ANOVA
Evidence based thinking.

Uso de Algoritmos de Machine Learning en la Construcción de Mapas de Pobreza: Venezuela 2012-2016

Omar Zambrano

Taller regional sobre desagregación de estadísticas sociales mediante metodologías de estimación en áreas pequeñas. ONU-CEPAL.

Santiago de Chile
22/11/2018

Contenido

1. Motivación
2. Datos disponibles
3. Construcción de mapas de pobreza usando algoritmos de aprendizaje
   a. Aspectos metodológicos
   b. Resultados 2012-2016
4. Lecciones y Consideraciones Finales

ANOVA
Evidence based thinking.

Motivación

- Venezuela enfrenta la recesión más severa y prolongada de su historia moderna.
- El episodio recesivo ha tenido como correlato de crisis fiscal y externa que ha deprimido la actividad económica, cortado las cadenas de abastecimiento, suprimido el financiamiento externo, y producido un agudo brote inflacionario.
- En este contexto las capacidades del Estado para mitigar los efectos negativos del choque de ingresos están limitadas y se han erosionado al extremo las condiciones del mercado laboral.
- La apropiada evaluación de los efectos socio-económicos de este episodio, sobre todo para la porción más vulnerable de la población, es aun limitado o inexistente.

Motivación (II)

- Encuestas de hogares no oficiales han ofrecido respuestas parciales sobre magnitud del choque de ingresos. Sin embargo, datos oficiales no disponibles o descontinuados.
- En este contexto poco se ha dicho sobre la dimensión distributiva y los aspectos territoriales en la contracción económica.
- ¿Se han incrementado la inequidades horizontales y espaciales del fenómeno de la pobreza? ¿hay heterogeneidad en los resultados espaciales observados? ¿hay efectos de concentración o elementos territoriales que informen estrategias de focalización de las políticas?

Datos Disponibles

- Datos oficiales:
  1. Encuestas de Hogares por Muestreo del Instituto Nacional de Estadísticas 2012 – 2105. Principal investigación estadística del país. Muestreo probabilístico, estratificado bifásico. Con base a segmentos censales. Representatividad estadal con N=38.000 cada semestre.
  2. Censo de Población y Vivienda 2010: 7.140.040 hogares, 31 millones personas (proyección demográfica)
- Datos no Oficiales:
  1. ESSEV: Encuesta de situación socioeconómica de los Venezolanos 2016. N=4.900
  2. ENCOVI: Encuestas de condiciones de Vida 2016. N=6.400

ANOVA
Evidence based thinking.

Aspectos metodológicos

- Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003), Micro–Level Estimation of Poverty and Inequality. Econometrica, 71: 355–364.
- Zhao, Q. and Lanjouw, P. (2009), Using PovMap2: A User´s Guide. World Bank.
- Poverty Mapping: procedimiento estadístico para combinar las dos fuentes de datos, aprovechar el grado de detalle de la encuesta de hogares, y la cobertura de los datos censales.

ANOVA
Evidence based thinking.

Aspectos metodológicos

- La idea básica:

$$(1) \quad Y_{h,c} = \beta X_{c,h} + \mu_{c,h}$$
$$(2) \quad \mu_{c,h} = \gamma_c + \varepsilon_{c,h}$$

$Y_{h,c}$: *Ingreso por hogar*
γ_c : *Cluster effect*
$\varepsilon_{c,h}$: *Household effect*

- De Elbers *et al*

  γ_c es una variable aleatoria con varianza definida

$$\text{var}\left(\widehat{\sigma_\eta^2}\right) \approx \sum_c [a_c^2 \text{var}\left(u_{c.}^2\right) + b_c^2 \text{var}(\widehat{\tau_c^2})] \approx \sum_c 2[a_c^2\{(\widehat{\sigma_\eta^2})^2 + (\widehat{\tau_c^2})^2 + 2\widehat{\sigma_\eta^2}\widehat{\tau_c^2}\} + b_c^2 \frac{(\widehat{\tau_c^2})^2}{n_c - 1}].$$

  $\varepsilon_{c,h}$ es ur $\widehat{\sigma}_{\varepsilon,ch}^2 = \left[\frac{AB}{1+B}\right] + \frac{1}{2}\widehat{\text{Var}}(r)\left[\frac{AB(1-B)}{(1+B)^3}\right]$. anza definida como:

ANOVA
Evidence based thinking.

Aspectos metodológicos

- *Machine Learning* y mapas de pobreza: puntos básicos.
  1. Problema básico de dominio y representatividad de las fuentes de datos que motiva metodologías SAE se mantiene.
  2. La estructura de los problemas típicos del enfoque *machine learning* es análoga al problema de los SAE:
     i. Estructura preferente de la información con *training set* (encuesta de hogares) y *scoring set* (censo).
     ii. Predicción probabilística basada en vectores de soporte común de variables entre dos fuentes de datos.
     iii. Foco en la predicción y clasificación fuera de la muestra de datos optimizan dimensionalidad y evita *overfitting*.
     iv. Formas no paramétricas relajan la necesidad de modelos predefinidos (aunque informan menos

Aspectos metodológicos

- Implementación del enfoque de *Machine Learning* para estimar mapas de pobreza:
  1. Construcción del vector de soporte común entre encuesta y censo. Definición de la variable objetivo.
  2. Tipos de algoritmos. Tipos de predicción. Sin modelos predefinidos, variables continuas (ingreso); variable dicotómica (pobre/no pobre); multinomiales (A, B...D).
  3. Ejemplos:
     a. Support Vector Machine: Separa el conjunto de datos clasificados con un hiperplano de n-1 dimensiones. Maximiza la distancia de los vectores de soporte al hiperplano central. Esta estructura funcional sirve como clasificadora.
     b. Random Decision Forest: Emplea un grupo de árboles de decisión creados aleatoriamente desde subconjuntos de covariates, evalúa sistemáticamente los nodos para escoger los más efectivos. Posibilidad de n árboles (forest) permite imputar probabilidad a cada observación

ANOVA
Evidence based thinking.

Aspectos metodológicos

a. Red Neuronal Artificial: Crea una serie de niveles interconectados por donde ”fluye” la información de los *covariates* de un nivel al siguiente. El aprendizaje ocurre al iniciar proceso de “backpropagation” permite identificar datos reales.
b. Naive Bayes Machine: Asume la independencia del impacto de cada variable sobre el valor de la variable dependiente. Adicionalmente, el modelo asume una distribución normal del valor de cada variable con respecto a la probabilidad de cada clasificación. Al realizar estas dos suposiciones, el modelo calcula la probabilidad de clasificación de la variable dependiente individualmente con respecto a cada variable explicativa, y luego utiliza la regla de Bayes para agregar los impactos de cada variable y producir una clasificación final.
c. Gradient Boosting Tree: árboles de decisión adaptables a variables contínuas, divide el espacio predictivo en estratos y segmentos de acuerdo a reglas arbitrarias. Asigna a cada obervación el promedio de la variable de interés del segmento en el que dicha observación finaliza

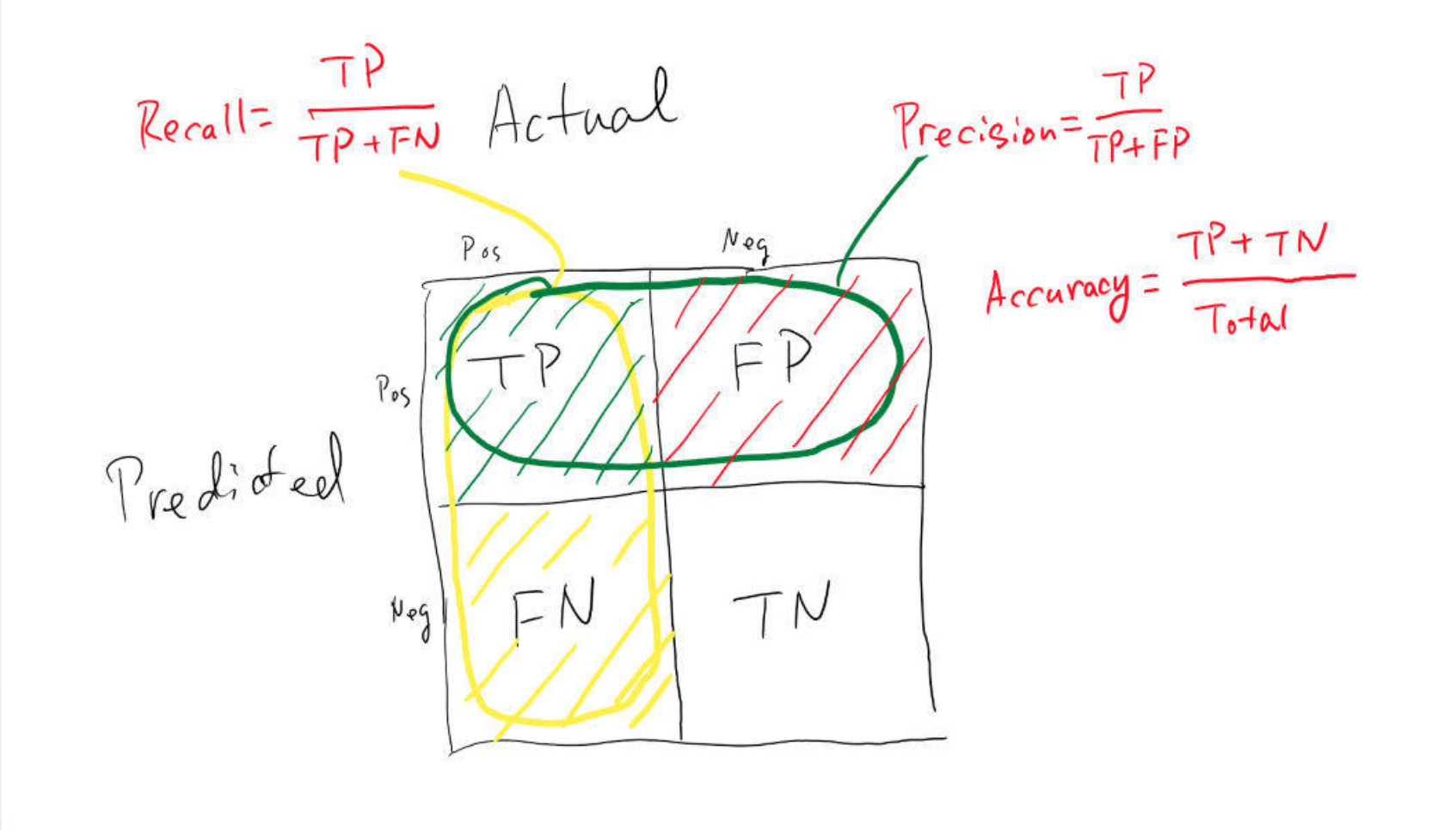
Aspectos metodológicos

- Criterios para evaluar poder de ajuste y predicción

1. Validación cruzada
   a. Fase de entrenamiento. Se divide el conjunto de datos en subconjuntos (4-10). Proceso iterativo, en una de las partes estima el modelo (conjunto de entrenamiento), las otra partes evalúan el modelo (conjunto de pruebas).
   b. Este proceso se repite con todas las partes y se aplica con todos los modelos para escoger el de mejor performance predictivo.

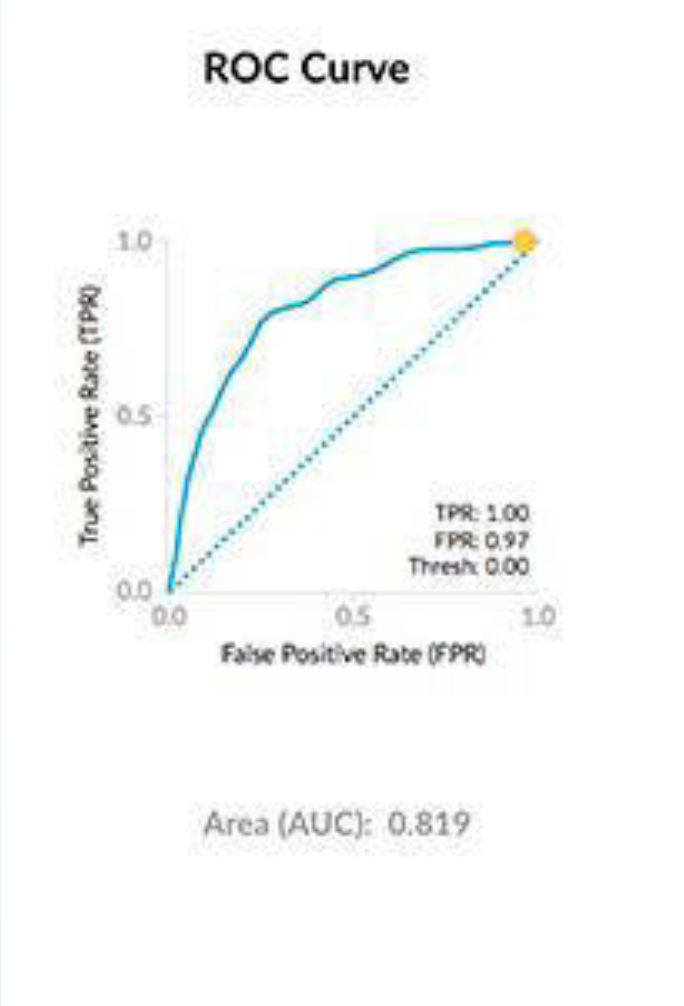
Aspectos metodológicos

2. Matriz de Confusión



Aspectos metodológicos

3. Curva ROC - AUC

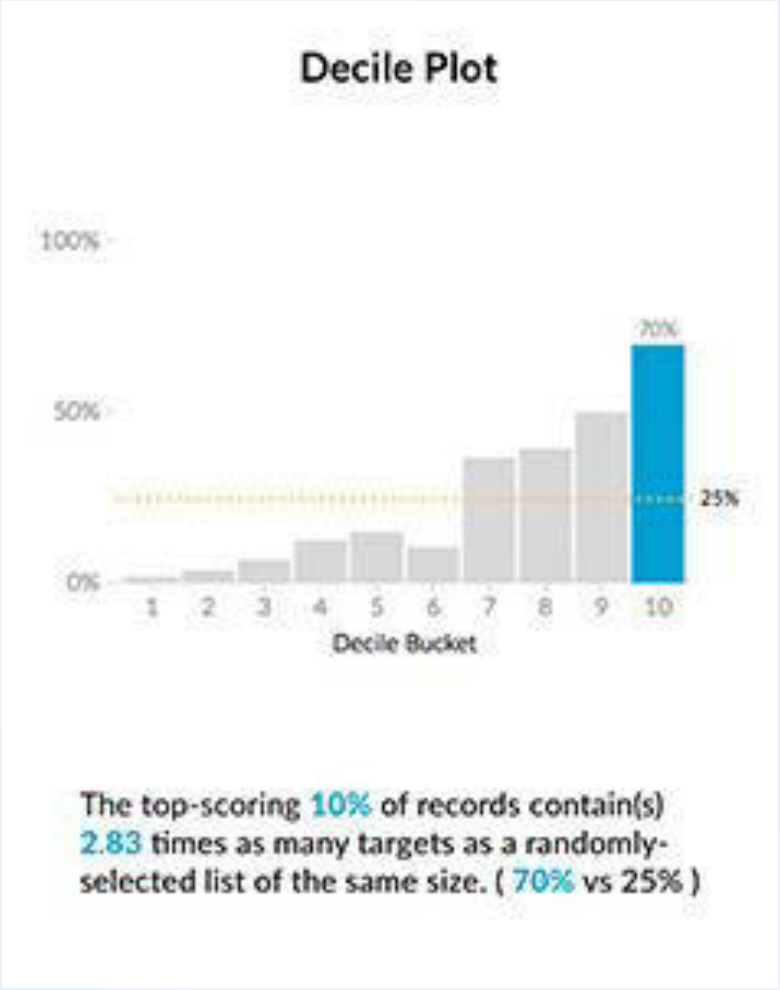


- Vértice superior izquierdo es el máximo posible.
- Área bajo la curva es un indicador de eficiencia global del modelo.

ANOVA
Evidence based thinking.

Aspectos metodológicos

4. Gráfico de deciles



- Indicador de rendimiento análogo a Logit/Probit
- Permite manejar la incertidumbre

ANOVA
Evidence based thinking.

Aspectos metodológicos

4. Caso Venezuela
   a. Algoritmo Random Decision Forest para el caso discreto. Score probabilidad.
   b. Algoritmo Gradient Boosting Tree para el caso continuo. Ranking.
   c. Reproducción de información de dominios representativos en la encuesta.

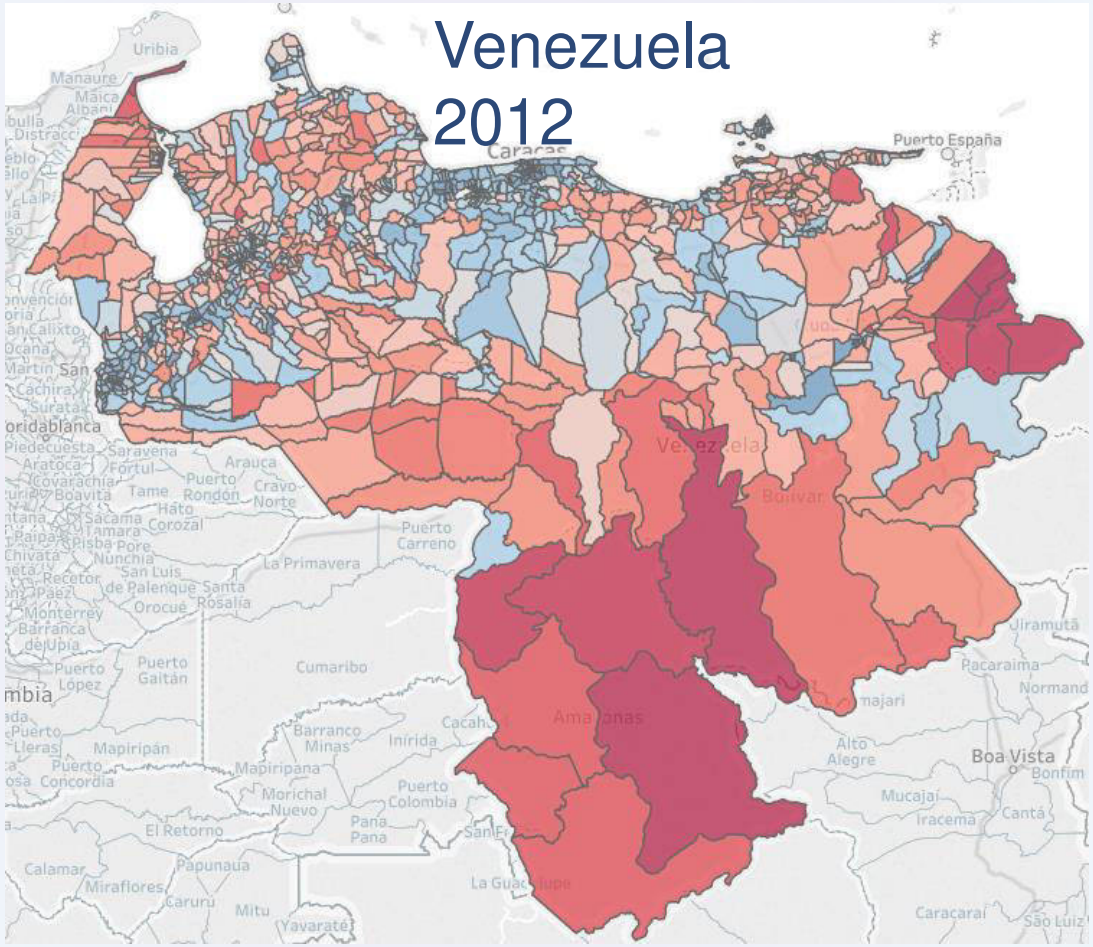
- Métricas: AUC>0.81; ACC>0.80; REC>0.72, PRE>0.66
- Luego de la fase de entrenamiento, evaluación y selección de modelos, se lleva a cabo la imputación de la clasificación o ingreso de +7.3 millones de hogares.
- Resultado: Estimación de FGT0, FGT0ext, FGT1, y FGT2 para 1125 parroquias de Venezuela.

ANOVA
Evidence based thinking.

Mapas de Pobreza Parroquial

Amplias porciones del territorio con pobreza moderada o inexistente. El eje Centro Norte Costero y Andes.

Tasa de pobreza de las 10 parroquias menos pobres: 4,1%.

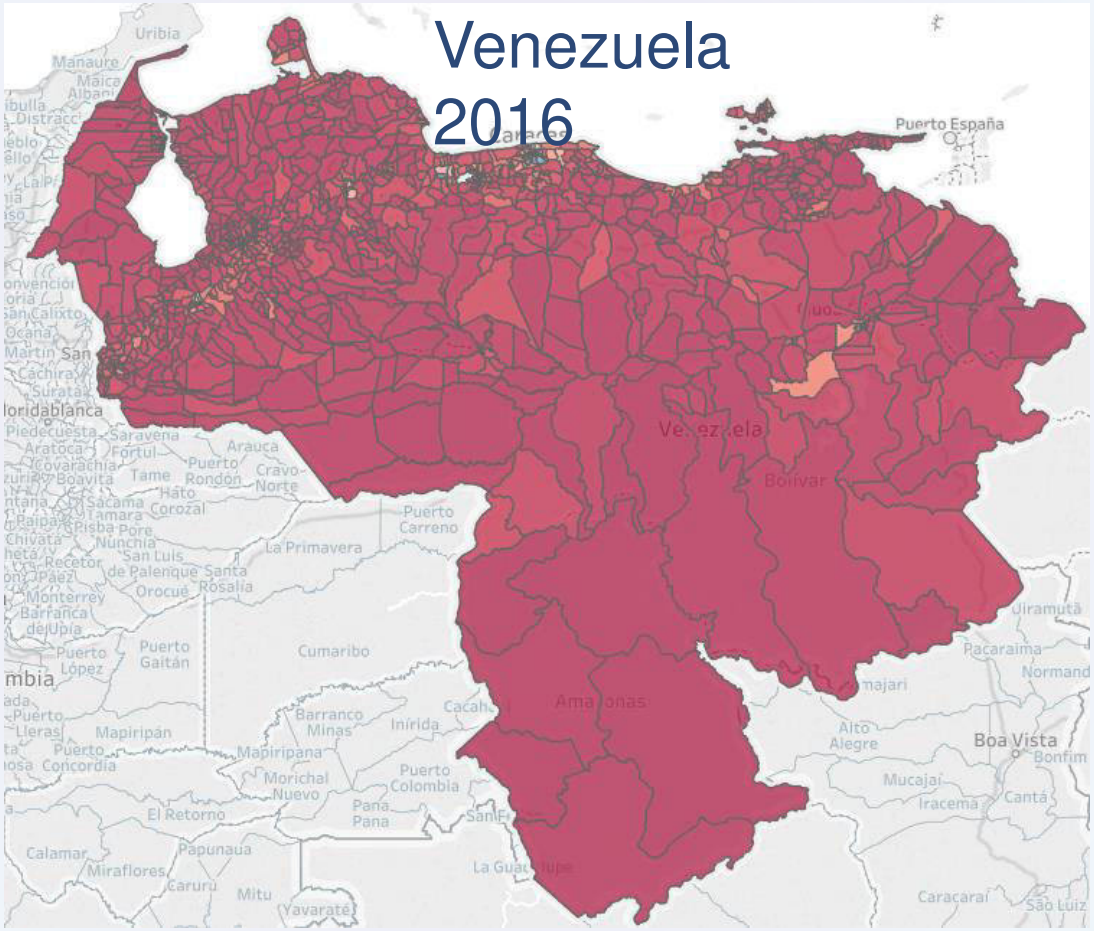


ANOVA
Evidence based thinking.

Mapas de Pobreza Parroquial

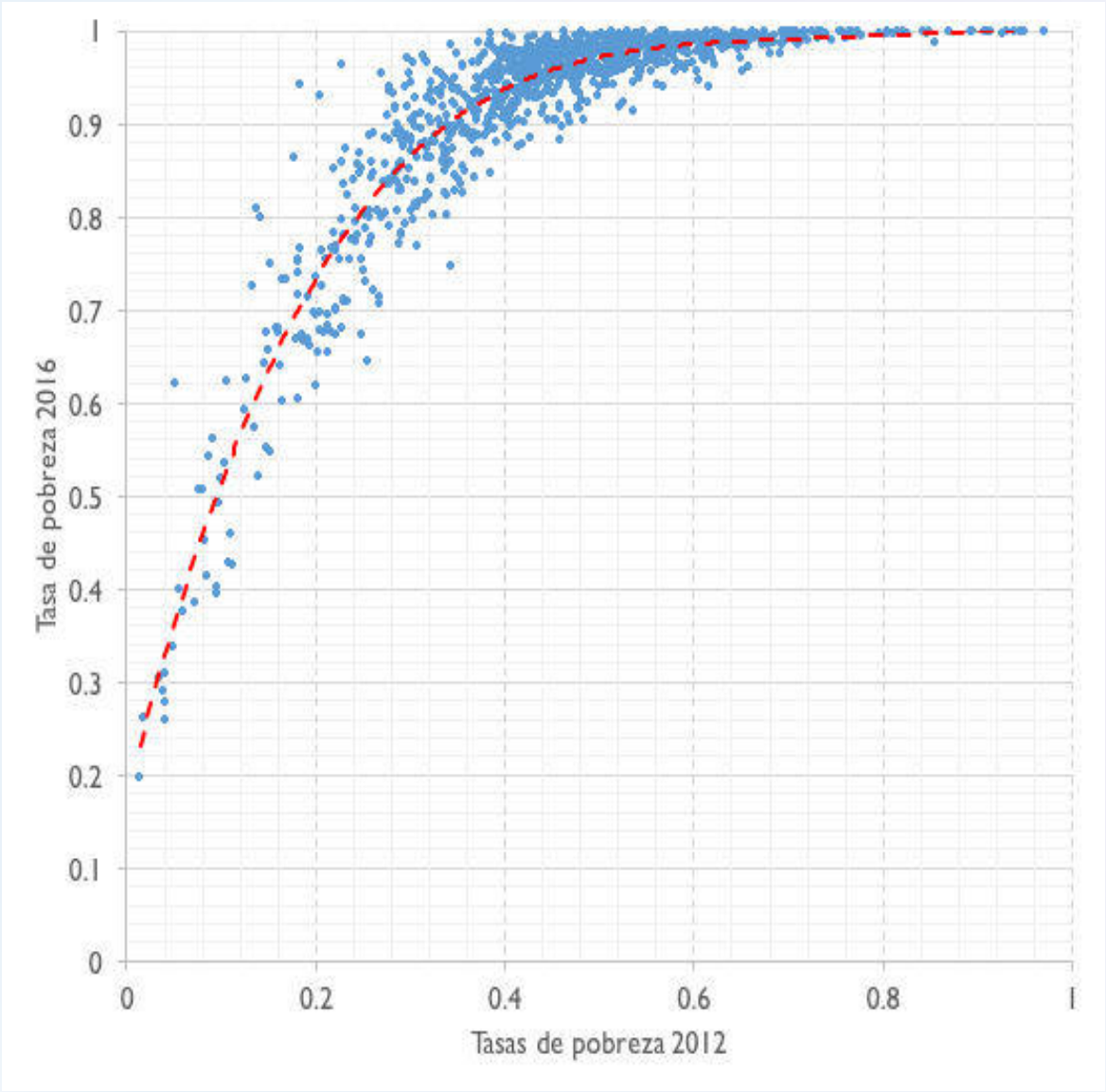
Dominancia absoluta de tasas de pobreza altas. Pobreza un fenómeno extendido y generalizado.

Tasa de pobreza de las 10 parroquias menos pobres: 29,9%.



19,8 100

Mapas de Pobreza 2012-2016

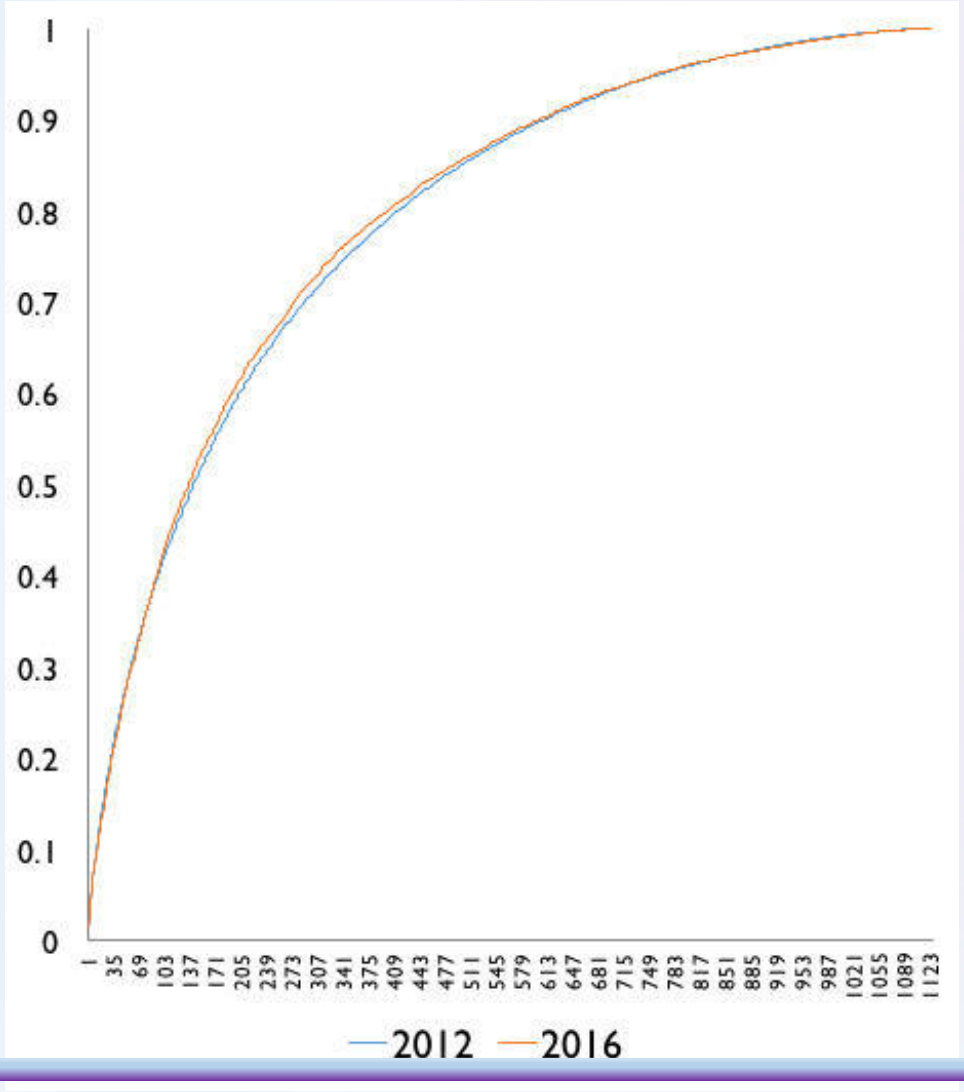


Empobrecimiento territorial generalizado, con un desplazamiento de toda la distribución hacia tasas altas de pobreza o pobreza total

93 parroquias con tasas de pobreza de 100%.

ANOVA
Evidence based thinking.

Mapas de Pobreza 2012-2016

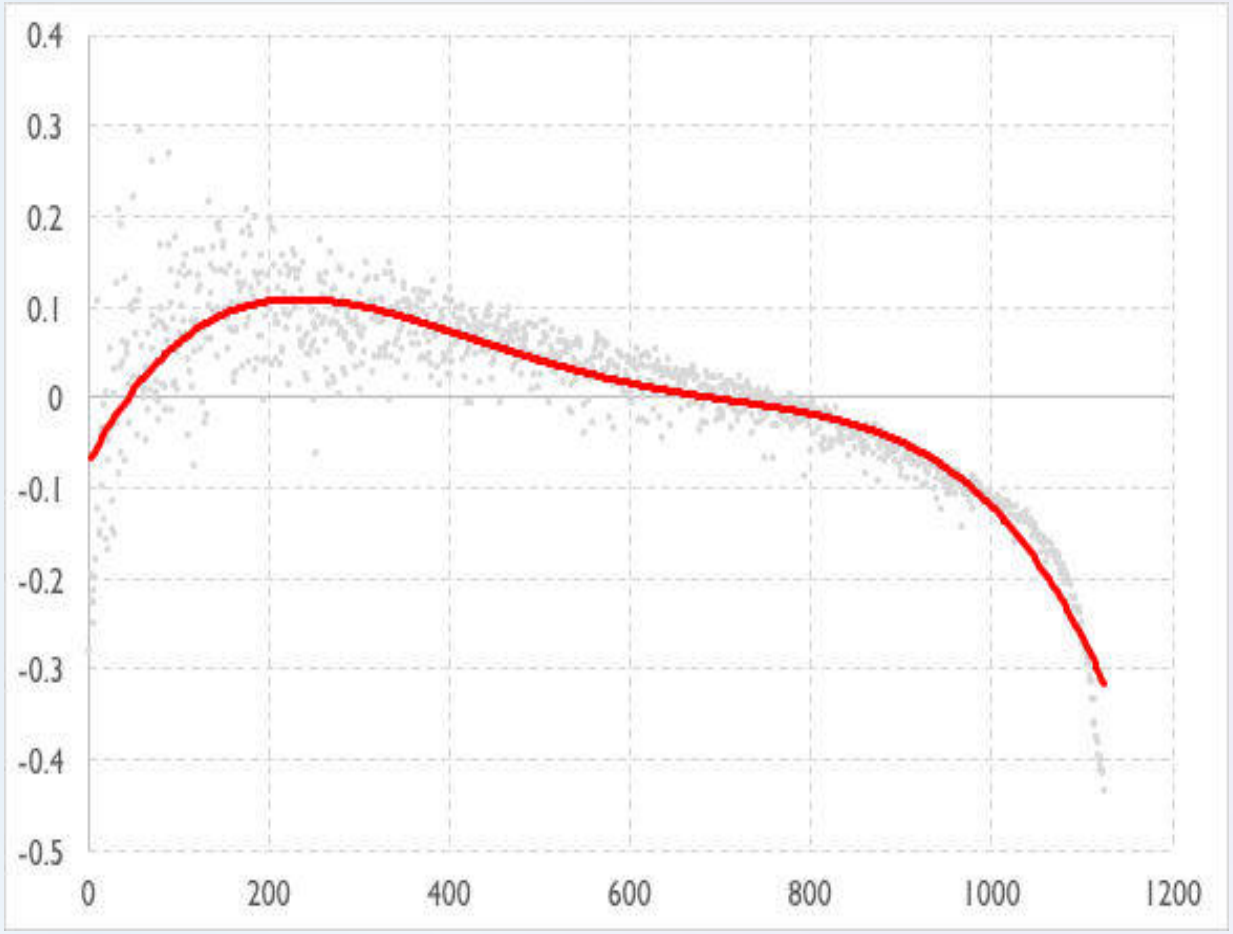


Leve (débil) tendencia a la concentración de pobres en conglomerados urbanos más densamente poblados.

Focalización territorial todavía es posible.

ANOVA
Evidence based thinking.

Mapas de Pobreza 2012-2016



Incidencia de mayor pobreza parroquial tuvo un sesgo anti-pobre a nivel parroquial.

Cambio en tasas de pobreza parroquial describe un U-invertida donde las parroquias con tasas relativamente más altas de pobreza inicial se deterioraron relativamente más.

Centros urbanos intermedios.

ANOVA
Evidence based thinking.

Mapas de Pobreza 2012-2016

	2012	2016
Tasa de Pobreza	0.362	0.814
Número de Pobres	9.95 millones	25.25 millones
Parroquias Pobreza > 80%	25	985
Parroquias Pobreza 60%-80%	164	104
Parroquias Pobreza 40%-60%	563	22
Parroquias Pobreza 20%-40%	311	13
Parroquias Pobreza < 20%	62	1
Pobres en 150 parroquias grandes	0.516	0.532
Pobres en 300 parroquias grandes	0.711	0.725

ANOVA
Evidence based thinking.

Lecciones/Consideraciones finales

- Mapa de pobreza permite una comprensión granular inédita de la dinámica socioeconómica de los países. Dimensión territorial.
- El uso de técnicas de Machine Learning es una alternativa viable para estimación de indicadores socioeconómicos a pequeña escala (SMA).
  - Implementación relativamente sencilla
  - Foco en la preparación de VSC
  - No recetas: distintas estimaciones requieren distintos modelos.
  - No solo pobreza, potencialmente todos los indicadores de EH.

ANOVA
Evidence based thinking.