Predicción efecto globo cultivos de cocaína periodo 2014-2020: Predicción de los cultivos de coca en Colombia después de intervenciones estatales

Luisa Fernanda Cuéllar, Isabella Riveros Usta y Daniel Mendivelso. Big Data and Machine Learning for applied economics. Universidad de los Andes. Bogotá 2022. Github del proyecto: <https://github.com/luifercue/Trabajo_final>

**Should not be longer than 8 (eight) pages (not including the title page with abstract, references, and appendix)**

**Resumen (200 palabras)**

En este trabajo se pretende predecir los movimientos de los cultivos de coca en Colombia en los años 2016 y 2020 con el fin de prevenir la aparición de nuevos cltivos, a través del desarrollo de un modelo de predicción bajo la metodología Logit, random forest y xgboots utilizando datos del Centro Nacional de Memoria Histórica en el periodo 2014-2020. En principio se utiliza una metodología de balanceo de la muestra para posteriormente someter los resultaods a un sistema de votación simple en la predicción final (Hirose y Zaman, 2009). El modelo sugiere que las zonas con mayor propensión a presentar cultivos después de cada intervención son las que están alrededor de XX km de los cultivos iniciales, las cuales en su mayoría son zonas asignadas a las comunidades indígenas o zonas con fácil acceso a vías terciarias, en particular en promedio los cultivos se mueven entre XX-XX kilómetros a la redonda. Adicionalmente, las zonas con mayor probabilidad de presentar cultivos son X, X, X, debido a que …

**Introducción**

Introduction. It should contain at least: the problem/research question clearly defined, antecedents, the added value (i.e., why your project is interesting, novel, and different), and a preview of results and takeaways.

La erradicación de coca puede ser un juego absurdo o de estrategia, lo que se ha observado en Colombia los últimos veinte años es un patrón en el cual se erradican los cultivos en un lugar y paulatinamente crece en otro.

**Datos**

Data. Treat this section as an opportunity to present a compelling narrative to justify and defend the data choices. Describe it accordingly with descriptive stats, graphs, etc.

Se procede a estudiar el NBI (necesidades básicas insatisfechas), las distancias a centros urbanos y mercados, el índice de ruralidad y los ingresos totales e ingresos totales per cápita del municipio. Los datos muestran que el promedio nacional de NBI3 de los municipios cocaleros es casi un 20% más alto que el de los no productores (vea la gráfica A en el apéndice). Se realizó el mismo ejercicio a nivel regional, debido a que las distintas variables socioeconómicas cambian de manera significativa entre regiones. Se encontró que los municipios cocaleros tienen para todos los casos excepto uno (región Amazónica) un NBI más alto (vea gráfica B)

La literatura ha encontrado que uno de los principales determinantes de los cultivos de coca es la marginalización de una región (Rocha 2010), a continuación se pueden observar las distancias entre los municipios y la capital del departamento respectivo, y el mercado mayorista de alimentos más cercano.

4 3 La medición del NBI va a ser el porcentaje de personas que incumplen con al menos una de las cinco dimensiones del indicador

4 Son distancias lineales tomadas desde los centroides de los municipios

Asimismo, se estudian los ingresos municipales totales para estudiar las diferencias económicas, además de una aproximación a la calidad de las instituciones económicas estatales, entre los cocaleros y los no productores.

presencia de grupos armados al margen de la ley (FARC, AUC, ELN) en los municipios donde hay cultivos ilícitos de coca

Datos trabajo Sebastián: Con el objetivo de encontrar regiones potencialmente cocaleras se utiliza un panel de datos a nivel municipal para los años 2010 a 2014, años para los cuales hay información disponible para todas las variables consideradas en este trabajo (en la sección 3.2 son explicadas). Los datos de hectáreas de cultivos ilícitos de coca provienen de la Oficina de Naciones Unidas contra la Droga y el Delito (UNODC por sus siglas en inglés), información que proviene del Sistema integrado de monitoreo de cultivos ilícitos (SIMCI). Para construir la base de datos con las características sociales, económicas, geográficas y de conflicto de cada región se utilizan las fuentes estadísticas que provee el panel municipal CEDE. Datos sobre erradicación de cultivos e incautaciones de coca, pasta y cocaína provienen del Ministerio de Justicia. Finalmente, los datos climáticos provienen del IDEAM.

**Modelo**

Model. Present and describe your preferred model specification. Argue why this is the best one for the task and describe it. Explain how it was trained, the hyper-parameters selection, and other relevant information. Be explicit about 1 how you define ”best.” If it is, by comparison, present the results of the different specifications and explain them briefly (details of non-preferred models can go into the appendix). In your comparison, be explicit about the metric chosen to compare among specifications.

**Resultados**

Here you present your results. Interpret these as an economist. When reporting the results is always a good idea to follow Nikolov’s (2022) advice: *A good rule to follow regarding reporting results is that less is usually better. Novice researchers (or graduate students) tend to over-include or report many secondary parameter estimates from most regression specifications. Although such a “kitchen sink” approach has some merits (e.g., it shows the audience the extensive analysis performed or that the researcher has examined various aspects regarding the stability of the results), this particular approach has a significant drawback. Pages of parameter estimates usually muddy the main message and* *story of the paper. It can significantly detract from the comprehensiveness of your central contribution, which is most important in your paper.*

The reader will get either lost, bored, or annoyed. – Present only results and parameter estimates that directly bolster your main takeaway and story. –Do not present secondary analyses or relegate such analyses to an Online Appendix if needed. I understand that a semester is a short time to have a full paper, so preliminary results are acceptable.

Esta investigación se aproxima al problema planteado a través de un análisis supervisado5 con el fin de hallar patrones y pronósticos sobre las regiones susceptibles a las consecuencias del Efecto Globo. La variable dependiente de este modelo es una variable dicótoma que toma el valor de 1 si el municipio hizo la transición a convertirse en cocalero, 0 en caso contrario, esto significa que es un problema de clasificación. De manera general, la metodología mencionada utiliza algoritmos que se encargan de enseñarle a la “máquina” como se caracteriza una observación que toma el valor de 1 y una observación con valor de 0 para después aplicar lo aprendido por el modelo creado a nuevas observaciones y poder clasificarlas (predecirlas) de manera correcta.

Un municipio potencialmente cocalero se define en este trabajo como aquel que no produce cultivos ilícitos de coca pero que tiene la combinación de características y factores que permiten la proliferación de los mismos. Para encontrar estos municipios se aglomeró el panel de datos a un corte transversal, es decir que las observaciones se convirtieron a nivel de municipio y no de año-municipio. Para lograr lo anterior se promediaron los valores de las variables entre el 2010 y el 2013 por municipio (se utilizan los datos del 2014 como base de prueba al ser la información más actualizada y completa posible). Posteriormente, se marcaron con un uno a los que cambiaron de estado, es decir los que empezaron a producir de un año a otro. Ya que el objetivo es encontrar los municipios potencialmente cocaleros se eliminan de la base los municipios que fueron productores de coca entre el 2010 y 2013, en particular esto reduce el ruido en los modelos de aprendizaje ya que los municipios potencialmente cocaleros solo pueden ser elegidos del conjunto de municipios que no producen coca. El proceso anterior reduce 185 municipios, los cuales produjeron constantemente durante los 4 años, de un total de 1,122. De esta nueva base de 937 municipios, 69 son los que cambiaron a ser productores de coca en algún momento del periodo de estudio. Esta base es usada para entrenar el modelo de clasificación.

Esta misma base también tiene un problema que genera sesgo en los clasificadores, la variable dependiente está fuertemente desbalanceada, esto significa que hay muy pocos positivos con relación a los negativos o viceversa. Para este caso de estudio hay muy pocos positivos, es decir 12 que el número de observaciones que pasaron por la transición de transformarse en productores de coca en el lapso del estudio son muy pocas en relación con el número de municipios que no hicieron la transición. Lo anterior hace que los modelos se sesguen a pronosticar dentro del grupo mayormente representado. Para corregir este problema se utiliza subbagging (Antonini, Elisseef y Paleologo, 2010), metodología que será explicada en la subsección 3.3.

**Conclusiones y recomendaciones**

Conclusions and recommendations. In this section, you state the main takeaways of your work.

**Referencias**

Moreno, S. (2018). “El Efecto Globo: identificación de regiones propensas a la producción y grupos de coca”. CESED. Universidad de los Andes. Bogotá. Colombia. Recuperado de; <https://repositorio.uniandes.edu.co/bitstream/handle/1992/41027/dcede2018-28.pdf?sequence=1&isAllowed=y>

Raffo, L. Castro, J. y Díaz, A. (2015). “Los efectos globo en los cultivos de coca en la Región Andina (1990-2009)”. Universidad del Valle. Cali. Colombia. Recuperado de: <https://revistas.uptc.edu.co/index.php/cenes/article/view/3426/5198>

Revista 070. “Inteligencia artificial para ganarle a la coca”. Junio 2018. Recuperado de: <https://cerosetenta.uniandes.edu.co/28378-coca-inteligencia-artificial/>

**Anexos**