos utilizadas se encuentran en el servidor interno de la Dirección de Desarrollo Rural Sostenible (DDRS) y cuentan con las siguientes características:

Observaciones	GEIH	CNA
Número de Viviendas	XX	XX
Número de Hogares	1.837.647	1.543.134
Número de Personas	XX	XX
Número de Variables	XX	XX

[2] (Varian, 2013)

Se procede a validar cuales variables existen en común para ambas fuentes de información, como resultado se obtiene la siguiente tabla:

Variable GEIH	Variable CNA	DESCRIPCIÓN GEIH	DESCRIPCIÓN CNA
VIV009	p_s15p162	Material predominante en las paredes exteriores	Material predominante de las paredes exteriores:
VIV012	p_s15p163	Material predominante en los pisos	Material predominante de los pisos:
VIV017	p_s15p164_sp1	¿La vivienda cuenta con servicio de energía eléctrica?	¿Con cuáles de los siguientes servicios públicos; privados o comunales cuenta la vivienda? Energía eléctrica
VIV016	p_s15p164_sp2	¿La vivienda cuenta con servicio de alcantarillado?	¿Con cuáles de los siguientes servicios públicos; privados o comunales cuenta la vivienda? Alcantarillado
VIV015	p_s15p165_sp3	¿La vivienda cuenta con servicio de acueducto?	¿Con cuáles de los siguientes servicios públicos; privados o comunales cuenta la vivienda? Acueducto
PER001	p_s15p167	Parentesco de las personas encuestadas con el jefe del hogar recodificado a la clasificación anterior al 2001	Parentesco con el jefe de hogar
PER004	p_s15p168	Género	Sexo

PER007	p_s15p169	Edad	¿Cuántos años cumplidos tiene?
EDU010	p_s15p173	Alfabetismo. ¿Sabe leer o escribir?	¿sabe leer y escribir español?
EDU001	p_s15p174	Asistencia Escolar	¿Actualmente estudia? (asiste actualmente a preescolar; escuela; colegio o universidad)
EDU004	p_s15p175a	Nivel educativo más alto alcanzado	Nivel educativo más alto alcanzado
SALU004	p_s15p176	¿A cuál de los regímenes de seguridad social en salud está afiliado?	En salud está afiliado(a) a:
SALU001	p_s15p176==4	¿Está afiliado o es cotizante o beneficiario a una EPS o ARS?	En salud está afiliado(a) a:

Selección de variables explicativas al ingreso por hogar

Las anteriores variables son pertenecientes a características de la vivienda, de los hogares y de las personas (entendiéndose que en una vivienda pueden existir más de un hogar) y luego validando para la GEIH cuáles variables pueden explicar mejor el ingreso (variable h_ing003) con el método de regresión Stepwise utilizando el acercamiento de “eliminación hacia atrás” obteniendo el siguiente resultado:

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***
Baldosín	0.008***	Especial	-0.011***
Mármol	0.056***	Subsidiado	-0.007***
Madera Pulida	0.017***	No sabe	0.021***
Alfombra	0.079***	Efectos por departamento	Sí
Energía	0.002***	Observaciones	73,311,708
Alcantarillado	0.001***	R-cuadrado	0.221
Acueducto	0.000***	Errores estándar en paréntesis	
Hombre	0.001***	*** p<0.01, ** p<0.05, * p<0.1	

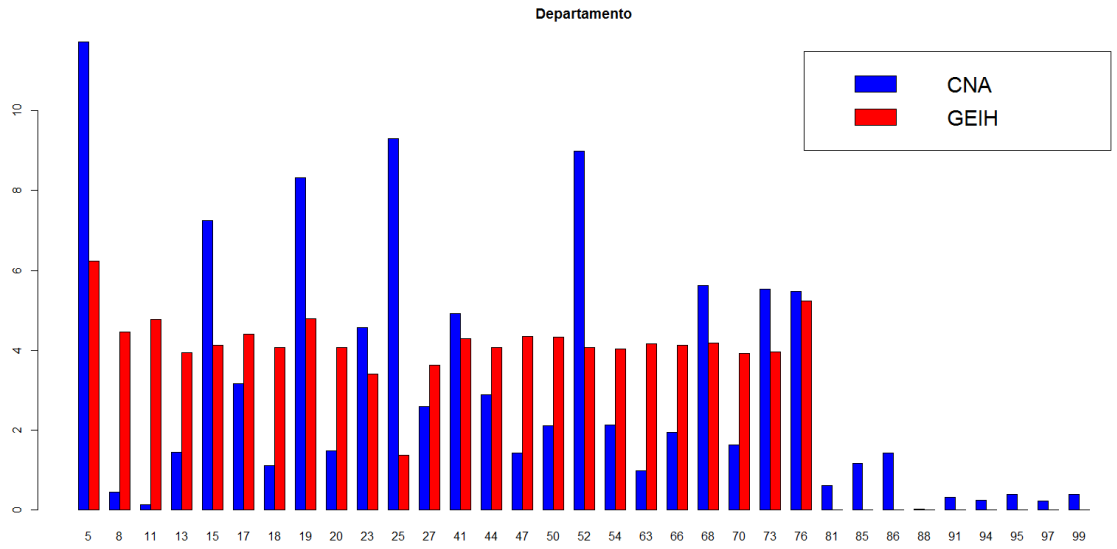
Identificando aquellas variables categóricas como dicotómicas en el modelo (razón por la que algunas variables tienen _x al lado de su nombre) y resultado ser todas significativas con un p<.01.

Adicionalmente, al querer pronosticar los ingresos por hogar, es necesario realizar alguna transformación a las variables que están a nivel personas, para ello, se decidió tratar estas variables como dicotómicas contando el número de personas dentro del hogar que cumplían con cada opción de respuesta y para la única variable continua, la edad, se transformó a categórica dividiendo el espectro de años en tres posibles conjuntos:

- Personas menores de 12 años (población dependiente)
- Personas entre los 12 y los 65 años (población económicamente activa)
- Personas mayores a 65 años (población dependiente)

Luego de convertirla en categórica se aplicó la misma lógica de conteo por hogar. Después de realizar una homologación entre categorías para ambas bases de datos (dado que las variables no contaban con el mismo orden de categorías, e incluso, en algunos casos fue necesario recodificar la clasificación para ambas fuentes de información) se procedió a validar el comportamiento de su distribución y realizar para cada variable una prueba estadística llamada (Student's t-test) para determinar si dos conjuntos de datos son significativamente diferentes entre sí, obteniendo los resultados adjuntos en el anexo 1.

Adicionalmente se identificó la variable departamento como la forma de modelar el componente espacial de las viviendas, obteniendo la siguiente distribución entre ambas fuentes de información:



Paired t-test	
t = 8.3588e-19, df = 32, p-value = 1	
alternative hypothesis: true difference in means is not equal to 0	
95 percent confidence interval:	-0.992772 0.992772
sample estimates:	
mean of the differences	4,07E-13

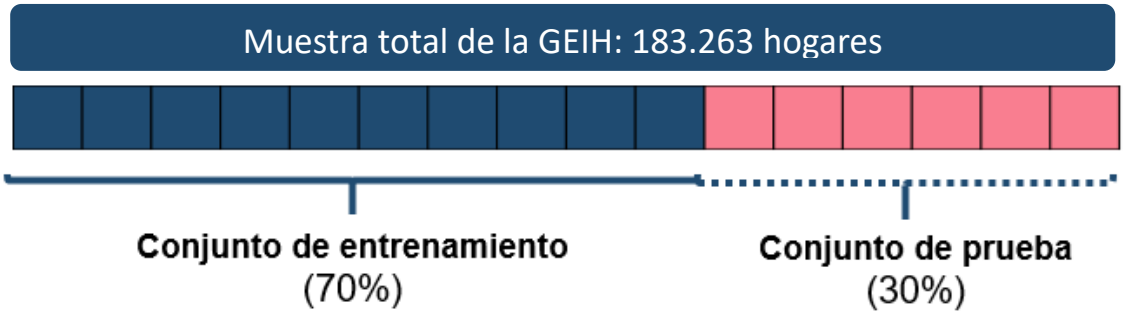
Como se puede observar en la gráfica anterior, la GEIH solamente tiene datos para 24 departamentos -representados por su código DANE en el eje X-, por lo que al

CNA se le realizó una división en dos partes, la primera parte con un número de observaciones de 1.468.925 que representa los datos del CNA que contienen los departamentos de la GEIH y una segunda parte con 74.209 observaciones para los departamentos que no aparecen en la GEIH (también mencionados anteriormente como territorios antiguos).

Lo anterior se realiza con el fin de crear dos redes neuronales por separado. Ambas redes neuronales parten de trabajar primero con los datos de la GEIH, la primera red se entrena con la variable departamento como parte del modelo mientras que la segunda se entrena sin esta variable para luego ser utilizadas en cada subconjunto de datos respectivamente.

Creación de las redes neuronales

En machine learning es necesario dividir la muestra en dos subconjuntos:



Dicha división se da con el fin de poder tener un grupo de datos con los que la máquina aprenderá los patrones detrás de los datos y otro grupo de control con el que se valida lo aprendido comparándolo con los datos observados.

Se utilizó el paquete *"nnet"* de R para crear una red neuronal con una capa única oculta y partir de múltiples combinaciones entre número de neuronas y número de iteraciones en que la red se entrena para ir disminuyendo el término de error en cada iteración.

Para este ejercicio se utilizaron las siguientes combinaciones:

neuronas <- c(5,10,15,25,50,100,150)

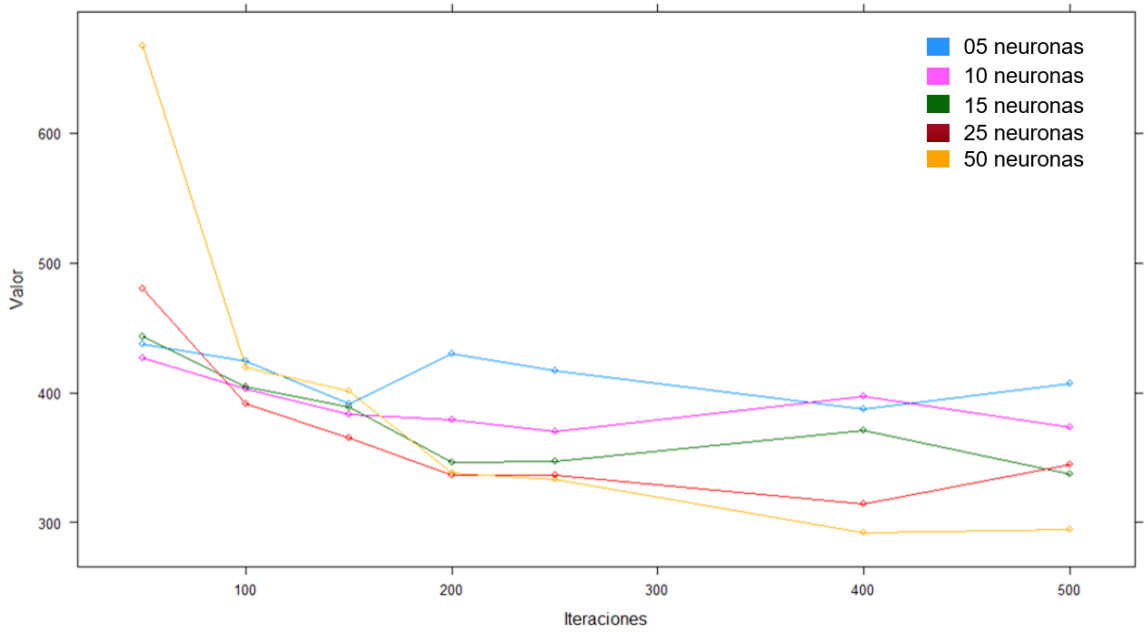
iteraciones <- c(25,50,100,150,200,225,250,400,500,550,600)

Luego se registró el resultado final del error para cada posible combinación y posteriormente se pronosticó el ingreso para el subconjunto del 30% para calcular la suma de errores al cuadrado entre el dato proyectado y el dato real del ingreso.

Es importante resaltar que para "nnet"[3] los valores del ingreso fueron transformados puesto que el paquete inicialmente no lograba realizar las iteraciones que le permitiera aprender los patrones detrás de los datos debido a la magnitud del valor en pesos colombianos. Para esto se transformó la variable de ingresos y se creó una nueva variable *Y* con la división del valor ingresos sobre el valor máximo de ingresos en toda la GEIH.

Los resultados fueron los siguientes:

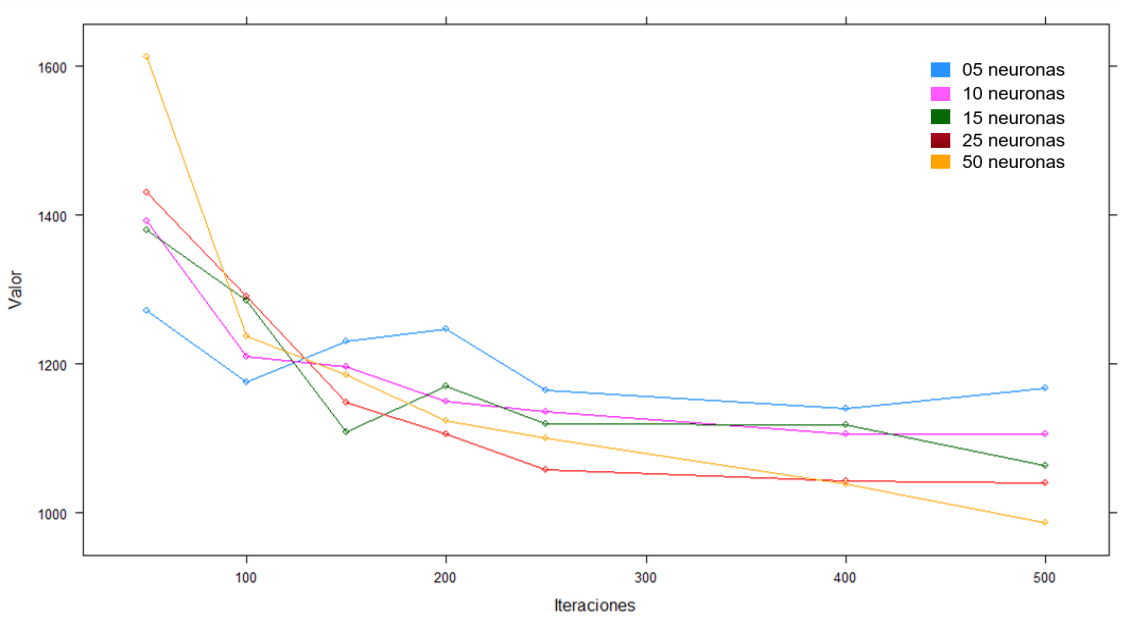
Resultados sobre el conjunto de entrenamiento



En el eje X se encuentra el número de iteraciones y el eje Y es el valor final registrado por cada iteración.

[3] (r Brian Ripley [aut, 2016)

Resultados sobre el conjunto de prueba



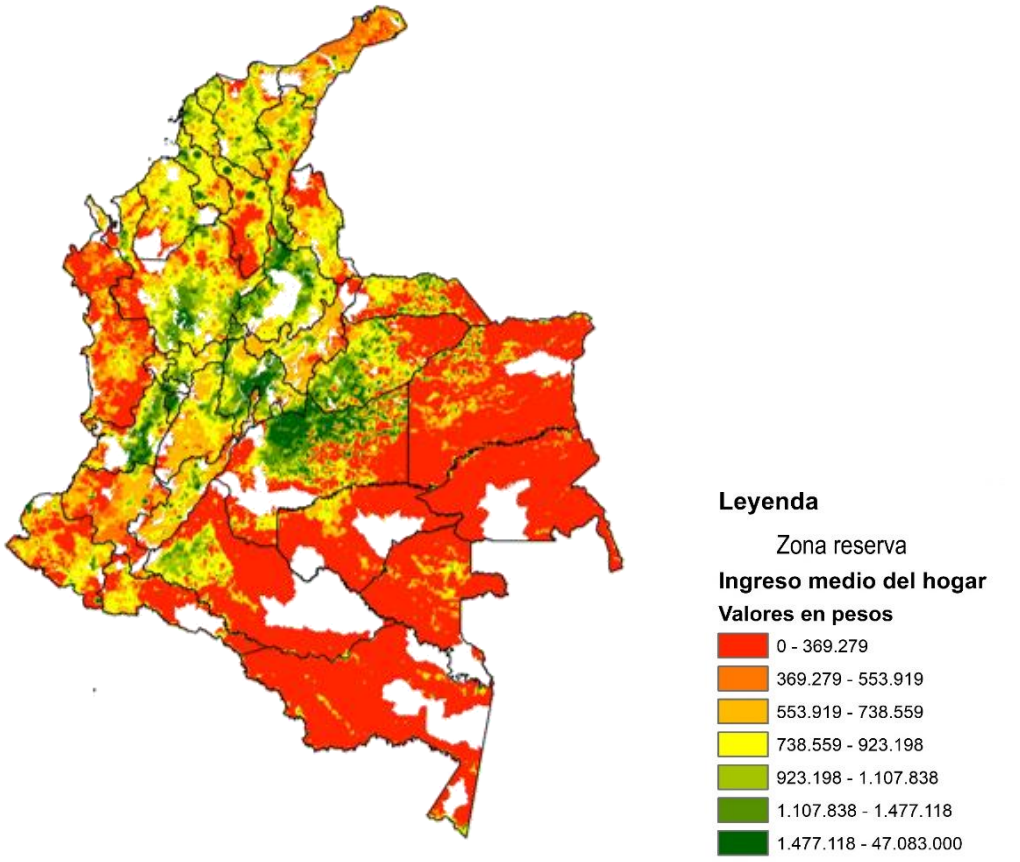
En el eje X se encuentra el número de iteraciones y el eje Y es el valor final registrado para la suma de errores al cuadrado en el set de prueba.

La selección de neuronas e iteraciones fue de 25 – 250 respectivamente puesto que resalta la estabilidad y el comportamiento de la curva suavizada (preferible al momento de entrenar una red neuronal y evitar overfitting de los datos).

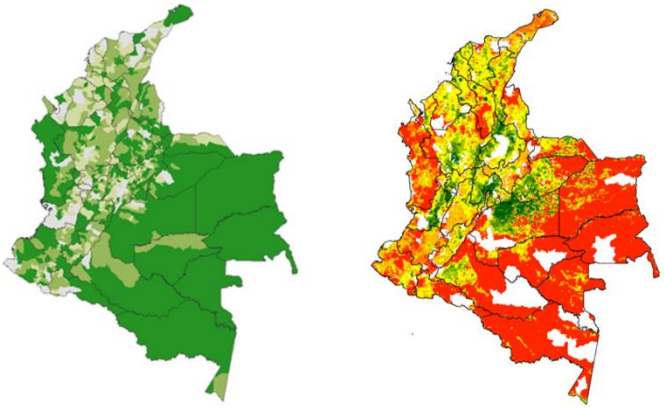
Resultados

Para los resultados se obtuvo un primer acercamiento visual del resultado utilizando las coordenadas geográficas de las viviendas en el CNA y con el software ARCGIS se procedió a vectorizar sobre el mapa departamental de Colombia. Para tener una mejor representación de los resultados, se utilizó una grilla de 5x5 kilómetros sobre todo el territorio nacional y la mediana del conjunto de puntos que se encontraran dentro de cada cuadro de la grilla.

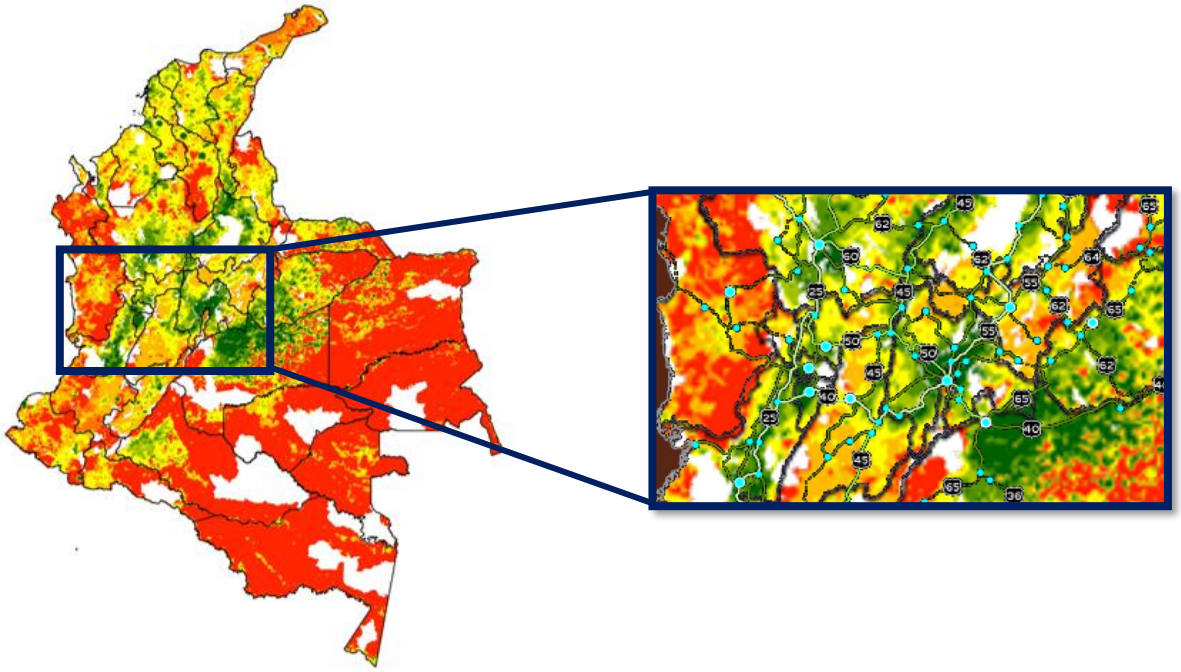
Mapa con la proyección de ingresos mensuales por hogar



Se puede evidenciar que las zonas con mayor proyección de ingresos por hogares se encuentran cercanas a las principales ciudades y aglomeraciones del país. Mientras que las zonas con menores ingresos se encuentran en zonas rurales dispersas tal y como se encuentra estructurada la categorización municipal definida anteriormente.



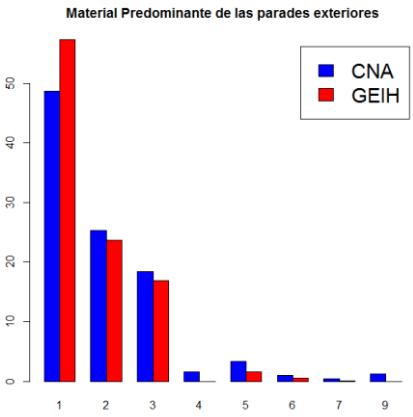
De hecho, si se hace un acercamiento a las zonas cercanas a las principales ciudades del país se puede observar una similitud entre las zonas con ingresos altos y la malla vial primaria:



Como primer acercamiento a la influencia de las condiciones territoriales para la generación de ingresos en los hogares rurales colombianos, se puede ver que en la medida que Colombia mejore su conectividad vial y existan más zonas conectadas a la malla vial primaria, al menos con el análisis visual, pueden mejorar los ingresos mensuales provenientes de actividades rurales.

Anexos

Anexo 1: Análisis exploratorio de las variables seleccionadas



Paired Test

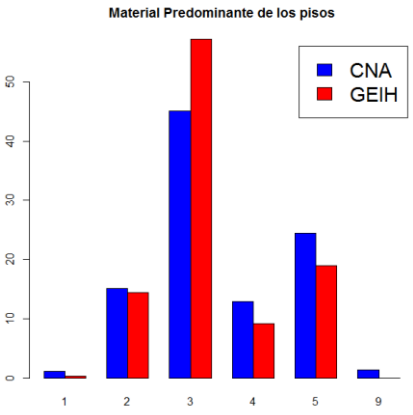
t = -6.6653e-16, df = 7, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -2.929417 2.929417

sample estimates:

mean of the differences: -8.257284e-16



Paired Test

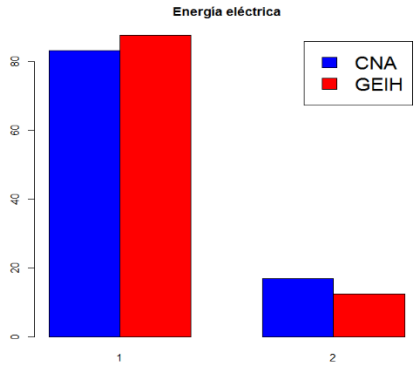
t = 3.8835e-16, df = 5, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -6.491344 6.491344

sample estimates:

mean of the differences: 9.806789e-16



Paired t-test

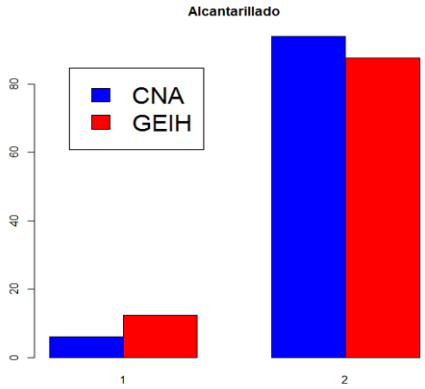
t = -1.5978e-15, df = 1, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -56.50451 56.50451

sample estimates:

mean of the differences -7.105427e-15



Paired t-test

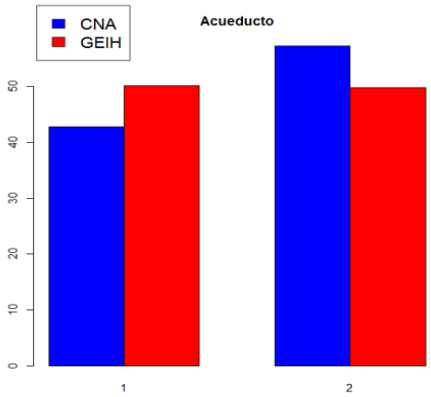
t = -7.0932e-17, df = 1, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -79.55045 79.55045

sample estimates:

mean of the differences -4.440892e-16



Paired t-test

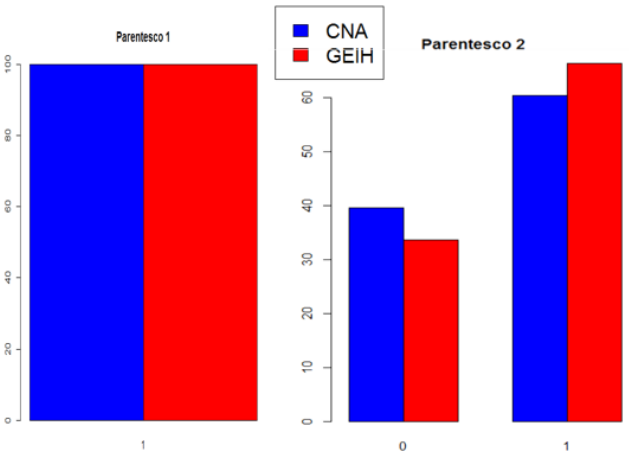
t = 0, df = 1, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -94.46529 94.46529

sample estimates:

mean of the differences 0



Paired t-test

t = 7.2971e-16, df = 2, p-value = 1

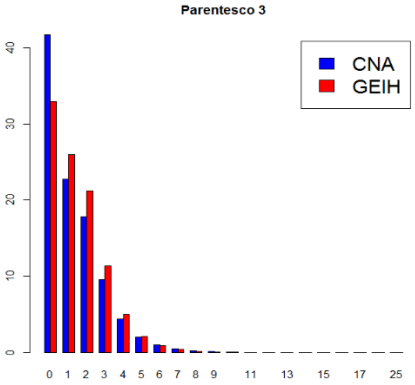
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-14.85198 14.85198

sample estimates:

mean of the differences
2.518818e-15



Paired t-test

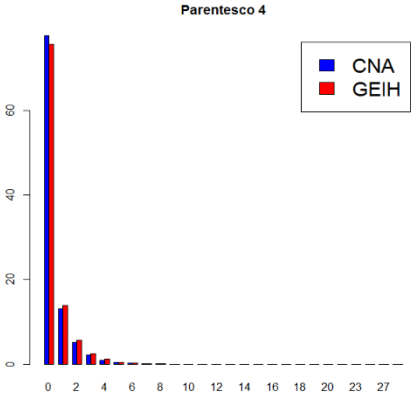
t = 6.0941e-17, df = 19, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -1.080357 1.080357

sample estimates:

mean of the differences 3.145607e-17



Paired t-test

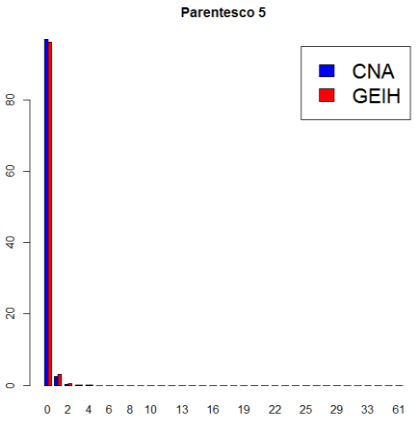
t = 3.2741e-15, df = 25, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.179266 0.179266

sample estimates:

mean of the differences 2.849797e-16



Paired t-test

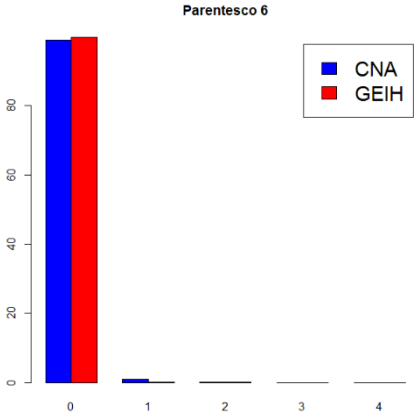
t = -6.9761e-15, df = 34, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.06110283 0.06110283

sample estimates:

mean of the differences -2.097487e-16



Paired t-test

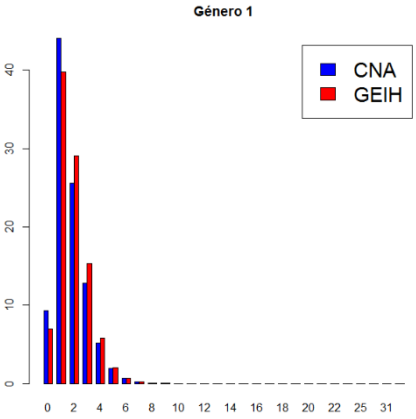
t = -8.953e-15, df = 4, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.6502365 0.6502365

sample estimates:

mean of the differences -2.096768e-15



Paired t-test

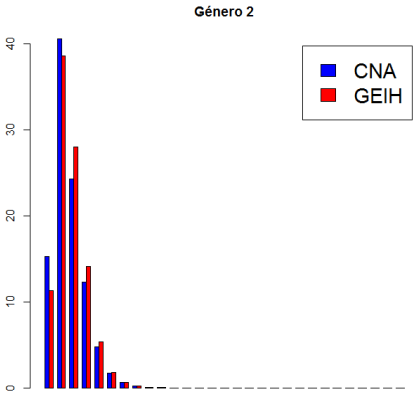
t = -1.616e-15, df = 27, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.488135 0.488135

sample estimates:

mean of the differences -3.844508e-16



Paired t-test

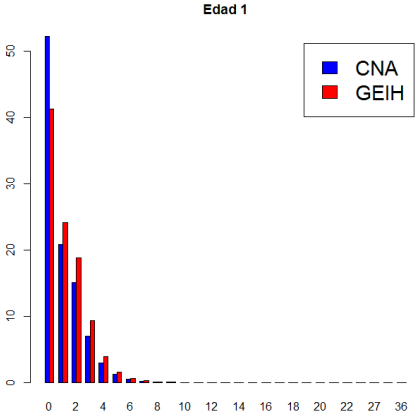
t = 9.1578e-16, df = 28, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.4360385 0.4360385

sample estimates:

mean of the differences 1.949396e-16



Paired t-test

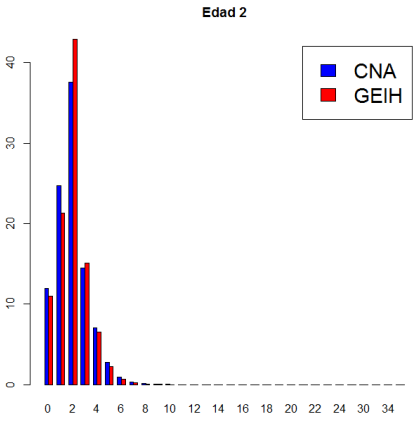
t = -5.7438e-16, df = 26, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.9549025 0.9549025

sample estimates:

mean of the differences -2.6683e-16



Paired t-test

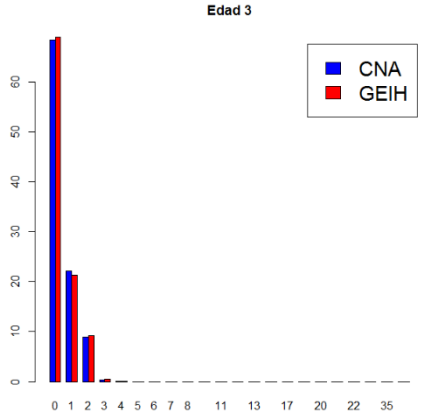
t = -8.2473e-16, df = 29, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.4511801 0.4511801

sample estimates:

mean of the differences -1.819374e-16



Paired t-test

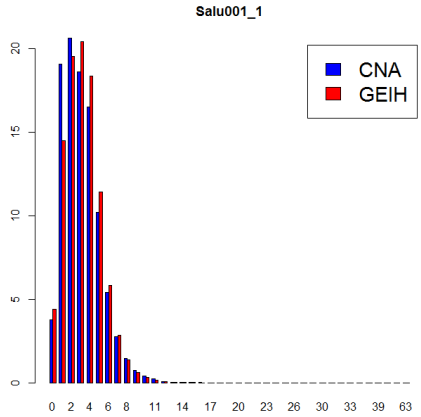
t = -1.5019e-14, df = 21, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.09972893 0.09972893

sample estimates:

mean of the differences -7.202341e-16



Paired t-test

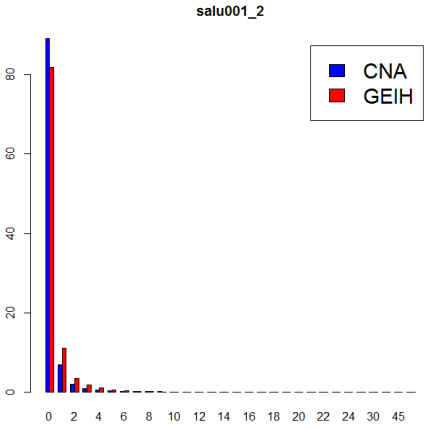
t = 3.8587e-16, df = 38, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.2926636 0.2926636

sample estimates:

mean of the differences 5.578464e-17



Paired t-test

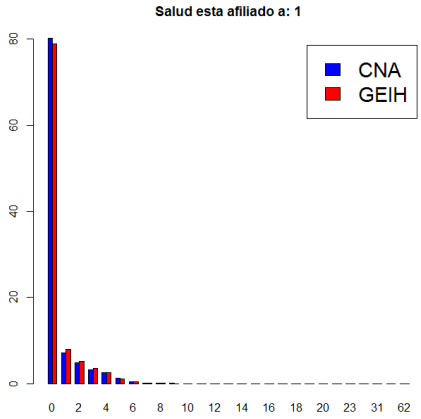
t = 4.1048e-16, df = 29, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.5958777 0.5958777

sample estimates:

mean of the differences 1.195927e-16



Paired t-test

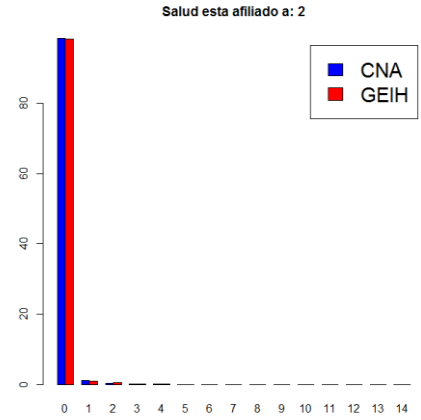
t = 8.7145e-16, df = 26, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.119462 0.119462

sample estimates:

mean of the differences 5.064647e-17



Paired t-test

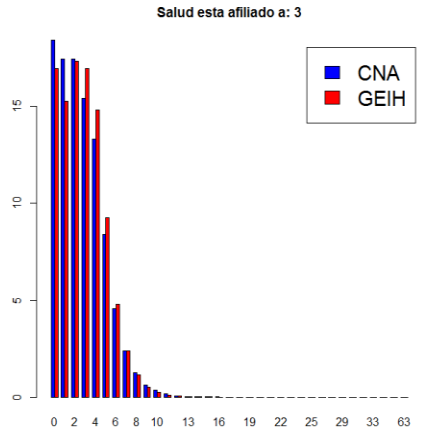
t = -1.8582e-14, df = 14, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.06154631 0.06154631

sample estimates:

mean of the differences -5.332278e-16



Paired t-test

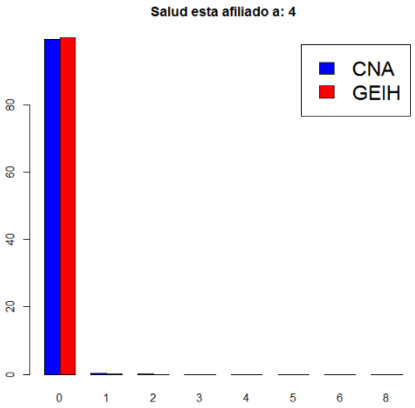
t = 2.1595e-15, df = 34, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.207412 0.207412

sample estimates:

mean of the differences 2.204031e-16



Paired t-test

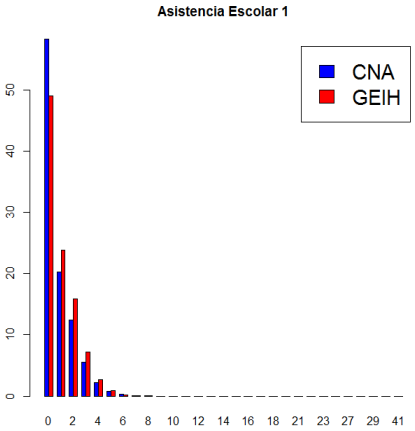
t = 1.4899e-14, df = 7, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.1245065 0.1245065

sample estimates:

mean of the differences 7.844702e-16



Paired t-test

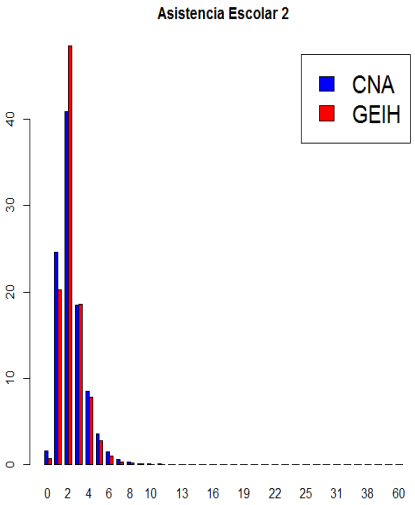
t = -2.7932e-16, df = 28, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.7617913 0.7617913

sample estimates:

mean of the differences -1.038792e-16



Paired t-test

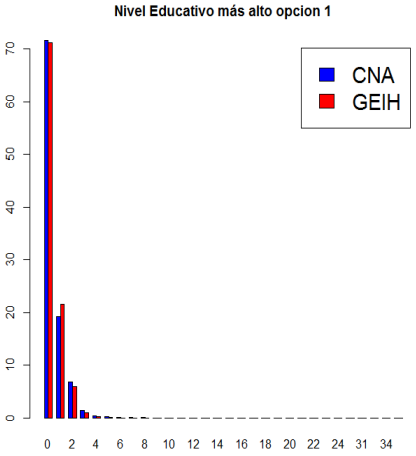
t = -5.9088e-16, df = 34, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.5220089 0.5220089

sample estimates:

mean of the differences -1.517766e-16



Paired t-test

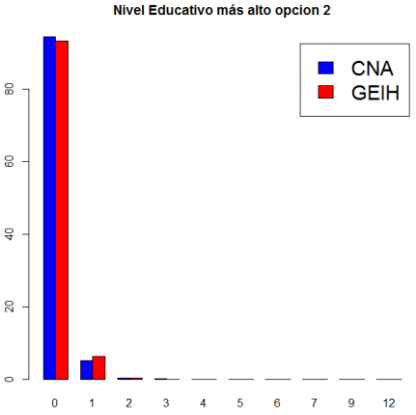
t = 1.0634e-15, df = 29, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.1827728 0.1827728

sample estimates:

mean of the differences 9.503246e-17



Paired t-test

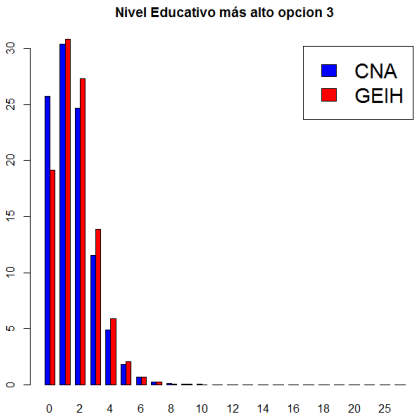
t = -3.9652e-15, df = 9, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.3747023 0.3747023

sample estimates:

mean of the differences -6.567899e-16



Paired t-test

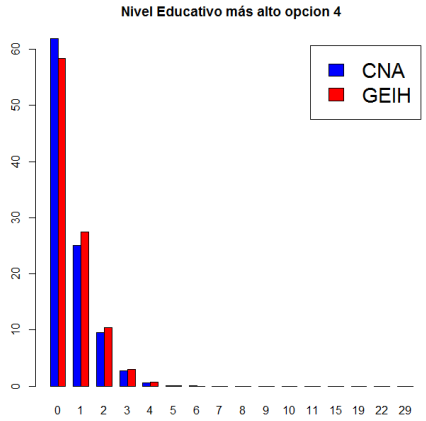
t = -2.677e-16, df = 23, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.6655122 0.6655122

sample estimates:

mean of the differences -8.612236e-17



Paired t-test

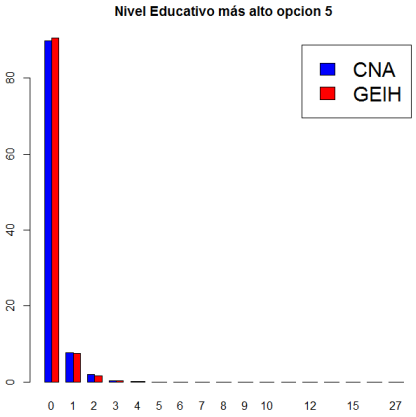
t = 9.1652e-16, df = 15, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.5992709 0.5992709

sample estimates:

mean of the differences 2.576869e-16



Paired t-test

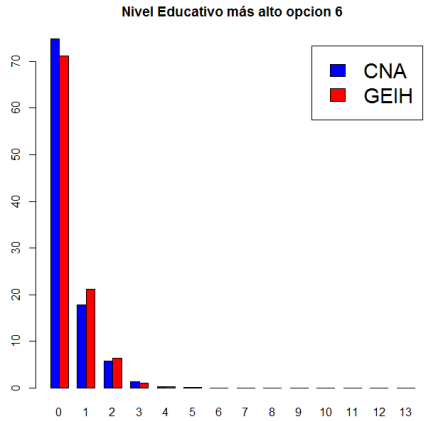
t = -9.1366e-15, df = 16, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.1133999 0.1133999

sample estimates:

mean of the differences -4.887434e-16



Paired t-test

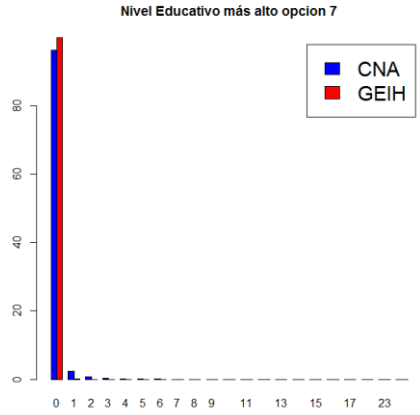
t = 8.4327e-16, df = 13, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.8161953 0.8161953

sample estimates:

mean of the differences 3.185914e-16



Paired t-test

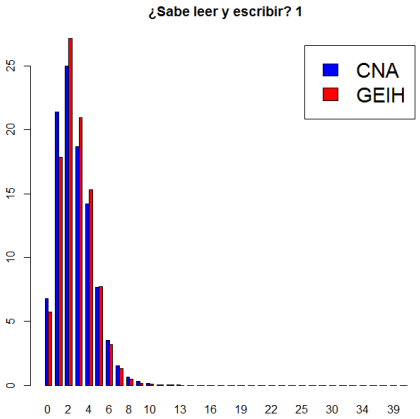
t = -6.799e-16, df = 20, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.4369517 0.4369517

sample estimates:

mean of the differences -1.4242e-16



Paired t-test

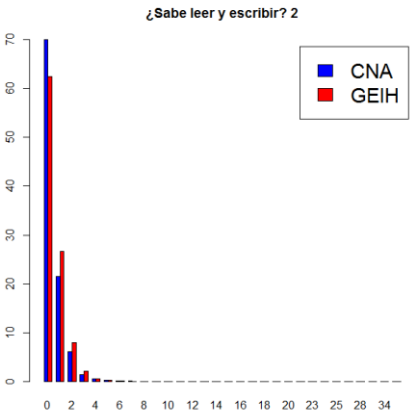
t = 1.199e-15, df = 35, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.2856024 0.2856024

sample estimates:

mean of the differences 1.686792e-16



Paired t-test

t = 2.2152e-16, df = 29, p-value = 1

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.6456029 0.6456029

sample estimates:

mean of the differences 6.992694e-17

Bibliografía

DNP. (2015). *Definición de Categorías de Ruralidad.* Bogotá DC: DNP. Obtenido de www.dnp.gov.co.

Varian, H. R. (2013). Big Data: New Tricks for Econometrics. *University of Berkeley*.