R E V I S T A
DERECHO ELECTORAL

SEGUNDO SEMESTRE **2023** NÚMERO 36

ISSN: 1659-2069

Utilización de ciencias de datos en análisis de resultados electorales: un ejemplo aplicado a los resultados de la segunda ronda electoral del 2018 en Costa Rica

Francisco José Mora Cordero*

https://doi.org/10.35242/RDE_2023_36_3

Nota del Consejo Editorial

Recepción: 3 de mayo de 2023.

Revisión, corrección y aprobación: 3 de julio de 2023.

Resumen: La ciencia de datos nace de la combinación de enfoques matemáticos, estadísticos y computacionales, cuyo objetivo es procesar datos mediante la aplicación de algoritmos y, de esta forma, obtener información difícil, o imposible, de lograr por otros medios. En este artículo ejemplificamos la aplicación de ciencias de datos, específicamente del algoritmo K-means a datos relacionados con los resultados de la segunda ronda electoral en Costa Rica del año 2018 en conjunción con datos sobre el desarrollo social a nivel de distritos en Costa Rica en el año 2017. Asimismo, se presentan algunas alternativas para la visualización de la información resultante.

Palabras clave: Procesamiento de datos / Resultados electorales / Balotaje / Elecciones presidenciales / Geografía electoral / Distribución de electores / Desarrollo social / Desarrollo humano.

Abstract: Data science is born from the combination of mathematical, statistical, and computational approaches, whose objective is to process data through the application of algorithms and thus obtain information difficult or impossible to achieve by other means. In this article, we exemplify the application of data science, specifically the K-means algorithm, to data related to the results of the 2018 runoff of Costa Rica in conjunction with data on social development at the district level in Costa Rica in 2017. The author presents some alternatives to visualize the resulting information.

Key Words: Data processing / Electoral results / Balloting / Presidential elections / Electoral geography / Electoral distribution / Social development /Human development.

* Costarricense, informático, correo fmorac@tse.go.cr. Graduado del Instituto Tecnológico de Costa Rica, con una licenciatura en calidad de software de la Universidad Nacional Estatal a Distancia y estudios en la maestría académica en ciencias de la computación del Instituto Tecnológico de Costa Rica. Cursó los programas de experto en ciencias de datos y de formación en minería de datos, de Promidat. Actualmente labora como administrador de proyectos en la Oficina de Proyectos Tecnológicos del Tribunal Supremo de Elecciones.,.

ISSN: 1659-2069

1. INTRODUCCIÓN

Chavarría (2022) en su artículo "Una mirada cantonal mediante estadística espacial al efecto del desarrollo humano sobre el apoyo electoral en la segunda ronda de la elección presidencial de 2018" señala la importancia de la distribución geográfica en el apoyo a los dos partidos participantes en la segunda ronda en las elecciones del año 2018, la cual analiza por medio de la aplicación de un modelo de autorregresión espacial, considerando los datos de los resultados electorales y del índice de desarrollo humano de cada cantón.

Siguiendo el enfoque de dicho artículo, planteamos un tratamiento alternativo, en el cual se aplican conceptos de ciencias de datos, particularmente el agrupamiento (*clustering*) y la visualización de datos, considerando tanto los resultados electorales de la segunda ronda del 2018 como datos relacionados con los índices de desarrollo de las unidades geográficas.

El objetivo de este artículo no es analizar las causas de los resultados obtenidos, lo cual corresponde al campo de las ciencias sociales, sino mostrar la aplicación de la ciencia de datos en el análisis de fenómenos relacionados con los procesos electorales.

Con el fin de proveer información para un análisis más detallado, los datos se toman a nivel de distrito y no a nivel de cantón, lo cual permitiría al investigador establecer hipótesis de interés sobre unidades geográficas más pequeñas y, posiblemente, homogéneas. El procesamiento de los datos se realizó en el lenguaje R[1] en el entorno de programación R Studio[2].

2. CIENCIA DE DATOS

La ciencia de datos nace de la conjunción de la estadística, las matemáticas y las ciencias computacionales, y tiene como finalidad obtener, a partir de datos en bruto, comprensión, visión y conocimiento (Wickman y Garret, 2016).

De esta amalgama de campos del conocimiento surgen diversos métodos para el tratamiento de los datos (Hastie, Tibshirani y Friedman, 2009), que han demostrado ser de gran utilidad en múltiples campos, entre otras aplicaciones podemos citar:

[1] R es un entorno y lenguaje de programación con un enfoque al análisis estadístico (https://www.r-project.org/).
[2] R Studio es un entorno de desarrollo integrado para el lenguaje de programación R, dedicado a la computación estadística y gráficos (https://www.rstudio.com/).

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

- Predecir si una persona que sufrió un ataque cardíaco sufrirá un segundo ataque.
- Predecir el precio de acciones en la bolsa de valores.
- Identificar letras y dígitos manuscritos.
- Estimar la cantidad de glucosa en la sangre de una persona, a partir de la absorción del espectro infrarrojo de su sangre.
- Identificar el riesgo de sufrir cáncer de próstata a partir de la información clínica y datos demográficos de una persona.
- Clasificar el correo electrónico como *spam* o no *spam*.
- Reconocimiento facial y de imágenes.
- Reconocimiento de voz.
- Identificar las opiniones positivas o negativas sobre un producto o servicio.
- Determinar la temática de una publicación.
- Establecer si una transacción con tarjeta de crédito es fraudulenta.

Una gran cantidad de las aplicaciones citadas interactúan con nosotros casi diariamente y de forma transparente, ya sea en reconocimiento de voz en los teléfonos celulares, en las recomendaciones para compras en los sitios de comercio electrónico, en dispositivos médicos, al comprar con una tarjeta de crédito o al solicitar un préstamo en una entidad financiera.

3. MÉTODO DE AGRUPAMIENTO K-MEANS

Uno de los enfoques utilizados en la ciencia de datos es el de agrupamiento (*clustering*). Según Mirkin (2005), el método de agrupamiento consiste en definir una métrica de similitud entre las entidades bajo consideración, agrupando aquellas similares en un grupo y manteniendo las disímiles en otros grupos[3].

[3] El problema de encontrar la partición óptima de un conjunto de datos n-dimensionales en *K* grupos es de índole combinatoria, lo que hace que sea impráctico intentarlo; por lo anterior, los algoritmos utilizados buscan hallar soluciones *suficientemente buenas*, que se aproximen al óptimo.

ISSN: 1659-2069

Cada entidad cuenta con una cantidad *n* de atributos[4], los cuales se expresan numéricamente, de forma que cada entidad se modela como un punto en un espacio *n*-dimensional. Un agrupamiento es la partición de ese espacio en *K* grupos que cumpla con las siguientes condiciones: (a) cada entidad pertenece a un grupo (y solo a uno) y (b) todos los miembros de un grupo están más cerca del centroide[5] del grupo que del centroide de cualquier otro grupo.

Uno de los algoritmos diseñados para agrupar datos es el *K-means* el cual requiere que se indique un valor para el parámetro *K*, que determina la cantidad de grupos que se desea obtener. A partir del valor *K*, el algoritmo selecciona (generalmente de manera aleatoria) *K* puntos como centroides iniciales, y selecciona los puntos más cercanos a cada uno. Posteriormente, utilizando dichos puntos calcula un nuevo centroide; este proceso se repite hasta que el proceso converja a *K* conjuntos estables (es decir, hasta que los puntos de cada conjunto no varíen entre una interación y otra) o hasta que se alcance un número máximo, preestablecido, de iteraciones.

Debido a que la selección inicial de los centroides puede incidir en la calidad de los grupos encontrados, se suele repetir el proceso varias veces, seleccionando distintos centroides iniciales, y seleccionando el que presente una mejor calidad en términos de la inercia intraclases e interclases[6]. Aunque el algoritmo encontrará siempre la mejor partición[7], sí brinda aproximaciones suficientemente buenas.

En cuanto al valor *K*, este puede seleccionarse de forma arbitraria, pero para lograr que la cantidad de clústeres asegure la mejor calidad, se recomienda realizar un experimento para determinar cuál valor de *K* presenta mayor calidad en los resultados.

4. DATOS

Una de las fuentes de datos utilizadas es la de los resultados electorales de la segunda ronda del 2018. Dicha información fue tomada del cuadro 3.3 del *Cómputo de votos y declaratorias de elección febrero y abril de 2018:*

[4] Un atributo puede verse como una variable en términos informáticos o estadísticos.
[5] El centroide de un conjunto de puntos es el punto que minimiza la suma de las distancias a cada punto.
[6] La inercia intraclase es una medida de homogeneidad de los elementos de la clase, al maximizar la inercia intraclase de cada clase, se minimiza la inercia interclases.
[7] La única forma de asegurar encontrar la mejor partición es por medio de la fuerza bruta, lo cual requiere probar todas las opciones de agrupamiento y seleccionar la mejor; sin embargo, este es un problema combinatorio, imposible de resolver en términos prácticos.

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

presidente, vicepresidencias y diputados a la Asamblea Legislativa (TSE, 2018), en el cual se muestran los siguientes datos: provincia, cantón, distrito, cantidad de juntas, electorado, el total general de votos, el porcentaje de participación, el total de votos válidos, total de votos para el partido Acción Ciudadana (PAC), total de votos para el partido Restauración Nacional (RN), total de votos en blanco, total de votos nulos, abstencionismo absoluto y abstencionismo porcentual.

La otra fuente de datos es el índice de desarrollo social del año 2017 (Ministerio de Planificación Nacional y Política Económica [Mideplan], 2018), del cual se tomó como base la tabla del anexo 3. Los datos contenidos en la tabla mencionada son los siguientes: código de distrito, distrito, y las dimensiones económicas, salud, educativa, seguridad y de participación electoral; a partir de las cuales el Mideplan calcula el índice de desarrollo social de cada distrito.

Para el procesamiento de los datos, nos inclinamos por no usar los valores absolutos de votación y por emplear valores porcentuales: el porcentaje de abstencionismo; y calculamos el porcentaje de votos obtenido por el PAC y el RN[8]. Con estos dos porcentajes, calculamos la diferencia entre el porcentaje del PAC y el porcentaje de RN (esto debido a que el PAC fue el partido ganador de la segunda ronda). El objetivo de estos cálculos es usar valores relativos que faciliten la comparación entre distritos, independientemente de su población y de su extensión.

Por último, para la visualización de los mapas, se utilizaron los datos geográficos de los distritos, provistos por Programa Iberoamericano de Formación en Minería de Datos (Promidat) en formato *shapefile*[9], logrando así combinar los datos geográficos con los demás datos distritales.

[8] Estos porcentajes se calculan sumando los votos del PAC y los votos del RN, y dividiendo cada una de esas cantidades por el resultado de la suma; dejamos de lado los votos nulos y los votos en blanco, ya que no representan una cantidad apreciable.

[9] *Shapefile* es un formato sencillo y no topológico que se utiliza para almacenar la ubicación geométrica y la información de atributos de las entidades geográficas Este archivo en este formato nos fue entregado por la empresa Promidat (https://promidat.website), como parte del curso de Visualización de Datos. Ellos a su vez obtuvieron dichos datos del sitio del Sistema Nacional de Información Territorial (https://www.snitcr.go.cr/).

ISSN: 1659-2069

5. RESULTADOS

Una vez cargados y preprocesados, los datos resultantes se almacenaron en un *dataframe*[10], cuyas columnas se muestran en las tablas 1 y 2, junto con sus distribuciones en cuartiles, valores menores y mayores y sus promedios.

La **¡Error! La autoreferencia al marcador no es válida.**, con base en los datos obtenidos del índice de desarrollo social del año 2017; y la Tabla 2

Descripción de las variables de los resultados electorales por distrito presenta los porcentajes de votos del PAC, de RN, de abstencionismo y la diferencia entre los porcentajes del PAC y RN.

No se muestra en las tablas 1 y 2 la existencia del código del distrito, ya que este se toma como un identificador y no como un dato propiamente dicho.

Tabla 1

Descripción de las variables del índice de desarrollo social por distrito

Económica	Participación	Salud	Educación	Seguridad	IDS2017
Min.: 0.00	Min.: 0.00	Min.: 0.00	Min.: 0.00	Min.: 0.00	Min.: 0.00
1st Qu.: 18.09	1st Qu.: 46.36	1st Qu.: 62.90	1st Qu.: 48.84	1st Qu.: 87.36	1st Qu.: 52.03
Median: 27.86	Median: 55.15	Median: 70.84	Median: 62.01	Median: 93.20	Median: 63.03
Mean: 31.44	Mean: 56.36	Mean: 69.63	Mean: 58.92	Mean: 90.50	Mean: 62.79
3rd Qu.: 41.41	3rd Qu.: 65.34	3rd Qu.: 78.65	3rd Qu.: 70.67	3rd Qu.: 97.14	3rd Qu.: 72.98
Max.: 100.00	Max.: 100.00	Max.: 100.00	Max.: 100.00	Max.: 100.00	Max.: 100.00

Nota: Min. es el valor mínimo que toma la variable, 1st Qu. es el valor del primer cuartil, median es le mediana, mean es el promedio, 3rd Qu. es el tercer cuartil y Max. el valor máximo.

[10] Un *dataframe* es una estructura de almacenamiento de datos, similar a una tabla, que permite realizar diversas operaciones sobre los mismos datos.

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

Tabla 2

Descripción de las variables de los resultados electorales por distrito

PorcPAC	PorcRN	PorcAbs	DifPorc
Min.:18.84	Min.:10.03	Min.:18.23	Min.: -60.778
1st Qu.:48.33	1st Qu.:28.91	1st Qu.:27.42	1st Qu.:-2.303
Median:61.80	Median:37.20	Median:32.83	Median:24.490
Mean:58.85	Mean:40.00	Mean:34.45	Mean:18.845
3rd Qu.:69.96	3rd Qu.:50.42	3rd Qu.:41.17	3rd Qu.:41.212
Max.:88.67	Max.:79.61	Max.:59.50	Max.:78.610

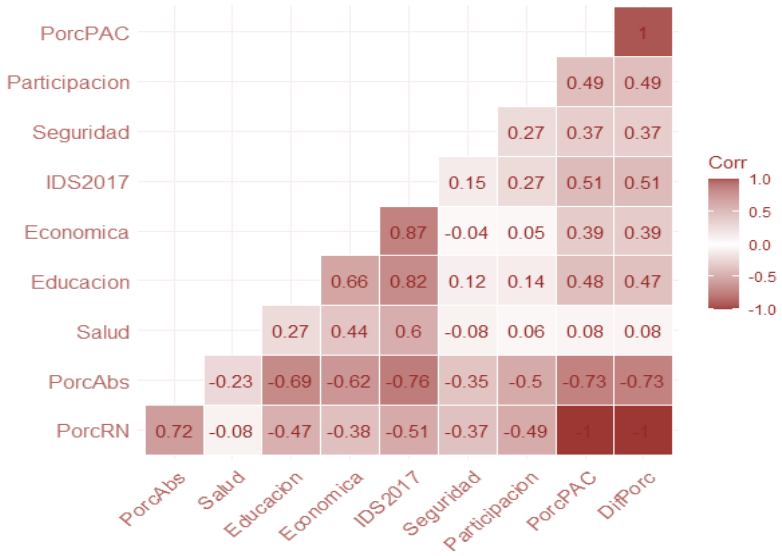
Además de la información estadística mostrada en las tablas 1 y 2, podemos visualizar las correlaciones existentes entre las variables[11], tal como se muestra en la figura 1.

[11] En este caso se utilizó una correlación de Spearman es una medida no paramétrica de la correlación de rango (dependencia estadística del *ranking* entre dos variables).

ISSN: 1659-2069

Figura 1

Correlación entre las variables



La figura 1 muestra, de manera muy clara, las correlaciones existentes entre las variables, por ejemplo, la correlación negativa entre el índice de desarrollo social y el porcentaje de votos para RN y la correlación positiva entre el índice de desarrollo social y el porcentaje de votos para PAC. Resalta, también, la correlación positiva entre el porcentaje de abstencionismo y el porcentaje de votos para RN.

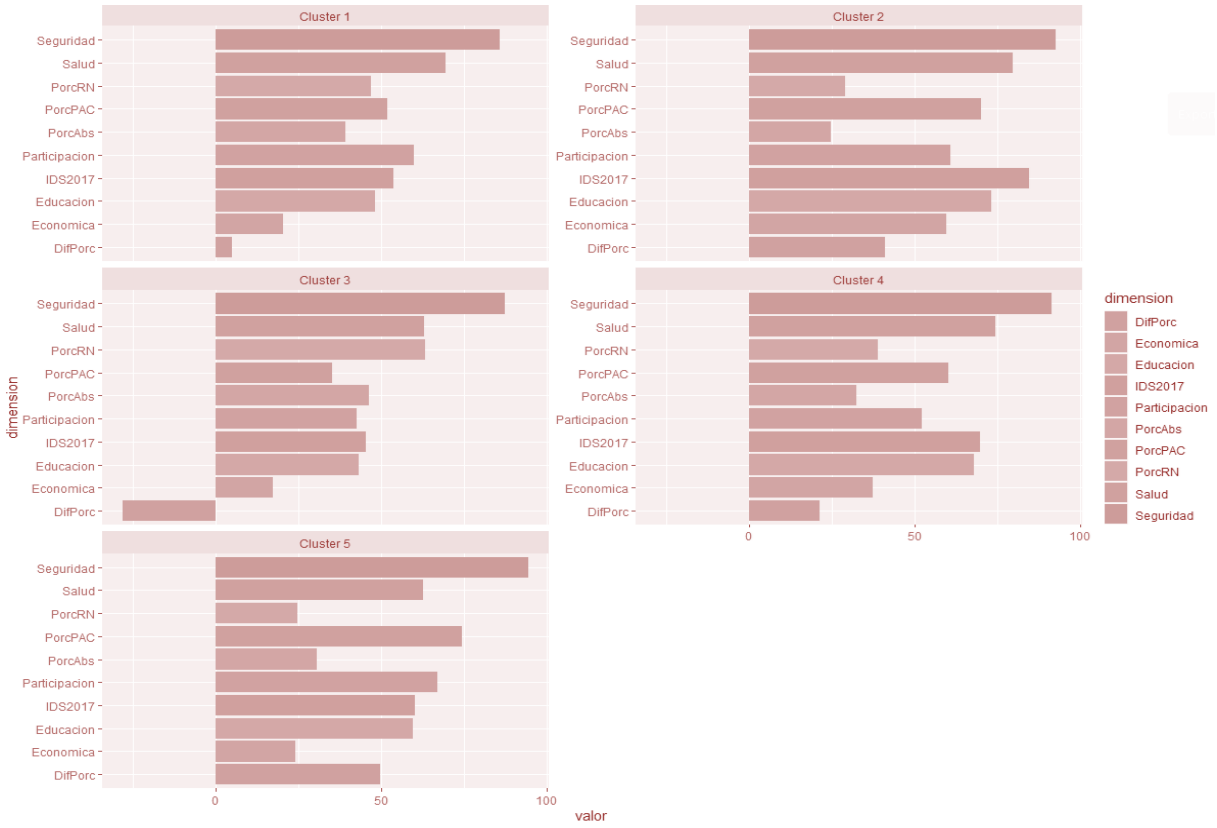
Considerando los datos e información obtenidos hasta el momento, se agrupó a los individuos (distritos) en 5 grupos (clústeres), por medio del método *K-means*, y se obtuvo una partición de los datos[12] con las características que se muestran en la figura 2.

[12] Los parámetros utilizados para la selección de los grupos fueron: K= 5, iteraciones= 500, ejecuciones= 500.

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

Figura 2

Valores de las variables del centroide de cada grupo



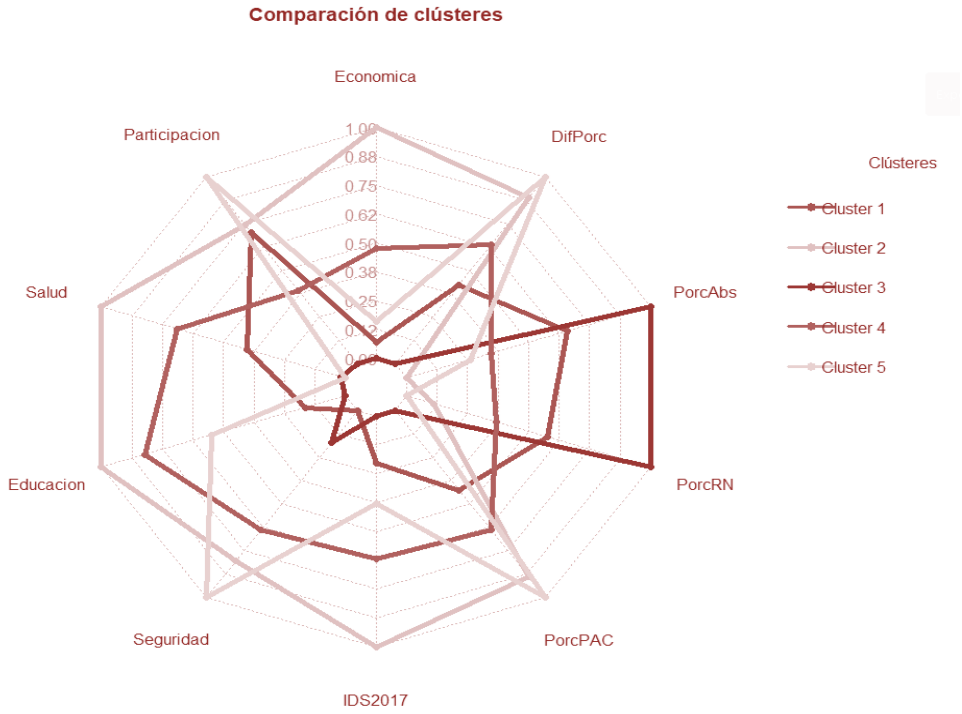
Se puede notar que en el grupo 3 se agrupan los distritos en los que RN obtuvo más votos que el PAC y los que, a su vez, tienen el menor índice de desarrollo social. En el grupo 2 se juntan los distritos en los que la diferencia porcentual de votos a favor del PAC es de las mayores y posee los mayores índices de desarrollo social. Observaciones de naturaleza similar pueden realizarse para los demás grupos y dimensiones (variables).

Otra forma de visualizar las características de los grupos es por medio de un gráfico de radar, tal como se presenta en la figura 3.

ISSN: 1659-2069

Figura 3

Distribución de los valores de los centroides de los grupos



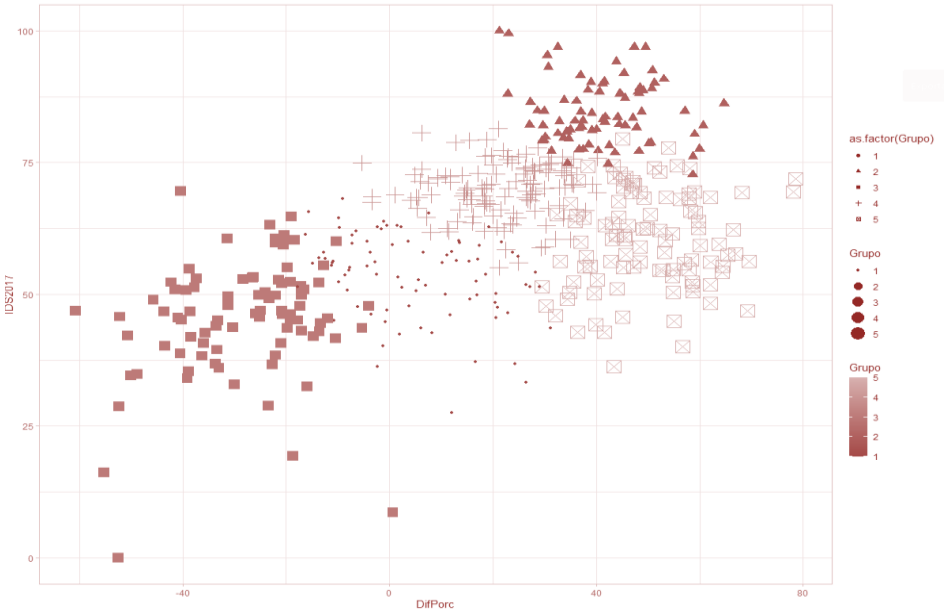
Por medio de la figura 3, se visualiza más claramente que el grupo 3 tiene los índices de desarrollo más bajos, con los porcentajes para RN y abstencionismo más altos. Igualmente, el grupo 2 se caracteriza por tener buenos índices de desarrollo humano (con el mejor índice general), un alto porcentaje de votos a favor del PAC y un bajo porcentaje de abstencionismo. El grupo 5 aparenta ser un caso interesante, ya que presenta una diferencia porcentual muy alta a favor del PAC, pero con índices de desarrollo no muy buenos.

La polarización de los grupos, sobre todo del 2 y el 3 se puede notar más claramente en la figura 4, en la que se grafican los distritos, utilizando como eje *x* la diferencia porcentual (% PAC - % RN) y como eje "*y*" el índice de desarrollo social.

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

Figura 4

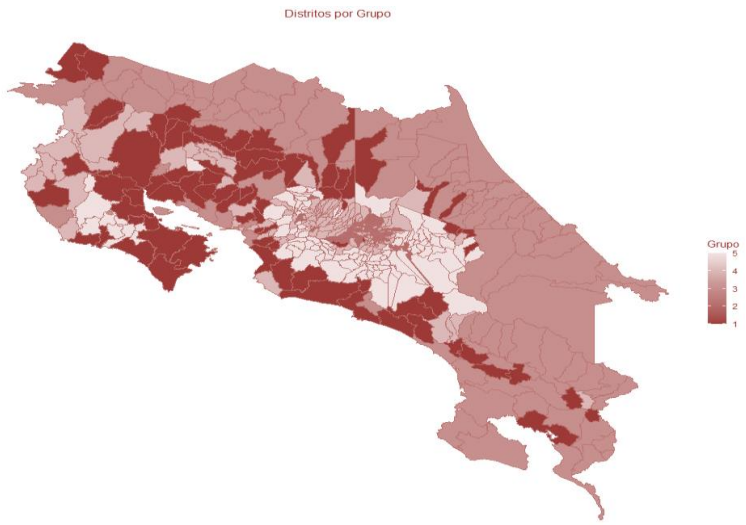
Distribución de los distritos, según índice de desarrollo y diferencia porcentual de votos (PAC - RN)



Habiendo caracterizado cada grupo, podemos visualizar su distribución geográfica en la figura 5.

Figura 5

Distribución geográfica de los grupos

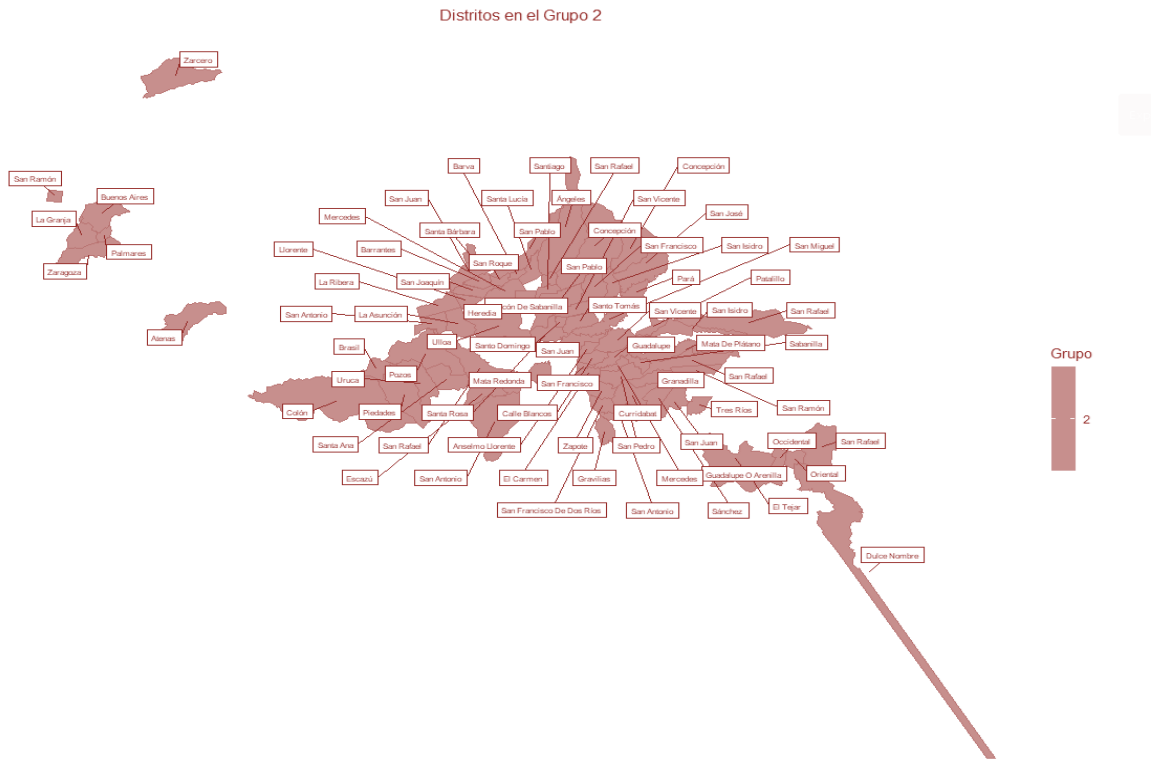


ISSN: 1659-2069

Si deseamos conocer los nombres de los distritos de cada grupo, podemos usar gráficos etiquetados que son subconjuntos del gráfico mostrado en la figura 5. Por ejemplo, para el grupo 2 tenemos los distritos que se presentan en la figura 6.

Figura 6

Distritos del grupo 2

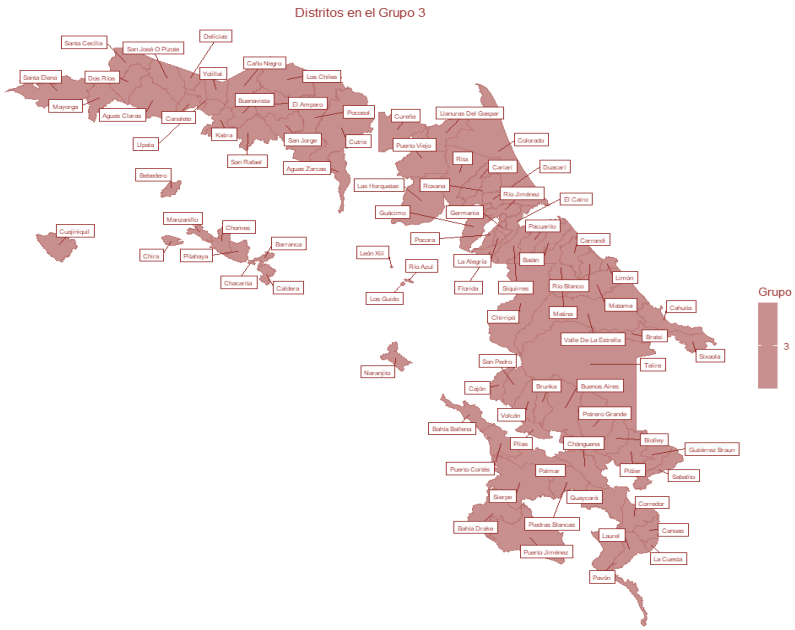


Y para el grupo 3, la figura 7 presenta los resultados.

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

Figura 7

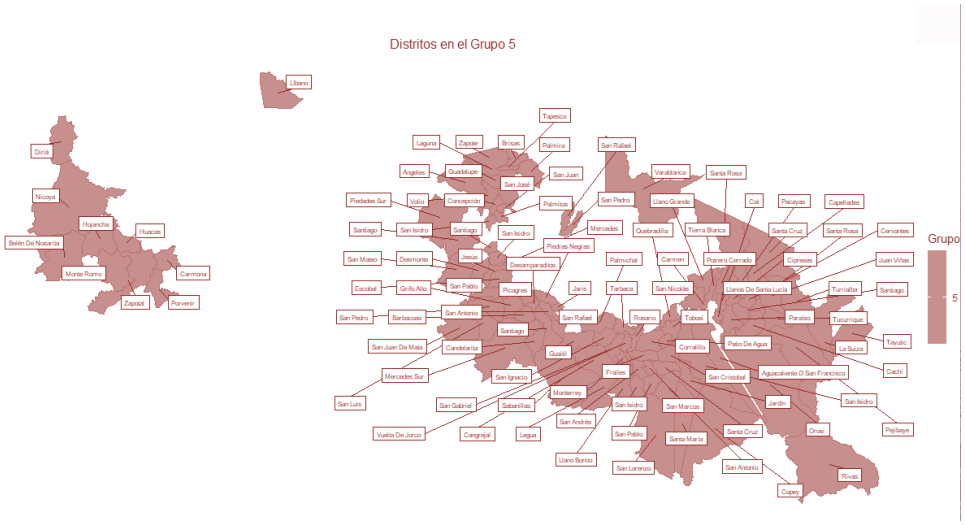
Distritos del grupo 3



Otro grupo interesante es el número 5, cuyos resultados se pueden observar en la figura 8.

Figura 8

Distritos del grupo 5



ISSN: 1659-2069

6. CONCLUSIÓN

El tratamiento de los datos, por medio de técnicas y métodos de ciencias de datos, permite a los investigadores y analistas detectar patrones que, de otra forma, podrían permanecer ocultos o no serían tan evidentes.

En el caso de los resultados electorales, hacer uso de los datos públicos provistos por el Tribunal Supremo de Elecciones, y complementarlos con datos provenientes de otras instituciones (en nuestro caso el MEIC e Instituto Geográfico Nacional) u otras fuentes, nos brinda dimensiones adicionales que enriquecen las posibilidades de análisis.

Adicionalmente, contar con diversas opciones para visualizar los datos, y las relaciones existentes entre ellos, puede llevar a los investigadores a plantear hipótesis o descubrir nuevas líneas de investigación.

Debido a las limitaciones propias de una publicación en una revista, presentamos únicamente gráficos estáticos que no permiten la interacción con un usuario. En el entorno de R Studio es posible crear gráficos interactivos, así como aplicaciones que brindan al usuario la libertad de parametrizar los cálculos y las visualizaciones por medio de las facilidades de publicación como R-Markdown[13] o Shiny[14].

Para finalizar, el enfoque presentado no posee características predictivas, pero podría servir como un punto de arranque para la búsqueda de modelos predictivos que permitan responder preguntas asociadas a elecciones futuras, por ejemplo, acerca de la polarización electoral y del comportamiento del abstencionismo.

[13] https://rmarkdown.rstudio.com/
[14] https://shiny.rstudio.com/

SEGUNDO SEMESTRE 2023 NÚMERO 36
ISSN: 1659-2069

REFERENCIAS

Chavarría M, E. (Ene.-jun., 2022). Una mirada cantonal mediante estadística espacial al efecto del desarrollo humano sobre el apoyo electoral en la segunda ronda de la elección presidencial de 2018. *Revista de Derecho Electoral*, (33), 81-90.

Ministerio de Planificación y Política Económica de Costa Rica, Área de Análisis del Desarrollo (2018). *Índice de desarrollo social 2017*. MIDEPLAN.

Tribunal Supremo de Elecciones de Costa Rica (2018). *Cómputo de votos y declaratorias de elección febrero y abril de 2018: presidente, vicepresidencias y diputados a la Asamblea Legislativa*. Tribunal Supremo de Elecciones.

Hastie, T., Tibshirami, R., Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference and Prediction* (Second Edition). Springer-Verlag.

Mirkin, B. (2005). *Clustering for Data Mining, A Data Recovery Approach*. Chapman & Hall/CRC.

Wickman, H. y Grolemund, G. (2017). *R for Data Science*. O'Reilly Media Inc.