Investigaciones Latinoamericanas en Ingeniería y Arquitectura

Revista ILIA N°2, 2025, pp: 45-52
URL: https://revistas.uca.edu.sv/index.php/ilia/index
ISSN: 3078-2740 / e-ISSN: 3078-2759 / ISSN-L: 3078-2740

Análisis clúster con datos satelitales y sociodemográficos para clasificar el territorio salvadoreño

Cluster Analysis with Satellite and Sociodemographic Data to Classify the Territory of El Salvador

DOI https://doi.org/10.51378/ilia.vi2.9659

F. Carranza[1], M. Aguilar[2]

[1]Maestría en Estadística Aplicada a la Investigación, Universidad Centroamericana José Simeón Cañas (UCA), El Salvador.

[2]Departamento de Organización del Espacio

E-mail: fcarranza@protonmail.com

Email: maguilar@uca.edu.sv

Fecha de recepción: 15 de mayo de 2025 Fecha de aprobación: 1 de noviembre de 2025

Resumen

Este estudio explora la posibilidad de subdividir el área rural de El Salvador en grupos de municipios, donde cada uno tenga sus propias características en cuanto a las variables PIB per cápita, consumo eléctrico per cápita, densidad poblacional, tasa de pobreza y luz nocturna. El estudio se hizo de manera horizontal considerando la distribución espacial de la luz. La luz se obtuvo con el procesamiento de imágenes satelitales en software de Sistemas de Información Geográfica (SIG). Con base en la naturaleza de los datos se optó por la aplicación de técnicas estadísticas avanzadas de *clustering* que apoyadas en las ventajas de la computación permitieran comparar 1000 posibilidades de *clusters* cambiando los parámetros de clasificación como el método y la distancia utilizados. El estudio concluye que, a nivel exploratorio, la subdivisión con técnica de *cluster* jerárquico es posible solo al incorporar la luz nocturna y técnicas avanzadas con t-SNE. Se encontró que el mejor modelo de subterritorios se agrupa en nueve categorías donde dos grupos son principalmente municipios con predominancia urbana y el resto con predominancia rural.

Palabras clave — SIG, teledetección, aprendizaje de máquina, agrupamiento, indicadores socioeconómicos.

Abstract

This study explores whether the rural area of El Salvador can be subdivided into groups of municipalities where each group has its own characteristics in terms of variables GDP per capita, electricity consumption per capita, population density, poverty rate and night light. The study was performed horizontally considering the spatial distribution of light. The light was obtained by processing satellite images with Geographic Information Systems (GIS) software. Based on the nature of the data, it was decided to apply advanced statistical clustering techniques that, supported by the advantages of computing, would allow comparing 1000 cluster possibilities by changing the classification parameters such as the method and the distance used. The study concludes that at the exploratory level, subdivision with hierarchical cluster technique is possible only by incorporating night light and advanced techniques with t-SNE. It was found that the best model of subterritories is grouped into nine categories where two groups are mainly municipalities with urban predominance and the rest with rural predominance.

Keyword — GIS, remote sensing, clustering, machine learning, socioeconomic indicators.

I. INTRODUCCIÓN

El Salvador se caracteriza por tener una minoría de municipios con predominancia de zona urbana mientras que la mayoría de los municipios presenta una predominancia de zona rural. El estudio tiene un objetivo, que es identificar agrupamientos de municipios (no necesariamente adyacentes) con características propias de tal forma que el área rural se pueda subdividir y que así se puedan diseñar políticas públicas que consideren territorios más focalizados. A nivel internacional diversos estudios han logrado agrupar municipios según sus características (Ej.: fiscales,

datos de movilidad, etc.). Un estudio previo nacional desarrollado por la Universidad Centroamericana José Simeón Cañas (UCA) y la Universidad Rafael Landívar identificó agrupamientos de municipios que llamaron territorios funcionales y que consisten en territorios subnacionales que funcionan como unidades funcionales o espacios autocontenidos desde una perspectiva económica, social y ambiental [1].

Para lograr el objetivo, se hizo un análisis estadístico riguroso usando técnicas de *clustering* (agrupamiento) con el objetivo de identificar categorías de agrupamientos de municipios. Finalmente, y con el objetivo de conocer la distribución geográfica de los municipios, se crearon visualizaciones de los municipios con software de Sistemas de Información Geográfica (SIG), lo cual permitió verificar su sentido sociodemográfico.

II. METODOLOGÍA

La investigación consistió en un estudio exploratorio de tipo cuantitativo donde se aplicaron técnicas de aprendizaje de máquina no supervisado conocidas como *clustering* a las siguientes variables: luz nocturna (obtenida mediante teledetección y SIG) y cuatro variables sociodemográficas y económicas: Producto Interno Bruto (PIB) per cápita, consumo eléctrico per cápita, tasa de pobreza y densidad poblacional, todas obtenidas de datos oficiales. En la Figura 1 se ilustra la metodología de esta investigación de manera general. La investigación fue no experimental, ya que los datos no se manipularon y se obtuvieron para el mismo año, 2007, al ser el año del último censo oficial al momento del estudio.

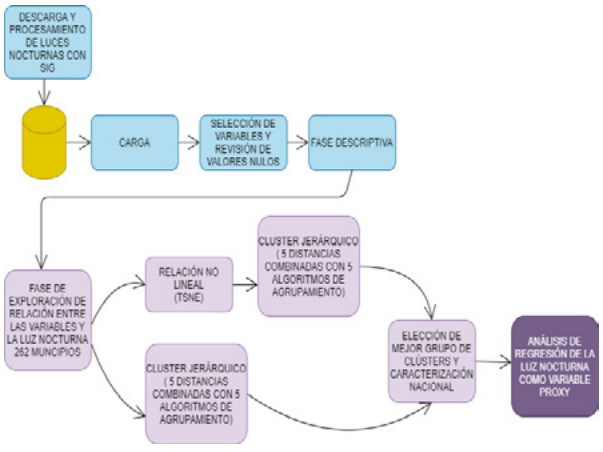

Fig. 1. Metodología general. Elaboración propia.

En el análisis de *clúster*, el método jerárquico agrupa a los individuos de forma jerárquica. Este puede tener como entrada pre-transformaciones como, por ejemplo: las dimensiones reducidas de los métodos PCA (*Principal Component Analysis*) o t-SNE (*t-distributed stochastic neighbor embedding*). Para el caso de las distribuciones que son asimétricas y que pueden afectar los agrupamientos [2] se pueden utilizar transformaciones. t-SNE, que es una herramienta utilizada para la reducción de dimensiones en especial en relaciones no lineales, que permite la generación de *clusters* de buena calidad [3]; el uso adecuado de los parámetros es necesario para evitar sesgos [4] al momento de su aplicación. Las técnicas de *clustering* que se aplicaron en este estudio fueron tres, todas aplicando *clustering* jerárquico: 1) Sin transformación logarítmica ni reducción de dimensiones, 2) Sin transformación logarítmica y aplicando reducción de dimensiones t-SNE y 3) Con transformación logarítmica y aplicando reducción de dimensiones t-SNE. Además, se trabajaron diferentes combinaciones cambiando el método de clasificación y el tipo de distancia. Los resultados de las tres técnicas se compararon aplicando los controles de calidad que se describen en el análisis estadístico exploratorio.

A. Recopilación de datos y software

Los datos sociodemográficos fueron obtenidos de una base de datos procesada por el departamento de Economía de la UCA a partir de datos del censo 2007. En la Fig. 2 se observan las otras fuentes de datos para este estudio. Todos los cálculos estadísticos, procesamiento de datos y visualizaciones con mapa fueron realizados con software libre y de código abierto. El análisis estadístico fue realizado con *software* R y los procesos de SIG con la versión estable QGIS 3.16.5 LTS Hannover.

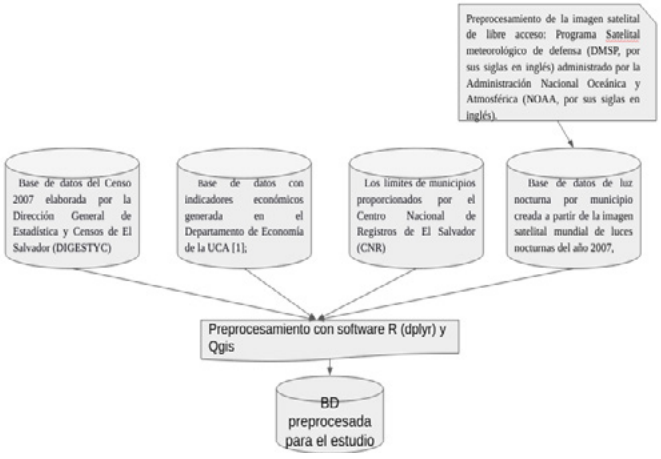

Fig. 2. Datos de entrada y preprocesamiento de datos. Elaboración propia.

B. Sistemas de Información Geográfica

El *software* de SIG se utilizó para tres tareas específicas: 1) el procesamiento de la imagen satelital para la obtención de la luz nocturna por municipio, 2) el mapeo de cada una de las variables de estudio por clases tanto en cuartiles como por rupturas naturales y 3) el mapeo de los agrupamientos de municipios para validar los resultados de los agrupamientos estadísticos y para confirmar el método de agrupamiento y la distancia utilizada.

Con relación al uso de la luz nocturna como *proxy* para indicadores sociodemográficos a nivel internacional, estudios en la última década concluyen que a nivel de distrito el PIB efectivamente se explica significativamente por las luces nocturnas en el área [5] y [6]. Uno de estos estudios encontró además que la no linealidad es mucho más fuerte para las ciudades metropolitanas donde los niveles de PIB son mucho más altos que en un modelo lineal. Por el contrario, en las zonas donde las actividades agrícolas y forestales son más altas, el uso de luces nocturnas en un modelo lineal sobreestima el PIB [7]. A nivel regional se ha realizado investigación de estos métodos indirectos en países como República Dominicana [8] y México [9] donde se ha comprobado su viabilidad en el primer caso y en el segundo se comprobó que las cifras oficiales muestran un crecimiento menor al obtenido por el método alternativo. A nivel nacional, este es el primer estudio que investiga con esta metodología, la viabilidad de utilizar la luz nocturna como proxy para estimar variables sociodemográficas en el contexto de El Salvador.

La imagen satelital de luces nocturnas es una imagen donde cada píxel representa aproximadamente 1 km2 y tiene un valor numérico en un rango de 0-63, que es un número entero positivo asignado a la respuesta de un sensor en relación con la intensidad de la señal recibida por el sensor. El número 0 representa la oscuridad completa y 63 es el valor para zonas más iluminadas. Para realizar el cálculo de la mediana de luz municipal se realizó en primer lugar un preprocesamiento de la imagen satelital. En segundo lugar, se realizó una intersección de este resultado con los límites municipales, con lo que se obtuvo para cada municipio cuadros completos con un valor de luz o fragmentos de estos, según la forma del municipio y su respectiva área. Con apoyo del *software* SIG, se visualizó cada uno de los resultados de agrupamientos de municipios para validar el significado sociodemográfico de los mismos según su distribución espacial.

C. Análisis estadístico exploratorio con R

Haciendo uso del método no supervisado de *clusterización*, se analizaron las distancias "*Euclidean", "maximum", "Manhattan", "Canberra"* y *"Minkowski"* y los algoritmos de agrupamiento siguientes: *"single", "complete", "average", "centroid"* y *"Ward"*. Se obtuvieron gráficas de siluetas de los *cluster*. Se procedió en dos vías de análisis (Fig. 1). Con el software R se procedió con los siguientes pasos:

1. Agrupación de los municipios mediante las variables: luz nocturna, PIB per cápita, consumo eléctrico per cápita, tasa de pobreza y densidad poblacional mediante clúster jerárquico.
2. Agrupación mediante una fase previa de reducción no lineal en dos dimensiones mediante t-SNE. Luego dichos puntos encontrados sirvieron de entrada para el clúster jerárquico. Para la selección de las mejores salidas de t-SNE, se generaron 50 ejecuciones con cada una de las distancias variando el parámetro de *perplexity* entre el rango de 5 a 50, theta de 0.0 y se hicieron dos ejecuciones una fue con 10,000 iteraciones máximas y la otra con 50,000 iteraciones máximas. Para la prueba de convergencia la selección fue visual, basada en la mejor visualización de los grupos en el plano xy, tomando en cuenta que lograra converger a lo largo de las 45 iteraciones de perplejidad. Además, se hicieron pruebas separadas con conversión y sin conversión logarítmica de las variables para saber si las distribuciones afectaban la formación de los *clusters*.

Los criterios de selección para el mejor resultado se basaron en las siguientes comparativas:

- Estadística: correlación cofenética, distancia de Gower, el número de *clúster* sugeridos por el gráfico de nivel de fusión, el resultado del gráfico de silueta. Además, se verificó el número de miembros mal agrupados.
- Criterio sociodemográfico: que el número de clúster tuviera sentido económico-sociodemográfico y estuviera acorde a las 5 variables de estudio con comprobación cartográfica.

Luego se analizaron los grupos mediante la prueba de Kruskal-Wallis y la prueba de Yuen para casos de valores atípicos. En todas las pruebas de hipótesis de la investigación se utilizó una significancia

de 0.05. Posteriormente, se procedió a caracterizar cada *cluster*, proporcionándole un sentido de criterio sociodemográfico a partir de las características de tendencia central, con base a la mediana y usos de diagramas de cajas y bigotes.

El alto valor de coeficiente de correlación cofenético [10] es uno de los indicadores para medir el cambio de la perturbación entre las distancias iniciales y las distancias finales con las que se agrupan los individuos en el *cluster*. El menor valor de la distancia de Gower es otro indicador para elegir el mejor modelo de aglomeración, esta medida proviene de la suma cuadrática de las diferencias entre las distancias de los datos iniciales y las distancias cofenéticas [11]. El gráfico de silueta permite validar la cercanía entre los individuos y sus vecinos en el *cluster*, a su vez permite identificar los miembros mal clasificados, para calcular y comparar contra otros cortes de *cluster* la distancia promedio de la silueta es un buen indicador [12].

III. RESULTADOS Y DISCUSIÓN

A. Resultados descriptivos

En las Fig. 3 y 4 se visualizan los mapas de cuartiles para cada variable sociodemográfica de estudio y la luz nocturna. En las Fig. 5 y 6 se visualizan los mapas de rupturas naturales (Jenks) para las mismas variables.

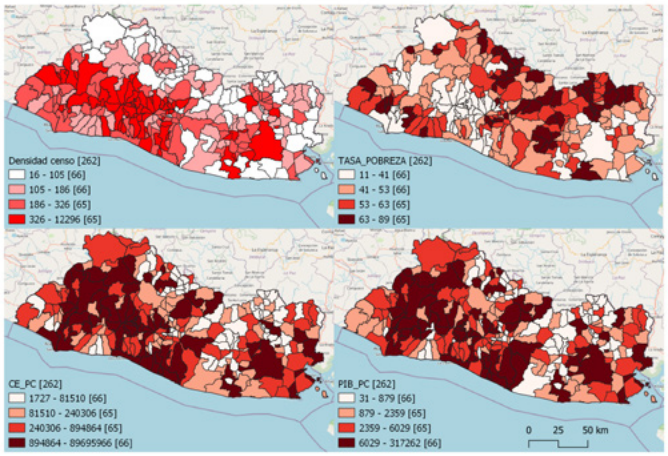

Fig. 3. Mapas de cuartiles para cada variable de estudio. Elaboración propia con mapa base de Openstreetmap.

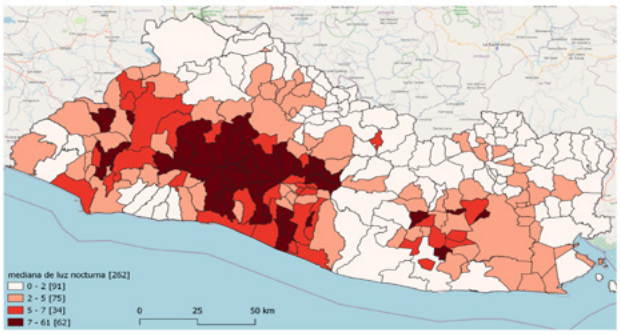

Fig. 4. Mapa de cuartiles para la variable mediana de luz nocturna, 2007. Elaboración propia con mapa base de Openstreetmap.

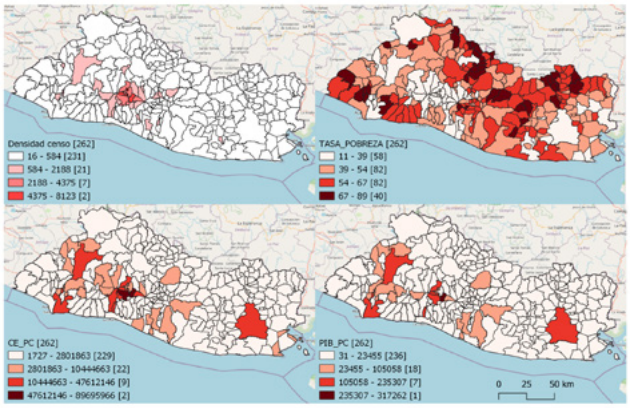

Fig. 5. Mapas de rupturas naturales (Jenks) para cada variable de estudio. Elaboración propia con mapa base de OpenStreetMap.

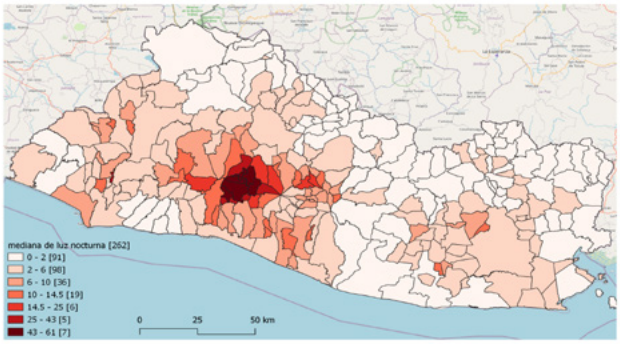

Fig. 6. Mapa de rupturas naturales (Jenks) para la variable mediana de luz nocturna, 2007. Elaboración propia con mapa base de Openstreetmap.

Como se puede observar en los mapas, el método Jenks permite visualizar clases que se agrupan mejor internamente y que se diferencian más entre sí cuando las distribuciones no son normales.

B. Resultados exploratorios

Los resultados del proceso de *clustering* sin t-SNE se muestran a continuación en las tablas 1 y 2 y las Fig. 7 y 8.

Tabla 1.: Clúster Jerárquico. Elaboración propia.

Metodología Opción lineal	Total de posibles clusters	Resultado
5 tipos de distancia 5 tipos de algoritmos 10 Cortes por combinación	250	El inconveniente encontrado con esta agrupación fue que no era posible seguir subdividiendo el área rural, por lo que se procedió a la generación de clusters jerárquicos a partir de un proceso previo de reducción de dimensiones de forma no lineal mediante el método de t-SNE teniendo como entrada la reducción automática PCA.

Tabla 2.: Clúster Jerárquico. Elaboración propia

Correlación Cofenética	Distancia Gower	Clúster Sugeridos	Observación
7.951.346	5602703	4-8 (Figura 8)	Agrupó mal algunos municipios (Figura 7) Mayores divisiones a 5 no produjeron separación de la zona rural La agrupación mostró consistencia a nivel nacional rural-urbano

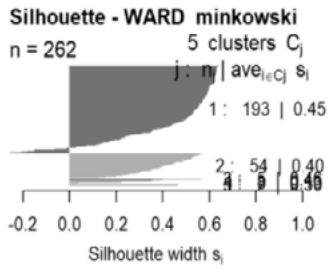

Fig. 7. Gráfico de Silueta para el método de agrupamiento Ward con distancia Minkowski. Elaboración propia.

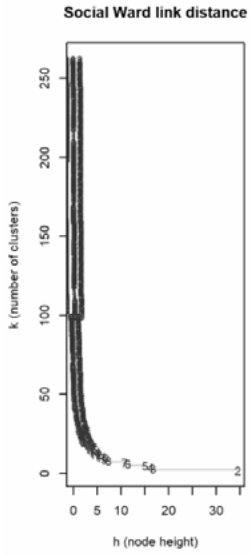

Fig.8. Gráfico de nivel de fusión Ward-Minkowski. Elaboración propia.

En la fase no lineal se obtuvieron los siguientes resultados:

Tabla 3.: Tabla de resultados de las ejecuciones del método t-SNE. Elaboración propia.

TRANSF	DISTANCIA	PERPLEXITY	ITERACIONES	CONVERGE	# PATRONES
LOG	EUCLIDEAN	5:50	10K	No	1
LOG	MAXIMUM	5:50	10K	No	1
LOG	MANHATTAN	5:50	10K	No	1
LOG	CANBERRRA	5:50	10K	SI	2-3
LOG	MINKOWSKY	5:50	10K	No	1
-	EUCLIDEAN	5:50	10K	No	1
-	MAXIMUM	5:50	10K	No	1
-	MANHATTAN	5:50	10K	No	1
-	CANBERRRA	5:50	10K	SI	2-3
-	MINKOWSKY	5:50	10K	No	1
LOG	EUCLIDEAN	5:50	5K	No	1
LOG	MAXIMUM	5:50	5K	No	1
LOG	MANHATTAN	5:50	5K	No	1
LOG	CANBERRRA	5:50	5K	SI	2-3
LOG	MINKOWSKY	5:50	5K	No	1
-	EUCLIDEAN	5:50	5K	No	1
-	MAXIMUM	5:50	5K	No	1
-	MANHATTAN	5:50	5K	No	1
-	CANBERRRA	5:50	5K	No	1
-	MINKOWSKY	5:50	5K	No	1

La salida elegida de la tabla 3, con transformación logarítmica y distancia Canberra para 10 mil iteraciones, se presenta en la Figura 9. La salida elegida de la tabla 3 sin transformación logarítmica y distancia Canberra para 10 mil iteraciones, se muestra en la Figura 10. De las 750 posibilidades de *clusters* a partir de los resultados de t-SNE con las distancias Canberra se eligió el mejor *cluster*. (Tabla 4, Figuras 11 y 12).

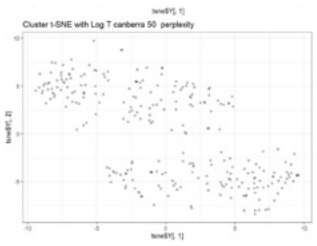

Figura 9. Gráfico de agrupamiento de municipios con el método t-SNE con conversión logarítmica y distancia Canberra. Elaboración propia.

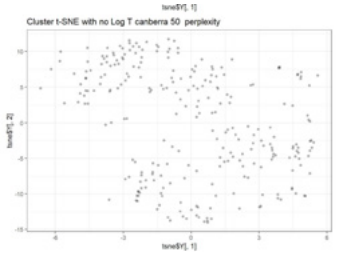

Figura 10. Gráfico de agrupamiento de municipios con el método t-SNE sin conversión logarítmica y distancia Canberra. Elaboración propia.

Tabla 4.: Tabla de resultados de las ejecuciones del método t-SNE. Elaboración propia.

RANK 1 WARD - CANBERRA - 9 CLUSTER (t-SNE)			
Correlación Cofenética	**Distancia Gower**	**Clúster Sugeridos**	**Observación**
8.152.862	4007521	3-7 (Figura 12)	Agrupó mal algunos municipios (Figura 11) Dado que se detectó la presencia de valores atípicos se comprobó mediante las pruebas de Yuend la diferencia entre agrupaciones desde 5 a 9 clústers, resultando a un 0.05 con significantes las agrupaciones La agrupación mostró consistencia a nivel nacional rural-urbano haciendo sub-niveles de valores de las tendencias centrales en la zona rural
RANK 2 CENTROID-EUCLIDEAN - 5 CLUSTER (t-SNE)			
Correlación Cofenética	**Distancia Gower**	**Clúster Sugeridos**	**Observación**
7.602.395	41387.1	4,5	Agrupó mal algunos municipios La agrupación mostró consistencia a nivel nacional rural-urbano, pero al dividir en mas cluster no logró separar el AMSS de los municipios rurales, siendo estos valores atipicos|

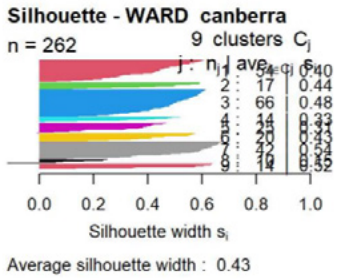

Fig. 11. Gráfico de Silueta para el método de agrupamiento Ward con distancia Minkowski. Elaboración propia.

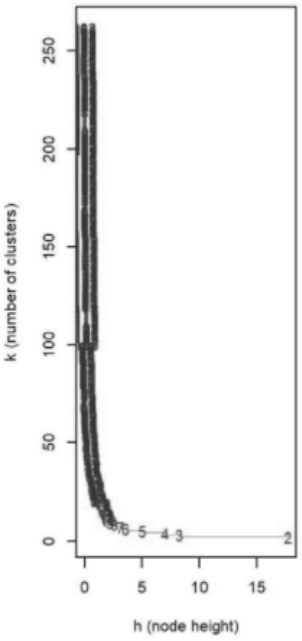

Fig.12. Gráfico de fusión Ward-Canberra. Elaboración propia.

En la Figura 13 se pueden observar los diagramas de cajas y bigotes de cada variable en 9 clústers. La distribución geográfica de los *cluster* se muestra en la Figura 14 y la tabla 5 muestra una caracterización general de estos.

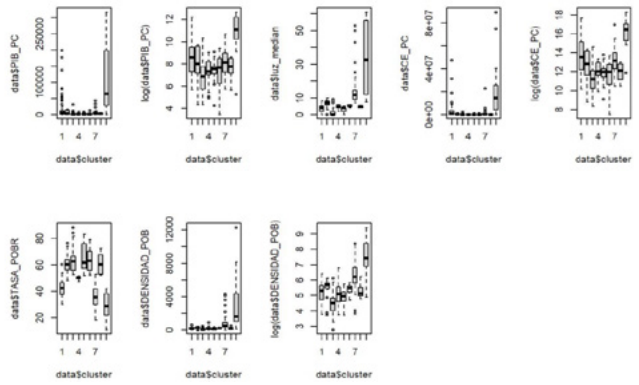

Fig. 13. Diagramas de cajas y bigotes. Elaboración propia.

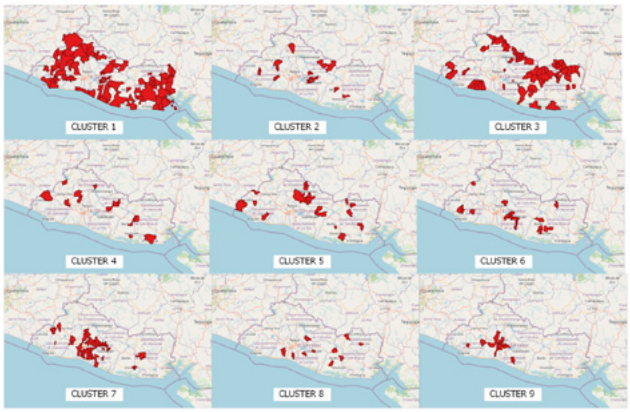

Fig. 14. Agrupamientos de municipios seleccionados, con 9 *cluster* y con el método Ward y distancia Canberra a partir de la reducción de dimensiones t-SNE. Elaboración propia con mapa base de Openstreetmap.

Tabla 5.: Categorización de los *cluster* encontrados con el método Ward y distancia Canberra - T-SNe. Elaboración propia.

Categoria	Cluster	Luz nocturna	PIB per cápita	Consumo eléctrico	Densidad poblacional	Tasa pobreza	# Municipio	Observación
A	9	1 lugar mayor	1 lugar mayor	1 lugar mayor	1 lugar mayor	1 lugar menor	14	
B	1	3 lugar menor	2 lugar mayor	2 lugar mayor	3 lugar menor		54	
C	7	2 lugar mayor	3 lugar mayor	3 lugar mayor	2 lugar mayor	2 lugar menor	42	
D	2	3 lugar mayor	4 lugar mayor	4 lugar mayor	3 lugar mayor	4 lugar mayor	17	
E	8	4 lugar mayor			4 lugar menor		10	mala clasificación
F	6		4 lugar menor	2 lugar menor	3 lugar mayor	1 lugar mayor	20	
G	5	2 lugar menor	3 lugar menor	3 lugar menor	2 lugar menor	3 lugar mayor	25	mala clasificación
H	4	4 lugar menor	2 lugar menor	4 lugar menor		4 lugar menor	14	
I	3	1 lugar menor	1 lugar menor	1 lugar menor	1 lugar menor	2 lugar mayor	66	

IV. CONCLUSIONES

A nivel exploratorio se logró clasificar con mayor detalle el área rural mediante *clustering* jerárquico sólo incorporando como entrada la reducción a dos dimensiones de tipo no lineal (t-SNE) e incorporando la luz nocturna como variable. El mejor modelo detectó nueve *clusters* que clasificaron de manera más detallada el área rural del país y esto se logró mediante el método Ward y la distancia Canberra. Los *cluster* tu-

vieron consistencia en un sentido sociodemográfico, siendo esto de gran importancia para la posible aplicación de políticas públicas. Además, la computación estadística permitió aplicar una metodología robusta donde se hizo uso de múltiples técnicas y parametrizaciones. Por otro lado, el mejor modelo de clasificación sin t-SNE (el cual fue descartado) obtuvo un *cluster* que agrupa a 193 municipios, lo que implicaría que el diseño de posibles políticas públicas sería el mismo para el 73% de los municipios. Las transformaciones logarítmicas de las variables para convertir sus distribuciones en normales arrojaron resultados deficientes con respecto al sentido sociodemográfico de las zonas de país por lo que no aportaron para esta investigación.

Con este estudio se ha dado un primer avance a nivel exploratorio para clasificar el país en agrupaciones de municipios que permiten dividir el área rural del territorio salvadoreño. Esto podría ser validado en una futura investigación mediante un método de clasificación supervisado que permita aceptar la hipótesis de que el área rural efectivamente se puede subdividir en agrupaciones con características sociodemográficas y económicas similares.

AGRADECIMIENTOS

Los autores agradecen el apoyo a César Sánchez del departamento de Economía de la UCA por el procesamiento de los datos sociodemográficos y económicos.

REFERENCIAS

[1] Cummings, Andrew et al. (2019). Identification and socioeconomic characterization of functional territories urban-rural in El Salvador, Central Americ .

[2] Keke Chen Ling Liu *Cluster* Rendering of Skewed Datasets via Visualization College of Computing, Georgia Institute of Technology.df

[3] Nicoleta Rogovschi; Jun Kitazono; Nistor Grozavu; Toshiaki Omori; Seiichi Ozawa t-Distributed stochastic neighbor embedding spectral clustering 2017.

[4] Martin Wattenberg, Fernando Viégas, Ian Johnson. How to Use t-SNE Effectively From https://distill.pub/2016/misread-tsne/

[5] Chen X, Nordhaus W (2011). Using luminosity data as a proxy for economic statistics. The Proceedings of National Academy of Sciences 108(21): 8589-8594. .

[6] Singhal A, Sahu S, Chattopadhyay S, Mukherjee A, Bhanja SN (2020) Using night time lights to find regional inequality in india and its relationship with economic development.

[7] Laveesh Bhandari y Koel Roychowdhury (2011). Night Lights and Economic Activity in india: A study using DMSP-OLS night time images.

[8] Cruz, L., & Penson, E. (2018). Midiendo el desarrollo de República Dominicana desde el espacio. Ciencia, Economía y Negocios, 2(1), 7-9.

[9] Guerrero, Víctor M., Corona F., Mendoza J.A. (2021). Enfoque de big data para generar y analizar datos de actividad económica en México.

[10] Sokal RR, Rohlf FJ. (1962) The comparison of ds.

[11] Jeroen van den Hoven (2016). Clustering with optimised weights for Gower's metric. Mobiquity (Amstelveen).

[12] Peter J. Rousseeuw (1986). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.