Mapas interactivos: una herramienta para el análisis exploratorio de datos ético

Interactive maps: a tool for ethical exploratory data analysis

Cristina Soto-Rojas[1]

Soto-Rojas, C. Mapas interactivos: una herramienta para el análisis exploratorio de datos ético. *Tecnología en Marcha*. Vol. 35, especial V Encuentro Bienal Centroamericano y del Caribe de Investigación y Posgrado. Junio, 2022. Pág 24-31.

https://doi.org/10.18845/tm.v35i6.6232

1 Maestría en Computación, Instituto Tecnológico de Costa Rica. Asistente de Investifación en el Centro Nacional de Alta Tecnología, San José, Costa Rica; Correo electrónico: csoto@cenat.ac.cr
https://orcid.org/0000-0001-9180-1628

Palabras clave

Mapas interactivos; ética; análisis exploratorio de datos; leaflet; R.

Resumen

Al iniciar el análisis exploratorio de datos de un nuevo proyecto, por lo general revisamos la estructura de las variables, la distribución, varianza, realizamos pruebas de normalidad, pero ¿estamos revisando si son éticos?. No siempre se corrobora que los datos no caigan en algún tipo de discriminación, y muchas veces tampoco es fácil verificar esto. Se propone el uso de mapas interactivos para la verificación de posibles sesgos geoespaciales en los datos, como un agregado al análisis exploratorio de los datos, para evaluar ese aspecto ético.

Keywords

Interactve maps; ethic; exploratory data analysis; leaflet; R.

Abstract

When starting the exploratory data analysis of a new project, we usually review the structure of the variables, the distribution, variance, we perform normality tests, but are we checking if they are ethical? It is not always corroborated that the data does not fall into some kind of discrimination, and many times it is not easy to verify this either. The use of interactive maps is proposed to verify possible geospatial biases in the data, as an addition to the exploratory analysis of the data, to evaluate this ethical aspect.

Introducción

El análisis exploratorio de datos (AED) es una técnica estadística utilizada en la primera parte del trabajo estadístico con un conjunto de datos nuevo [1]. Consiste en explorar los datos, analizando la estructura que tienen las variables, indagando la distribución que pueden seguir variables numéricas, buscando posibles patrones, entre otras técnicas que dependen del tipo de datos. No existe una guía puntual establecida, sin embargo, la gran mayoría permiten conocer las variables, observar posibles patrones existentes en los datos y así establecer hipótesis que luego se confirmarán o rechazarán con los métodos estadísticos aplicados a los datos posterior al AED.

Esta técnica es bastante popular y utilizada como fase inicial de investigación estadística, desde modelos sencillos, hasta modelos de inteligencia artificial. Es decir, este método es la base de muchos de los trabajos de investigación desarrollados hoy en día, que tienen impacto directo en nuestra sociedad. Por lo general, dentro de este AED se evalúa si los datos tienen algún tipo de inconsistencia, datos faltantes o alguna variable atípica, pues los modelos estadísticos pueden ser sensibles a este tipo de situaciones. Incluso se han creado herramientas como JENGA-A, por Schelter et al. [2] que permite evaluar el impacto de estas situaciones en los modelos, pues no siempre es posible controlar todos los errores en los datos, especialmente cuando hablamos de modelos en tiempo real.

Sin embargo, es importante ahora evaluar si con ese mismo afán en todos los casos se busca controlar que los datos sean éticos. Así como los errores en los datos van a tener un impacto directo en los resultados del modelo, los sesgos en los datos van a tener un impacto directo también. No son casos aislados en los que se ha reportado que un modelo tiene sesgo, por

ejemplo, [3] nos habla sobre el impacto de datos con un sesgo racial por años de historia con discriminación racial y el impacto que esto tiene en los modelos que ahora usen esos datos para la ayuda con toma de decisiones, decisiones que perpetúan esos sesgos.

Veamos ahora cómo nos pueden ayudar los mapas interactivos. Roth [4] realiza un amplio análisis sobre los mapas interactivos, enfocado a los fundamentos de la cartografía y la interactividad que tienen los mapas con los usuarios, recordando todo lo que los mapas pueden proveer a los usuarios que los utilizan. Luego, Andrienko et al. [5] nos expone un ejemplo del uso de mapas interactivos para el análisis de variables geoespaciales y cómo estas herramientas abren nuevas posibilidades de exploración de datos, y nos permiten ir más allá. Los mapas interactivos han sido muy desarrollados, pues son múltiples las aplicaciones que los utilizan día a día, por ejemplo, Waze, Google Maps, Uber, entre otras.

Ahora, más allá de esos usos, Adrienko [6] nos comenta sobre el uso para el análisis exploratorio, específicamente a través de una herramienta llamada DESCARTES, que buscaba ayudar al usuario, indiferentemente de si era cartógrafo o no, a analizar datos con referencias geoespaciales, para poder utilizar esto en el AED. Se le brindaba al usuario interactividad para que pudiera así explorar y establecer sus dudas de investigación o patrones. Podemos ver un ejemplo en la figura 1.

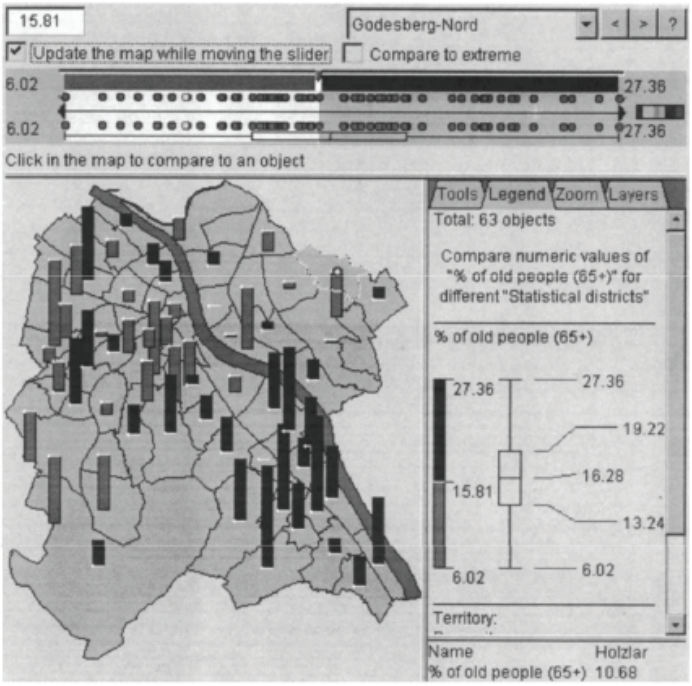

Figura 1. *Ejemplo de uso de la Interfaz DESCARTES.*

Note. Ejemplo de aplicación DESCARTES para ver las edades en los distintos distritos de Bonn. Tomado de *Interactive maps for visual data exploration*, por [6].

Considerando esta herramienta que puede ser de gran utilidad, analicemos cómo puede apoyarnos en el análisis ético que deseamos hacer. Como se mencionó anteriormente, existe múltiples ejemplos de diferentes sesgos que han sido evidenciados en muchos modelos estadísticos utilizados para facilitar la toma de decisiones. Existen sesgos raciales, de género, de clases, entre otros y el origen de ellos son los datos que son utilizados para entrenar a estos modelos.

Saiph [7] y las demás expositoras mencionan la importancia de considerar la evaluación ética de los modelos estadísticos en todas las etapas, desde el AED, hasta la etapa final, pues en cada una de ellas existe el riesgo de caer en sesgos. Es claro que no podemos eliminar por completo el sesgo en los datos y los modelos, pues el mundo sigue siendo un mundo lleno de sesgos, pero podemos intentar minimizarlo.

Y los mapas interactivos pueden darnos una perspectiva geoespacial que no siempre es tan evidente si revisamos los datos crudos o con otro tipo de visualizaciones, pues un mapa interactivo nos permitirá realizar ciertas comparaciones y evaluaciones entre las variables numéricas o categóricas de nuestros datos y la ubicación geoespacial de esos resultados, lo cual puede ayudarnos a identificar patrones que evidencien algún sesgo.

Metodología

A continuación se presenta una propuesta del uso de un mapa interactivo para el análisis exploratorio ético de los datos. El caso propuesto se realizó utilizando dos capas geoespaciales:

- Capa Casas Culturales: esta capa es del Ministerio de Cultura y Juventud de Costa Rica y provee la localización geográfica de las casas culturales en el país.
- Capa Índice de bienestar de niños y adolescentes: esta capa es del Instituto Nacional de Estadística y Censos (INEC) y muestra el índice de bienestar de niños y adolescentes a nivel distrital.

Ambas capas fueron extraídas del Sistema Nacional de Información Territorial [8] de Costa Rica. El objetivo de analizar ambas capas es evaluar si existe algún tipo de sesgo en la ubicación de las casas culturales y si realmente analizar los rendimientos que tienen estas casas permite determinar el aporte que hacen a la población de niños y adolescentes de forma equitativa.

Para la implementación de este gráfico se utilizó la herramienta R, que permite la elaboración de mapas interactivos con ayuda del paquete Leaflet [9]. Se le dio tratamiento a ambas capas ajustando la proyección geográfica para que calzara con la base de OpenStreet que Leaflet permite utilizar. La capa del índice de bienestar era una capa de polígonos que contenía la información del índice para cada cantón. Mientras que la capa de las casas culturales era una capa de puntos que presentaba información extensa sobre los contactos, miembros y ubicación de cada casa cultural.

Posterior a esto se utilizaron herramientas de Leaflet para fusionar ambas capas. Se seleccionó una gama de colores naranjas para representar el índice, de forma que entre más claro se veía el cantón, más bajo era el índice, para que esto fuera observable con la capa verde claro de OpenStreet y para resaltar una región se escogió un tono azulado para contrastar al naranja. Para los puntos se seleccionó un tono vino que fuera un poco complementario al naranja, pero notorio y permitiera contrastar con la capa verde de OpenStreet.

Resultados y discusión

A continuación la visualización del mapa interactivo obtenida.

Figura 2. *Visualización propuesta.* Nota: Visualización inicial al usuario del mapa interactivo.

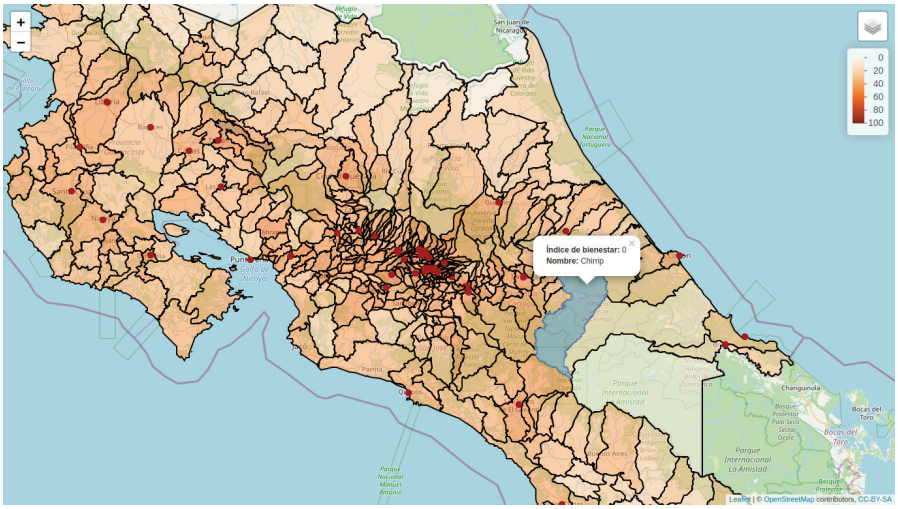

Figura 3. *Visualización de un cantón. Nota:* Selección de un cantón y muestra de información asociada.

Figura 4. *Visualización de una casa cultural. Nota*: Selección de una casa cultural y muestra de información asociada.

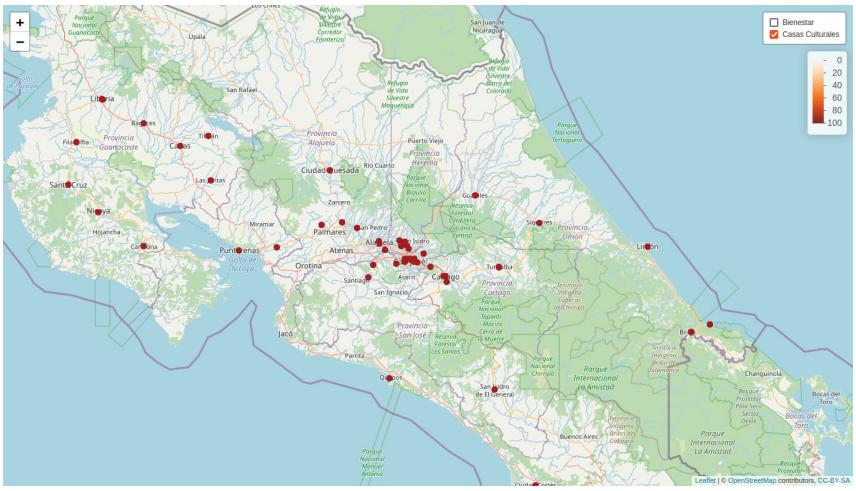

Figura 5. *Visualización de capas individuales. Nota:* Selección de la capa de casas culturales.

Figura 6. *Visualización de capas individuales. Nota*: Selección de la capa de índice de bienestar individual.

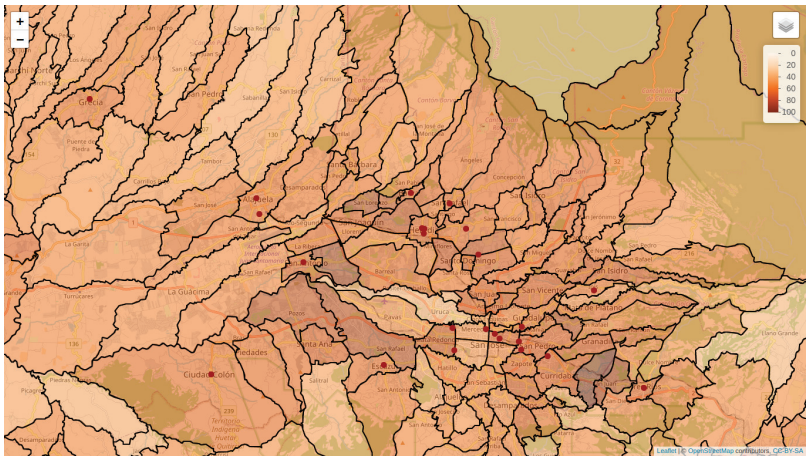

Figura 7. *Uso de herramienta de zoom. Nota:* Posibilidad de zoom sobre el mapa.

Como podemos notar en las visualizaciones, las casas culturales se encuentran bastante concentradas en la zona central del país, y en zonas más claras, es decir con índice de bienestar menor, como en la zona norte, podemos observar prácticamente nula presencia de casas culturales. Entonces, con solo esta visualización podemos empezar a notar un patrón geoespacial que nos muestra que puede existir un sesgo en la ubicación de estas casas culturales.

Entonces, si por ejemplo, analizamos el impacto de las casas culturales en el índice de bienestar de niños y adolescentes, no podemos asegurar con certeza que ese impacto es igual en todas las regiones del país. Más bien esta visualización abre la puerta a hacer propuestas sobre una distribución más equitativa de las mismas, es decir, de llegar a las manos indicadas, esta visualización nos ayudaría a promover eliminar un sesgo, que tal vez no estaba claro directamente de los datos crudos.

El análisis geoespacial de los datos nos da una visión de los datos que es difícil de obtener a través de otras visualizaciones, nos permite analizar variables sobre los datos acompañadas de la dimensión espacial, que puede llegar a transmitir esa información que deseamos compartir diferente. Además, los mapas interactivos permiten también modelar variables temporales y ver la evolución temporal del mapa. Es decir, podemos en una misma visualización evaluar tiempo, espacio y otras variables de nuestro interés de forma clara y amigable con el usuario. Es una herramienta que puede aportar mucho valor al AED.

Es importante recordar, que no siempre una crítica de sesgos es bien recibida, existen muchas empresas o entidades que quieren que estos sesgos se sigan promoviendo, o simplemente no les interesan, entonces es importante conocer que esta realidad existe, no para desanimarnos, si no para motivarnos a hacer nuestro trabajo de la forma más ética posible dentro de nuestras capacidades.

Consideraciones que podemos tener para velar por esto son: indagar sobre el origen de nuestros datos y si fueron tomados, por ejemplo, a poblaciones variadas; compartir el origen de nuestros datos si generamos datos abiertos y la forma en que fueron recolectados; evaluar el alcance que tienen nuestras aplicaciones generadas y si están alcanzando todas las poblaciones deseadas o solo algunas. Finalmente, velar porque el producto que estamos generando realmente tiene como propósito mejorar la calidad de vida humana.

Conclusiones

El análisis exploratorio de datos es una herramienta utilizada en muchísimas investigaciones estadísticas, y es de gran utilidad para detectar patrones, conocer los datos con los que se trabajará y establecer preguntas de investigación, es una parte que se aconseja seguir replicando. Sin embargo, se recomienda ampliar el análisis también al área ética.

Una de las opciones recomendada para ese AED ético, cuando los datos lo permitan, es uno acompañado de un mapa interactivo, cruzando las diferentes capas, para analizar si en los datos existe algún sesgo que no se está viendo de forma clara. Este tipo de análisis nos permite visualizar la información de una forma diferente y explorar según nuestro interés, por lo que provee más información para poder detectar estos sesgos.

En un futuro se planea explorar la creación de una herramienta que ya contenga diversas capas geoespaciales con metadatos generales, en la que el usuario pueda cargar su capa , cruzarla con otras capas y explorar, para que así pueda realizar un AED geoespacial y ético de una forma más amigable. Además se desea que contenga ejemplos como el presentado para ilustrar la importancia de este tipo de análisis.

Referencias

[1] J. T. Behrens, (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2(2), 131.

[2] S. Schelter, T. Rukat, & F. Biessmann. (2021). JENGA-A Framework to Study the Impact of Data Errors on the Predictions of Machine Learning Models. In EDBT (pp. 529-534).

[3] J.J. Avery. (2019). An uneasy dance with data: Racial bias in criminal law. S. Cal. L. Rev. Postscript, 93, 28.

[4] R.E. Roth. (2013). Interactive maps: What we know and what we need to know. Journal of Spatial Information Science, 2013(6), 59-115.

[5] G. Andrienko, N. Andrienko, & V. Gitis. (2003). Interactive maps for visual exploration of grid and vector geo-data. ISPRS Journal of Photogrammetry and Remote Sensing, 57(5-6), 380-389.

[6] G. Andrienko, & N. Andrienko. (1999). Interactive maps for visual data exploration. International Journal of Geographical Information Science, 13(4), 355-374.

[7] Ética & Inteligencia Artificial. Taller RIIAA 2020. https://www.youtube.com/watch?v=_erf6xVFRG4

[8] SNIT. (20 de abril de 2021) https://www.snitcr.go.cr/

[9] Leaflet (1 de abril de 2021) https://rstudio.github.io/leaflet/