# Desigualdad y vacunas

## *Análisis territorial de grandes problemas de Desarrollo en Guatemala con una mirada de vulnerabilidad socioeconómica utilizando aprendizaje de máquina no supervisado*


Con el apoyo de :

Laboratorio de Datos GT

OXFAM

VACUNAS PARA LA GENTE LATINOAMÉRICA

AGENDA REGIONAL PARA UNA VACUNA UNIVERSAL CONTRA LA COVID-19

Laboratorio de Datos GT

# Análisis territorial de grandes problemas de Desarrollo en Guatemala con una mirada de vulnerabilidad socioeconómica utilizando aprendizaje de máquina no supervisado

*Proyecto de investigación: Desigualdad y vacunas en Guatemala*

*Guatemala, 2023*

Por:
Paolo Doménico Marsicovetere Fanjul[1]
Oscar Chávez Valdez[2]
Laboratorio de Datos GT[3]

---

**Este estudio forma parte del proyecto de investigación "Desigualdad y Vacunas en Guatemala", una serie que incluye diagnósticos, estudios de caso y otros materiales de investigación realizado por OXFAMP**

---

1. Licenciado en Física, Universidad del Valle de Guatemala; Posgrado en Datos, Economía y Política de Desarrollo, MITx; Analista de Datos e Investigador del Laboratorio de Datos GT, Guatemala.
2. Ingeniero en Robótica, TEC de Monterrey; Máster en Administración Pública, EDG, Guatemala; Director Ejecutivo e Investigador del Laboratorio de Datos GT, Guatemala.
3. Laboratorio de Datos GT es un centro de pensamiento independiente dedicado a proyectos de desarrollo, investigación y análisis de datos para proponer soluciones basadas en evidencias en Guatemala: http://www.labdedatosgt.com

INTERMÓN y la Asociación Laboratorio de Datos GT en Guatemala. El mismo fue realizado por Laboratorio de Datos GT (Guatemala), bajo la dirección de la Dra. Karin Slowing Umaña y Oscar Chávez, investigadores a cargo del proyecto.

Nuestro objetivo es generar conocimiento sobre la respuesta institucional y el impacto a la sociedad de la emergencia por COVID-19 a nivel nacional y regional para entender y su relación con las desigualdades en poblaciones vulnerables y ofrecer recomendaciones para reducir las brechas. Nuestras acciones están encaminadas al goce efectivo del derecho a la salud, promoviendo el acceso justo a tecnologías sanitarias, el uso eficaz y transparente de los fondos públicos de los gobiernos y organismos multilaterales en la respuesta frente a esta y futuras pandemias.
Para obtener más información sobre la información publicada en este documento, por favor contactar: info@labdedatosgt.com



Diseño, diagramación e ilustración del informe a cargo de SOMOS3 studio.

# Contenido

# 1. Resumen

Guatemala es un país altamente desigual en materia de condiciones sociales, económicas, políticas, culturales y ambientales que están relacionadas entre sí. Esta desigualdad está asociada a vulnerabilidad -propensión a afectación y capacidad de respuesta ante una amenaza- diferenciada ante problemáticas de desarrollo a lo largo de todo el país. En este documento se presenta un ejercicio para clasificar los municipios del país, considerando variables socioeconómicas y de problemas de desarrollo (específicamente, COVID-19, violencia homicida y participación electoral) para ilustrar cómo se pueden incorporar factores asociados a la vulnerabilidad a la comprensión territorial de estos desafíos. Para tal efecto se utilizó un algoritmo de aprendizaje de máquina no supervisado llamado k-means. Específicamente para las problemáticas seleccionadas, se consiguió observar que existen cuatro perfiles de vulnerabilidad entre los municipios del territorio nacional según su promedio de pobreza, adscripción étnica y áreas geográficas para el análisis de la incidencia de COVID-19, la tasa de homicidios y el porcentaje de participación en elecciones: municipios rurales, ladinos y pobres; municipios rurales, indígenas y muy pobres; municipios urbanos, ladinos y poco pobres; y municipios urbanos, indígenas y pobres. Estas clasificaciones son un insumo, principalmente para el desarrollo de política pública, que podría facilitar el desarrollo de directrices de acción diferenciadas según las características de los municipios para responder a estas problemáticas. Además, con este ejercicio se ejemplifica la importancia de incorporar herramientas de investigación innovadoras para ir mejorando la comprensión territorial de las problemáticas que la población guatemalteca enfrenta y aprovechar una creciente disponibilidad de tecnología e información.

## 2. Introducción

El concepto de vulnerabilidad se puede entender como la probabilidad de daño que puede tener un individuo, un hogar o un colectivo producto de su exposición y/o interacción con condiciones y/o factores que amenazan su integridad (física, mental, social, económica, otras); así también, refiere a las capacidades que posee y condiciones existentes para evitar dichas situaciones (Slowing, K., Chavez, O., 2022). Por estas razones, el concepto de vulnerabilidad se emplea generalmente asociado al concepto de riesgo. Más recientemente, también se emplea junto al concepto de resiliencia, entendido éste como la capacidad para recuperarse de un evento adverso. Distintas disciplinas usan el término de manera relativamente diferente, poniendo énfasis en algunos aspectos más que en otros.

Ruiz Rivera, N. (2012) señala los elementos en común que se pueden encontrar en la mayor parte de definiciones de vulnerabilidad: 1) Se define siempre en relación con algún tipo de amenaza, sean eventos (como sequías, terremotos, inundaciones, enfermedades, accidentes, hambrunas o pérdida de empleo, entre otros); 2) La unidad de análisis (individuo, hogar, grupo, comunidad) se define como vulnerable ante una amenaza específica; 3) El análisis de la construcción de vulnerabilidad se hace en dos momentos distintos del proceso. Por un lado, en las condiciones que tiene la unidad de análisis antes de la situación que amenaza que lo hacen más o menos propenso a la afectación, y por el otro, están las formas que desarrolla para enfrentar la situación, una vez ésta ha ocurrido, y que le permiten también ajustarse posteriormente a la nueva situación creada.

Más recientemente, la comprensión de la causalidad de los principales grandes problemas que afectan a la sociedad se ha ido ampliando. Pasando así de una mono causalidad del origen de los problemas, a una comprensión multicausal, que reconoce la existencia de múltiples factores que interactúan para provocar el daño. La teoría multicausal obliga a considerar que no se trata solo de la presencia de factores directos, sino también que las condiciones sociales, económicas, políticas, culturales y ambientales en las cuales las personas que enfrentan un riesgo se encuentran, también son importantes para determinar la probabilidad y magnitud del daño de un fenómeno. La vulnerabilidad a cualquier evento se agudiza por las diversas problemáticas que

impactan en la misma población, el patrón de desarrollo vigente, y la incapacidad de los grupos más débiles de la sociedad para enfrentarlos, neutralizarlos u obtener beneficios de ellos (Pizarro, R. 2001).

Por ejemplo, en el caso de la salud, a partir del año 2012, la Organización Mundial de la Salud -OMS- reconoció e incorporó el concepto de "Determinantes sociales de la salud" en el análisis epidemiológico y en el diseño de las respuestas sanitarias ante los problemas de salud-enfermedad de la población. Estas se definen como "las circunstancias en que las personas nacen crecen, trabajan, viven y envejecen, incluido el conjunto más amplio de fuerzas y sistemas que influyen sobre las condiciones de la vida cotidiana". Se debe considerar que las condiciones anteriores pueden ser altamente diferentes para varios subgrupos de una población y pueden dar lugar a diferencias en los resultados en materia de una emergencia de salud.

Es bajo esta nueva comprensión de multicausalidad, que desde 2021, Laboratorio de Datos GT propone una nueva manera de entender y categorizar el territorio de Guatemala para abordar los problemas de política pública en el país (Slowing, K., Chávez, O., Maldonado, E., & García, A. L., 2021). El objetivo de este documento es presentar una nueva manera de categorizar los municipios del territorio guatemalteco, dentro del contexto de cada problemática de estudio, considerando las características de vulnerabilidad de cada municipio. Esto permite concebir soluciones, desde una visión que reconozca la heterogeneidad del territorio y de esta forma, superar las barreras que se crean a la hora de implementaciones planeadas de manera homogéneas a nivel nacional o que siguen la tradicional división departamental del territorio.

# 3. Metodología

## 3.1. Selección de variables

Para realizar la categorización del territorio, se utilizaron cuatro variables para cada problemática en cuestión: Tres variables de vulnerabilidad socio económica, sumada a una cuarta variable especifica al fenómeno de estudio.

Variable 1: incidencia de la pobreza total en los municipios o porcentaje de la población del municipio que se encuentra en situación de pobreza, calculado como: [cantidad de personas en situación de pobreza] / [población total del municipio]. A pesar de que Guatemala es un país con altos índices de pobreza, por lo que este dato es sumamente relevante para la política pública, la medición de la pobreza es una práctica poco frecuente. La última medición se realizó en 2014 mediante la Encuesta de Condiciones de Vida (ENCOVI), mientras que la última medición de pobreza a nivel municipal se realizó en 2002. Con el objetivo de minimizar el sesgo ocasionado por la desactualización de datos de pobreza en el país, en este estudio se decidió utilizar la incidencia de pobreza municipal por Figueroa, W., Peñate, M., y Marsicovetere, P. (2020) que se estima mediante un método de machine learning (algoritmos supervisados denominados random forest) que se entrenó con datos de la ENCOVI 2014 y se aplicó a datos del Censo 2018 (INE).

Variable 2: incidencia de % población rural en los municipios o porcentaje de la población del municipio que vive en área rural, calculado como: [cantidad de personas que vive en área rural] / [población total del municipio]. Este se estimó directamente a partir de los datos del XII Censo Nacional de Población y VII de Vivienda 2018 (INE).

Variable 3: incidencia de la población ladina en los municipios o porcentaje de la población del municipio que se identifica como ladina, calculado como: [cantidad de personas que se identifican como ladina] / [población total del municipio]. Este se estimó directamente a partir de los datos del XII Censo Nacional de Población y VII de Vivienda 2018 (INE) (pregunta: Según su origen o historia, ¿cómo se considera o auto identifica?).

## 3.2. Clustering

Utilizando estas variables, se implementó un algoritmo de Clustering o 'crear agrupaciones' en español, para crear la categorización de los municipios. Este es un algoritmo de aprendizaje de máquina no supervisado que ejecuta una serie de procedimientos mediante el cual extrae información (patrones, tendencias o conocimiento clave) de un conjunto de datos para clasificarlos. Ellos funcionan mediante la optimización de un parámetro a través de un proceso iterativo y/o analítico.

Un algoritmo de aprendizaje no supervisado produce una variable respuesta para un conjunto de datos que a priori no tiene una. Es decir, si se tiene un conjunto de datos con  variables independientes, el algoritmo de aprendizaje no supervisado es un proceso que toma información de esas variables para generar una variable respuesta[4]. En este caso, un algoritmo de clustering tiene por objetivo clasificar una serie de registros en clusters [grupos] o subconjuntos, buscando que los registros en un subconjunto posean similitudes entre sí y sean diferentes a los registros de otros subconjuntos.

**Recuadro R1.** Cálculo de la distancia euclidiana entre puntos para un algoritmo de k-means

*Variables predictoras - son empleadas por el algoritmo para generar la respuesta.*

*Cada fila representa un registro u observación en la base de datos.*

*Cada columna es una variable distinta de los registros.*

| ID | $X_1$ | $X_2$ | ... | $X_n$ | Y |
|---|---|---|---|---|---|
| 1 | 0.3 | 0.580 | ... | 0 | B |
| 2 | 0.4 | 0.283 | ... | 1 | A |
| 3 | 0.2 | 0.457 | ... | 0 | C |
| . | . | . | ... | . | . |
| . | . | . | ... | . | . |
| . | . | . | ... | . | . |
| m | 0.8 | 0.691 | ... | 1 | C |

Variables respuestas (por ejemplo, la clasificación por cluster). Estas son generadas por algoritmo no supervisado a partir de las tendencias que identifique en las variables predictoras.

**Fuente:** elaboración propia

4. En contraparte, los algoritmos de aprendizaje supervisados son aquellos que se aplican a conjuntos de datos cuando ya hay una variable respuesta conocida. Estudian las asociaciones estadísticas que hay entre la variable respuesta y  y una serie de variables predictoras  x_(1,) x_(2,)..., x_n a lo largo de todos los registros de un conjunto de datos. Ello puede servir para extrapolar tendencias a nuevos registros donde se desconoce la variable respuesta (por ejemplo, para hacer predicciones de clasificaciones por categorías o hacer regresiones numéricas), o para hacer inferencias con respecto al conjunto de datos mismo (estudiar la relación entre una variable predictora y la variable respuesta).

Uno de los algoritmos de clustering más comunes, y el que se ha utilizado en este caso, es *k-means* [*k-promedios*] y este sirve para clasificar un conjunto de datos en una predeterminada cantidad ($k$) de clusters[5]. La manera con la que este algoritmo determina que los datos dentro de cada subconjunto sean similares entre sí y diferentes a los de otros subconjuntos es calculando la distancia entre los datos[6] (Hartigan y Wong, 1979). Si el conjunto de datos posee $n$ variables, podría decirse que cada dato o registro es un punto en un espacio $n$ de dimensiones y la distancia podría calcularse por distintas métricas como la distancia euclidiana.

Sea un punto $a=(a_1,a_2,...,a_n)$ y un punto $b=(b_1,b_2,...,b_n)$, la distancia euclidiana entre ellos estaría dada por:

$$D_2(a,b)=\sqrt[2]{\sum_{i=1}^{n}(a_i-b_i)^2}$$

También podrían usarse otras métricas de distancia, como la distancia de taxi (o Manhattan). Es importante considerar cuál es la métrica de distancia empleada en un algoritmo de clustering porque tienen distintas particularidades. Por ejemplo, la distancia euclidiana es bastante intuitiva y práctica para conjuntos de datos con pocas (2 a 4) dimensiones pero se compleja cuando ellas aumentan (Gu, Angelov, Kangin y Principe, 2017:1). Además, requiere que todas las dimensiones (las variables) tengan escalas similares porque contrariamente las que tienen mayor magnitud dominarán la medida de distancia, haciendo que los clusters giren únicamente en torno ellas, o agregarán complejidad computacional al cálculo (Virmani, Taneja y Malhotra, 2015:2-3).

5. Este algoritmo fue desarrollado independientemente por Lloyd (1957) y Forgy (1965), por lo que también se le conoce como el algoritmo de Forgy-Lloyd.
6. Esto implica que el algoritmo de clustering solo es capaz de utilizar variables numéricas, en caso de ser necesario usar una variables categóricas, estas deben convertirse.

A continuación, se explican los pasos que conlleva el algoritmo de clustering de k-means. En los anexos se ilustra gráficamente un ejemplo de este proceso.

- **Paso 1:** definir en cuántas agrupaciones se quiere clasificar al conjunto de datos y la métrica de distancia a emplearse.
- **Paso 2:** a partir del número deseado de agrupaciones, se generan aleatoriamente igual cantidad de puntos en el espacio de los datos. A estos puntos se les llama centroides.[7]
- **Paso 3:** se calcula la distancia que cada dato del conjunto tiene a cada centroide. Cada dato se agrupará bajo el centroide que tenga más cerca (menor distancia).
- **Paso 4:** para cada agrupación, se calcula el punto promedio o el centro utilizando la métrica de distancia establecida.
- **Paso 5:** el punto central de cada agrupación se convierte en los nuevos centroides.
- **Paso 6:** iterar los pasos 3 a 5 (calcular la distancia que cada dato tiene a los centroides, agruparse bajo el que tenga más cerca, calcular el centro de las agrupaciones y actualizar el centroide) hasta que entre una nueva iteración y la anterior no existan cambios significativos en las agrupaciones.

Un desafío en la implementación de un algoritmo de clustering como k-means consiste en determinar la cantidad idónea de clusters k para clasificar el conjunto de datos. Si bien no existe una única forma objetiva de evaluar la mejor cantidad de clusters, existen métodos que pueden orientar su elección. Entre ellos es el método gráfico del punto codo: para diferentes valores de k, sumar el total del cuadrado de la distancia de cada punto (registro en la base de datos) al centroide del cluster al que pertenece (WSS por sus siglas en inglés). Esta magnitud decrece conforme k aumenta y la cantidad de clusters k sugerida es el punto de inflexión de esta relación. Con ello, se busca encontrar una k óptima tal que, al aumentarla, la disminución en WSS (que tan distante es cada punto al centro de su cluster) ya no amerite el aumento en la complejidad de los resultados por agregar un cluster adicional[8].

Los análisis de aprendizaje de máquina no supervisado tales como los algoritmos de clustering no necesitan una hipótesis predefinida, no necesitan una variable respuesta

7. En aplicaciones comunes del algoritmo, los centroides iniciales parten desde la ubicación de algunos de los puntos (registros) del datos del subconjunto seleccionados aleatoriamente.
8. En caso el punto codo no fuera posible de determinar o poco concluyente, puede emplearse el método de la silueta u otros. Por ejemplo, ver: https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

(la generan) y son flexibles a distintas distribuciones de datos. No suponen, por ejemplo, que las variables posean una distribución normal o que exista algún tipo de relación entre ellas. Estas características hacen que el clustering sea una herramienta efectiva y flexible para el análisis exploratorio – una forma de encontrar patrones en el conjunto de datos como una primera fase de análisis que motiva a subsiguientes preguntas de investigación – así como para realizar análisis definitivos de patrones a lo largo de múltiples variables.

Aplicaciones comunes de los algoritmos de clustering incluyen la segmentación de registros en distintas categorías (que a priori no se conocen): en investigación de mercados pueden emplearse para clasificar a consumidores según sus conductas (Punj y Stewart, 1983), en ecología pueden emplearse para inferir estructuras poblacionales a partir de datos genéticos (Oyelade et al., 2016) o en investigación educativa pueden emplearse para caracterizar patrones de rendimiento entre estudiantes (Tomy y Jacob, 2011)[9].

Es importante mencionar que en este caso, debido a la diversidad de fuentes de información utilizadas, en un primer paso, se normalizaron[10] las variables con el objetivo de que todas las variables sean comparable a través de sus medidas de dispersión; luego usando el diagrama de codo se determinó la cantidad óptima de clusters (se decidió utilizar 5 clusters o categorías[11]); y finalmente se implementaron tres algoritmos de k-means con distancia euclidiana (uno por problemática) para generar las 3 categorizaciones del territorio.

9. Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. En Journal of Marketing Research (Vol. 20, Issue 2, p. 134). JSTOR. https://doi.org/10.2307/3151680 https://www.jstor.org/stable/3151680; Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., & Adebiyi, E. (2016). Clustering Algorithms: Their Application to Gene Expression Data. En Bioinformatics and Biology Insights (Vol. 10, p. BBI.S38316). SAGE Publications. https://doi.org/10.4137/bbi.s38316; M, B., Tomy, J., A, U., & Jacob, P. (2011). Clustering Student Data to Characterize Performance Patterns. En International Journal of Advanced Computer Science and Applications (Vol. 1, Issue 3). The Science and Information Organization. https://doi.org/10.14569/specialissue.2011.010322

10. A cada dato de la distribución (el valor que cada municipio tiene en la variable) se le resta el promedio de la distribución (el promedio nacional de la variable) y luego se divide en la magnitud de la desviación estándar de la distribución: $\frac{x-\mu(x)}{\sigma(x)}$ Esto centra la distribución de cada variable en 0 y la magnitud resultante de cada dato es igual a la cantidad de desviaciones estándard que está del promedio de la distribución.

11. Cabe reiterar que la determinación de clusters mediante el diagrama de codo posee un grado de subjetividad. Se decidió trabajar con 5 categorías de manera empírica para reducir la dispersión interna de los clusters en cuanto a las variables de interés.

# 4. Clusters para el análisis territorial de diferentes problemas de Desarrollo en Guatemala, considerando la vulnerabilidad socioeconómica

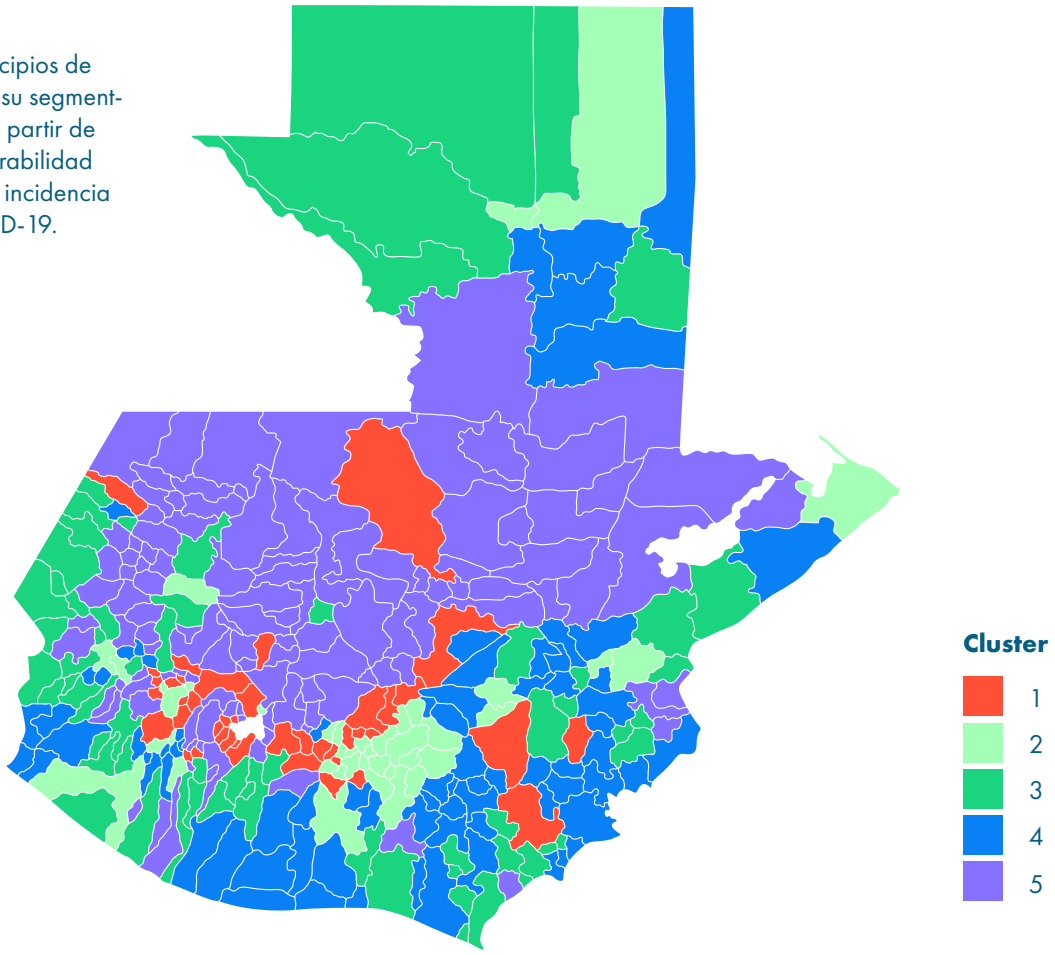
El surgimiento de nuevas tecnologías y sistemas de información que permiten la constante reportería y recopilación de datos estadísticos posibilita el análisis y formulación de propuestas ante problemáticas socioeconómicas a lo largo del territorio nacional. Por ejemplo, la desigualdad en el acceso a servicios públicos de salud, las brechas en la matrícula educativa y las expresiones de violencia en contextos electorales son fenómenos complejos que podrán afrontarse más efectivamente con una mayor producción de información respecto a ellos. Sin embargo, debido a la creciente disponibilidad de datos y la complejidad inherente en algunos de estos desafíos, los métodos de investigación cuantitativa tradicionales tienen limitaciones en su capacidad de analizar grandes conjuntos de datos y de discernir patrones subyacentes en ellos. Por tanto, es necesario implementar metodologías innovadoras en la investigación socioeconómica y de políticas públicas, tales como el aprendizaje de máquina.

Además, como se explicó anteriormente, reconociendo que Guatemala es un país con altos niveles de desigualdad, es necesario que la lectura de las problemáticas socioeconómicas considere las distintas causas, expresiones y consecuencias que estas pueden tener dependiendo de los factores de exposición y vulnerabilidad de los habitantes. Por tanto, a continuación, se desarrollan tres ejemplos de análisis de clustering donde se agrupan municipios en el contexto de tres problemáticas diferentes, pero para cada una, considerando también las variables de vulnerabilidad socioeconomía antes mencionadas (porcentaje de pobreza general, porcentaje de ruralidad y porcentaje de población ladina). Los tres problemas de interés son impacto de la pandemia por COVID-19 (incidencia de casos COVID-19); la violencia homicida (tasa de homicidios) y participación política electoral (participación en elecciones de 2023). El objetivo de estos ejercicios es generar un mejor entendimiento territorial de estos fenómenos dentro del contexto de vulnerabilidad socioeconómica del país con el propósito de contar con mejor información para estrategias de abordaje desde la institucionalidad pública.

## 4.1. Análisis territorial del impacto de la pandemia considerando la vulnerabilidad socioeconómica e incidencia de casos COVID-19

Para el primer caso se utilizó la incidencia municipal de casos de COVID-19, definida como la cantidad de casos confirmados de COVID-19 a julio de 2022, dividida por la cantidad de habitantes en cada municipio[12]. Las agrupaciones resultantes se ilustran en el mapa M1, mientras que en la gráfica G2 se muestran sus distribuciones de las variables del análisis mediante diagramas de caja y bigote. Cabe señalar que los clusters generados se etiquetan con un número ("1", "2", "3", "4" y "5"), pero este es arbitrario y su orden carece de significado[13].

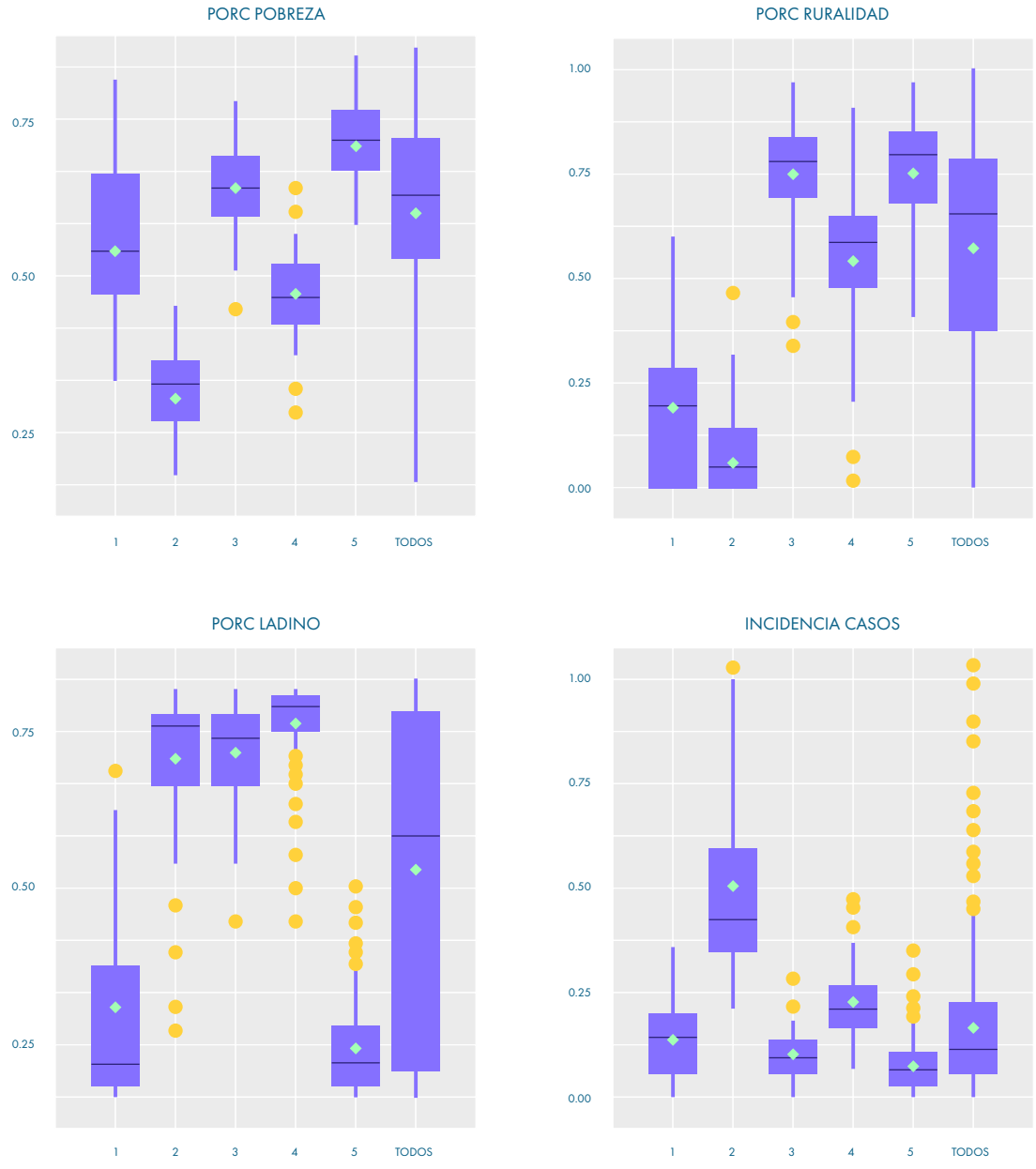
**Mapa M1:** Municipios de Guatemala según su segmentación en clusters a partir de variables de vulnerabilidad socioeconómica e incidencia de casos de COVID-19.




**Fuente:** elaboración propia con datos del MSPAS (2022), INE (2018) y ENCOVI (2014)

12. Fuente: tablero de Situación de COVID-19 en Guatemala, MSPAS -Ministerio de Salud Pública y Asistencia Social-, 2022: https://tableros.mspas.gob.gt/covid/
13. Surgen de la distribución inicial aleatoria de los centroides al inicio del algoritmo, por lo que únicamente se usan para etiquetar a las agrupaciones resultantes.

**Gráfica G2:** Diagramas de caja[14] y bigote por cluster de cada variable utilizada para generar la categorización: porcentaje de pobreza, porcentaje de ruralidad, porcentaje de población ladina e incidencia de casos de COVID-19.




**Fuente:** elaboración propia

14. Nota: el rombo rojo representa el promedio de las distribuciones, mientras que la línea horizontal en cada caja ilustra la mediana. Las cajas encierran al 50% de los datos de cada distribución (aquellos comprendidos entre el quintil 1 y 3), mientras que los puntos negros representan datos atípicos. En los anexos pueden encontrarse tablas con las magnitudes numéricas de las distribuciones.

Tal como se muestra en la gráfica G2:

**CLUSTER 1**

Está compuesto por **46 municipios** predominantemente urbanos e indígenas. En promedio, *el 21.5% de los habitantes de estos municipios residen en áreas rurales y el 17.2% de ellos se auto identifican como ladinos.* El promedio del porcentaje de pobreza de los municipios de este cluster es de 56.1%, colocándolo en un nivel de pobreza medio (en relación al promedio nacional de 59.6%)[15]. La mayoría de los municipios que fueron agrupados acá pertenecen a los departamentos de **Sacatepéquez, Chimaltenango, Sololá y Quetzaltenango,** y entre ellos se encuentran las cabeceras departamentales de **Sololá, Totonicapán, Santa Cruz de Quiché, Salamá, Cobán, Jalapa y Jutiapa,** y en total tienen **2.34 millones de habitantes**. El promedio de la tasa de incidencia de Covid-19 en los municipios de este cluster es de 1.11.

**CLUSTER 2**

Está compuesto por **42 municipios** que se caracterizan por tener los menores porcentajes de habitantes viviendo en áreas rurales (11.0%, en promedio) y en situación de pobreza (26.1%, en promedio), así como por tener habitantes con alta autoidentificación como ladinos (en promedio, 82.4%). *21 de estos municipios se encuentran en el departamento de Guatemala y Sacatepéquez (incluyendo Ciudad de Guatemala y Antigua Guatemala) y el resto corresponde a otras cabeceras departamentales (Guastatoya, Chimaltenango, Quetzaltenango, Retalhuleu, San Marcos, Flores, Puerto Barrios y Zacapa)* o a municipios que forman parte del área de influencia de las ciudades en ellos. Estos municipios suman **4.89 millones de habitantes**, y destacan por presentar los más casos de COVID-19, con una tasa de incidencia promedio de 4.00. Incluye a San Lucas y Guatemala que alcanzaron una tasa de incidencia de 7.43 y 7.10.

15. Según el indicador de pobreza desarrollado por Figueroa et al. (2020).

**CLUSTER 3**

Es una agrupación de **68 municipios ladinos** con niveles medio-altos de pobreza y ruralidad que a su vez mostraron una baja incidencia de COVID-19. En estos municipios hay **2.41 millones de personas y 83.8% de autoidentificación promedio como ladinos, 77.3% de residencia rural promedio y 66.1 % de pobreza promedio.** Muchos de estos municipios son fronterizos con otros países y la mayoría están en San Marcos (14), Suchitepéquez (8), Huehuetenango (8) y Jutiapa (8). En promedio, tienen una tasa de incidencia de 0.82.

**CLUSTER 4**

Tiene características similares al "3", excepto que los municipios de este son menos pobres *(en promedio, tienen una pobreza de 45.7%*), ligeramente más urbanos *(en promedio, 57.6% de habitantes en áreas rurales)* y ladinos *(en promedio, 90.3%)* y poseen una mayor incidencia de COVID-19 de 1.68, la cual solo está por debajo de la del cluster "2". El cluster acumula un total de **2.58 millones de habitantes** y figura 76 municipios, sobre todo de Escuintla (12), Santa Rosa (10), San Marcos (8) y Jutiapa (7).

**CLUSTER 5**

Es el más grande, contando con **108 municipios y 5.47 de millones de personas**, provenientes particularmente de las regiones del **Norte, Noroccidente y Altiplano del país**. Este cluster corresponde a municipios con población indígena (en promedio, **los municipios tienen 10.2% de población** que se autoidentifica como ladina), rural (76.9%), y con muy altos grados de pobreza (80.0%), a la vez que poseen las menores tasas de incidencia de COVID-19 con una tasa de incidencia promedio de 0.55

De manera general, los clusters pueden describirse con los siguientes perfiles: el cluster "1" son municipios urbanos, indígenas y pobres; el "2", municipios urbanos, ladinos y poco pobres; el "3" y "4", municipios rurales, ladinos y pobres (pero el "4" con más incidencia de Covid-19 que el otro); y el 5, municipios rurales, indígenas y muy pobres.

Los resultados de este ejercicio podrían utilizarse para desarrollar una estrategia de mitigación al impacto por la pandemia por COVID-19 diferenciada por municipio, ya que este facilita el análisis del territorio no solo por la variable de interés, si no por las características heterogéneas de cada territorio. La estrategia podría tomar en cuenta la incidencia de casos para planificar la magnitud de las intervenciones necesarias. En los municipios del cluster "2" podría implementarse intervenciones más agresivas puesto que poseen una incidencia significativamente más alta que los municipios de otros clusters; seguidos por nivel de incidencia de casos los municipios de los clusters "4" y "1", por último, los municipios de clusters "3" y "5".

Pero más importante aún, si bien estas intervenciones responden directamente a la incidencia de casos, presentan la oportunidad de planificar con pertenencia cultural, adecuándose a las características poblaciones de cada municipio; y considerando otro tipo de problemáticas que de otra manera son ignoradas:

- Por ejemplo, si bien los municipios del cluster "5" registran la menor incidencia de casos por COVID-19, también son los municipios más vulnerables (son los que registran mayor pobreza, ruralidad y mayor población indígena). Por esto, podemos suponer que presentan un limitado acceso a servicios de salud y privaciones en derechos, y también que existe un alto subregistro de casos (Slowing y Chávez, 2022). Desde el diseño de la estrategia se pueden concebir acciones para superar estas otras barreras que afectan a los territorios como de infraestructura, idioma, etc.
- Similarmente, la estrategia podría adaptar sus acciones en municipios rurales (clusters "3", "4" y "5") en términos logísticos para acceder a poblaciones con mayor dispersión en el territorio o con reducida infraestructura básica (por ejemplo, conseguir equipo especializado para transportar y mantener la cadena de frío (Slowing y Chávez, 2021)).

Una propuesta así fue realizada por el Laboratorio de Datos GT (2021) al MSPAS utilizando esta metodología[16].

Es importante señalar la flexibilidad de esta metodología, que permite utilizar la variable de incidencia de casos para crear los clusters, pero pudiera reproducirse con otra variable de interés, como podría ser, por ejemplo, la cantidad de muertes por COVID-19 o el exceso de mortalidad[17]. Además, esta categorización permite cierta dispersión en de las variables a lo interno de cada cluster. En la práctica, esta metodología podría facilitar el análisis para la identificación de municipios y casos paradigmáticos, para planificar acciones específicas dentro de cada cluster.

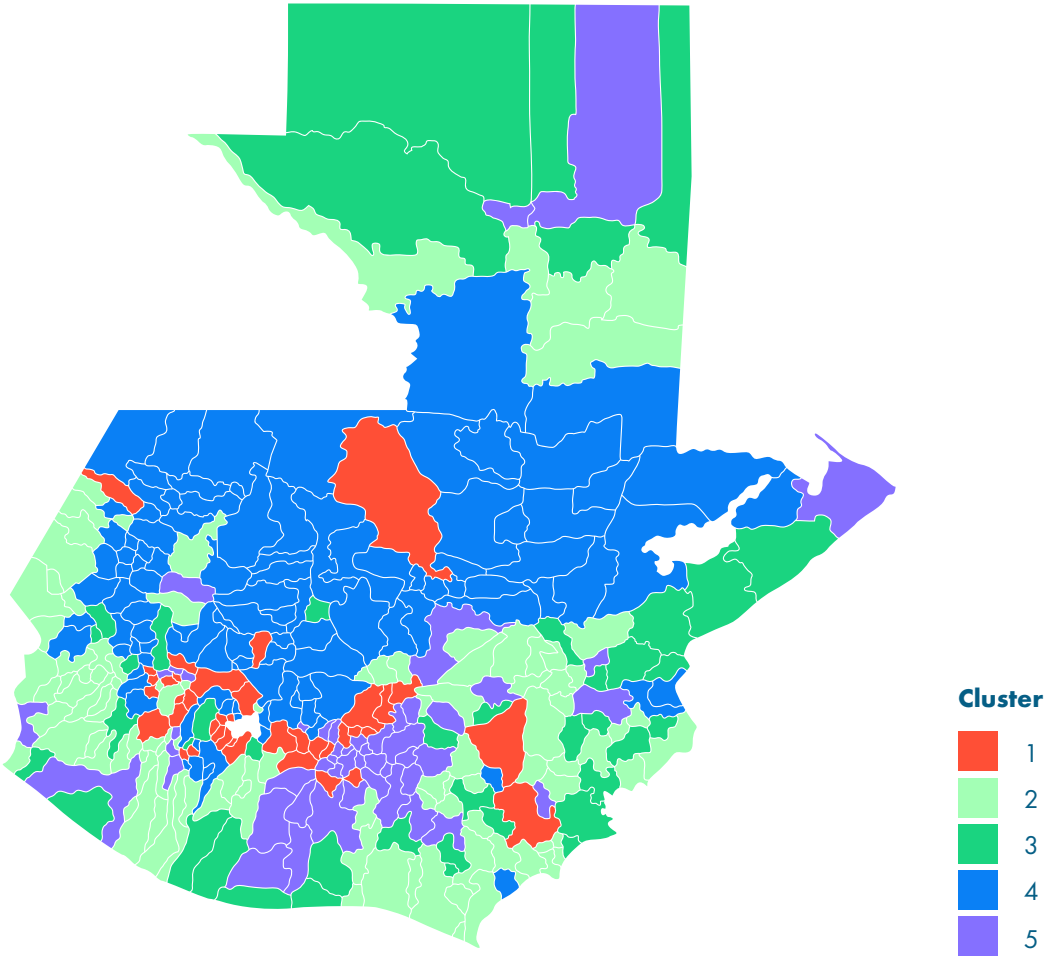
16. Ver https://lac.oxfam.org/lo-%C3%BAltimo/publicaciones/propuesta-para-fortalecer-la-aplicacion-equitativa-de-la-vacuna-contra
17. Ver https://labdedatosgt.com/exceso-de-muertes-durante-la-emergencia-por-COVID-19-en-guatemala-2020-2021/

## 4.2. Análisis territorial de la violencia homicida considerando la vulnerabilidad socioeconómica y tasa de homicidios por municipio

En este segundo caso de estudio, se vuelven a utilizar las mismas variables de vulnerabilidad socioeconómica (porcentaje de personas viviendo en condiciones de pobreza, porcentaje de personas viviendo en un área rural y porcentaje de personas que se autoidentifican como ladinos en los municipios), pero en el algoritmo, se combinan con una cuarta variable de interés sobre violencia homicida: tasa de homicidios. Esta última se obtiene sumando la totalidad de los homicidios registrados entre 2019 y 2022, y se divide por la cantidad de habitantes en el municipio[18], multiplicadas por 100,000. Los resultados del algoritmo de clasificación ahora se muestran en el siguiente mapa M2 y los diagramas en el gráfico G3.

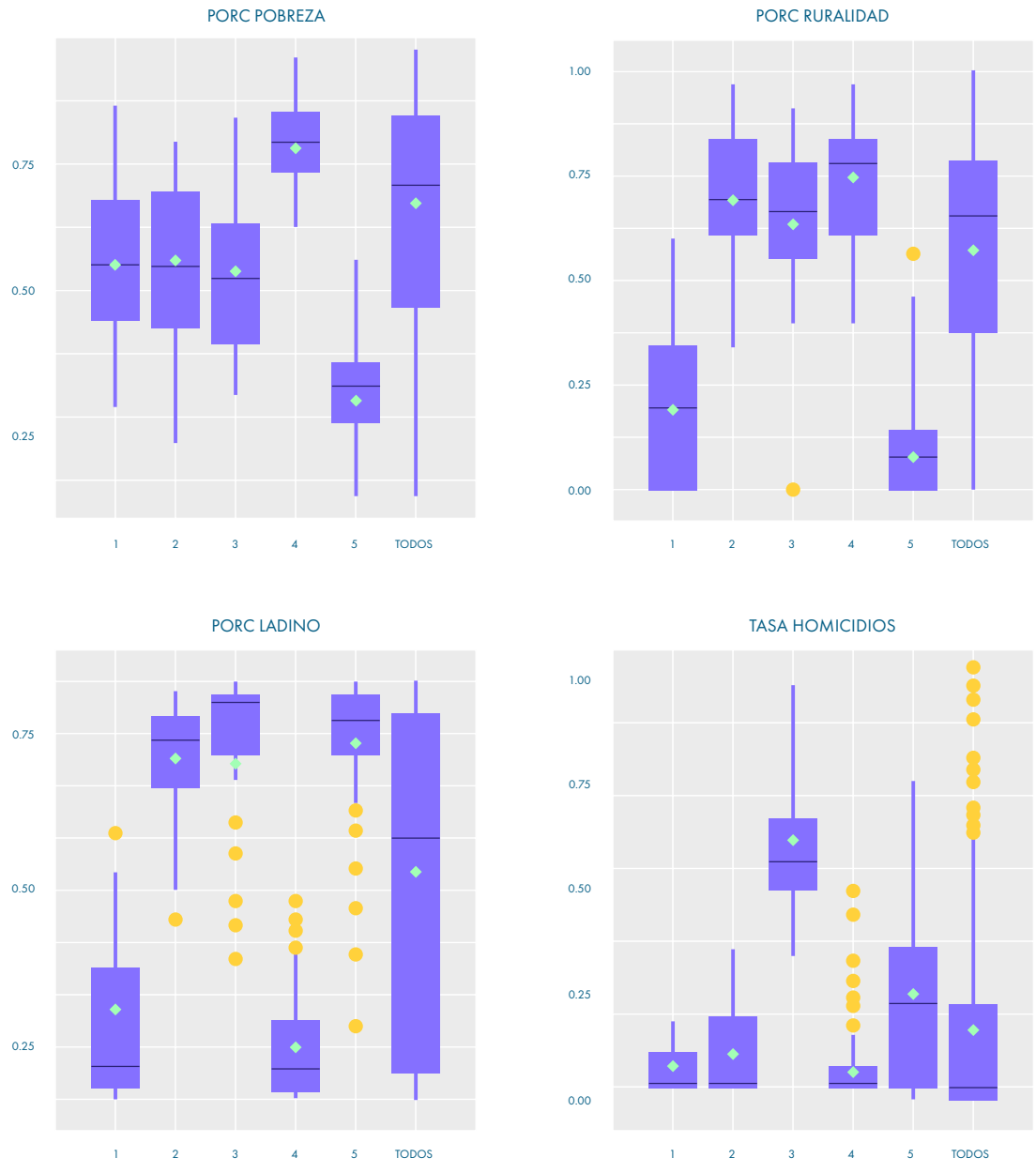
18. Estimaciones y proyecciones de población a 2022, Instituto Nacional de Estadística

**Mapa M2:** Municipios de Guatemala según su segmentación en clusters a partir de variables de vulnerabilidad y tasa de homicidios entre 2019 y 2022 por cada 100,000 habitantes




**Fuente:** elaboración propia con datos del TSE (2023), INE (2018) y ENCOVI (2014)

**Gráfica G3:** Diagramas de caja[19] y bigote por cluster de cada variable utilizada para generar la categorización: porcentaje de pobreza, porcentaje de ruralidad, porcentaje de población ladina y tasa de homicidios entre 2019 y 2022 por cada 100,000 habitantes




**Fuente:** elaboración propia con datos de PNC (2019-2022), INE (2018) y ENCOVI (2014)

19. Nota: el rombo rojo representa el promedio de las distribuciones, mientras que la línea horizontal en cada caja ilustra la mediana. Las cajas encierran al 50% de los datos de cada distribución (aquellos comprendidos entre el quintil 1 y 3) mientras que los puntos representan datos atípicos. En los anexos pueden encontrarse tablas con las magnitudes numéricas de las distribuciones.

En este caso:

El cluster con la mayor tasa de homicidios es el "2"[20], cuyos municipios tuvieron un promedio de **162.6 homicidios por cada 100,000 habitantes entre 2019 y 2022**. Los municipios que conforman este cluster se encuentran entre los municipios con la mayor tasa a nivel nacional (incluyendo a Tiquisate y Jerez con tasas atípicamente altas de 332 y 300) y en promedio poseen **50.9% de población en condiciones de pobreza, 59.7% de población en áreas rurales y 92.2% de población ladina** (casi todos son completamente ladinos, salvo algunas excepciones como Camotán y Olopa de Chiquimula). La mayoría de los municipios de este cluster se encuentran en Escuintla, Petén, Zacapa o Jutiapa.

El cluster "1" es relativamente similar al "2" en términos de porcentaje de pobreza, porcentaje de población ladina y porcentaje de población que reside en área rural (en promedio, respectivamente, 58.1%, 83.9% y 71.8%) pero contrasta con una tasa de homicidios promedio sustancialmente menor de 44.4. Lo componen 102 municipios alrededor del país que concentran 3.03 millones de habitantes.

El cluster "5" le sigue al "2" con **una tasa municipal de homicidios promedio de 79.7**.[21] Este cluster corresponde a 46 municipios urbanos, ladinos y poco pobres tales como las cabeceras departamentales y municipios de Sacatepéquez

20. De nuevo, la asignación de las etiquetas de los clusters ("1", "2","3", "4" y "5") son aleatorias; el cluster "1" en este caso no es el mismo cluster "1" que en el ejercicio anterior.
21. Cabe resaltar la diferencia entre tasa de homicidios y cantidad bruta. Algunos municipios tienen una alta tasa de homicidios debido a que tienen relativamente pocos habitantes por lo que municipios con menor población y escasos homicidios podrían registrar una mayor tasa de homicidios frente a municipios con mayor población y una gran cantidad de homicidios. Entre otros, este es el caso del municipio de San José, Petén, que con un 1 homicidio registrado en el período de análisis tiene una tasa de 13.7 por cada 100 mil habitantes mientras que el municipio de Guatemala con 146 registrados tiene una tasa de 12.0 por cada 100 mil habitantes. San José resultó agrupado en el cluster "3" junto a otros municipios presuntamente de alta alerta por homicidios, mientras que Guatemala en el cluster "5".

y Guatemala, y acumula 5.50 millones de habitantes. **Tienen un promedio de pobreza de 29.4% pobreza**, de 14.3% residentes en áreas rurales y de 85.7% de autoidentificación ladina.

Por otro lado, los clusters "3" y "4" poseen las menores tasas de homicidios con promedios de 20.2 y 12.1, respectivamente. De hecho, 25 de 105 (23.8%) municipios en el cluster "4" no registraron ningún homicidio en el período de análisis. Los municipios de ambos clusters tienen población mayoritariamente indígena, con un promedio de autoidentificación ladina de 14.5% para el "3" primero y 10.0% para el otro. La principal diferencia entre ambos es que el "1" corresponde a municipios urbanos y medianamente pobres (en promedio, 21.7% rurales y 56.6% pobres) como los de Quetzaltenango, Sacatepéquez, Guatemala y Sololá, mientras que el "4" corresponde a municipios rurales y muy pobres (en promedio, 76.9% rurales y 79.8% pobres), como los de Alta Verapaz, Quiché y Huehuetenango.

Análogamente al ejercicio con la variable de incidencia de COVID-19, los cinco clusters que se generaron con la variable de tasa de homicidios pueden resumirse bajo perfiles basados en las tres características socioeconómicas de los municipios en ellos. Un cluster urbano-indígena-pobre ("1"), un cluster urbano-ladino-poco pobre ("5"), dos clusters rurales-ladinos-pobres (uno con mayor tasa de homicidios -"3"- y otro con menor - "4") y un cluster rural-indígena-muy pobre ("4"). Si bien los análisis de clustering no son capaces de determinar la causalidad detrás de las variables utilizadas en cuestión, ellos son herramientas útiles para identificar patrones e interrelaciones entre observaciones de los fenómenos que pueden complementar esfuerzos más amplios para estudiarlos o abordarlos.

De los 11,946 homicidios registrados entre 2019 y 2022, solo 754 (6.3%) de ellos ocurrieron en los municipios urbanos e indígenas del cluster "3" y solo 780 (6.5%) en los municipios rurales e indígenas del cluster "4", a pesar que juntos reúnen a 150 de los 340 municipios (44%) del país. Salvo por algunas excepciones[22], la violencia homicida en estos municipios es relativamente esporádica y podría hipotetizarse que la organización comunitaria y el tejido social que yace en ellos es un factor determinante en su prevención. Asimismo, se puede hipotetizar que dichos eventos esporádicos se lleven a cabo por personas en estado de emoción violenta o de embriaguez[23].

Por otro lado, el cluster "5", que está conformado principalmente por municipios urbanos y ladinos, concentra 5,921 (49.6%) de todos los homicidios en el país (de estos, 2,172 o el 28.2% ocurrieron en Guatemala, Villa Nueva y Mixco). Los municipios de este cluster son relativamente ricos, pero también poseen altos niveles de desigualdad.[24] Por tanto, podría estudiarse la hipótesis de que la desigualdad ha impulsado la violencia homicida en estos municipios[25]. Otra hipótesis que podría estudiarse tanto en el contexto de estos municipios urbanos ladinos del cluster "5" como en los municipios rurales ladinos del cluster "3" es que las altas tasas de homicidio están relacionadas con la presencia de centros de operaciones, rutas y puertos que utilizan estructuras de crimen organizado.

22. San Juan Sacatepéquez y Jalapa, ambos del cluster "1", registraron 19 y 17 homicidios - más que la mitad de todo ese cluster. El Estor, del cluster "4", registró 10.
23. Según el documento Disminución de homicidios en Guatemala: una mirada desde la prevención (2014) de Carmen Rosa de León-Escribano, "Los altos niveles de violencia en el país están asociados a dos factores de riesgo: proliferación y fácil acceso a las armas de fuego y al alto consumo de alcohol." https://iepades.org/wordpress/wp-content/uploads/2020/08/Disminucion-de-homicidios-en-Guatemala.pdf
24. Una expresión de esa desigualdad, por ejemplo, es que, en los municipios de Guatemala, Villa Nueva y Mixco habían por lo menos 229 asentamientos informales, según el Censo de Asentamientos de TECHO-Guatemala (2016).
25. Existen diversos mecanismos con los que la desigualdad puede impulsar homicidios: "1) la experiencia de vivir en privación relativa genera sentimientos de frustración que pueden afectar las relaciones interpersonales [genera una sensación de injusticia entre las personas en desventaja que les lleva a buscar una compensación por otros medios]; 2) vivir en situación de privación relativa da lugar a subculturas de la hostilidad, la cual se canaliza hacia el círculo familiar o es adaptada al pequeño contexto urbano en la forma de un "código de la calle" (code of the street); 3) en contextos con altos niveles de privación económica hay un incremento en el número de oportunidades criminales porque los objetivos probables son mucho más visibles debido a la extendida desigualdad. " (Ramírez de Garay, 2014). Por otro lado, "la actividad criminal también se puede explicar por un análisis costo-beneficio; cuanto más escasas sean las oportunidades económicas para los más pobres y mayor sea la brecha de ingreso entre pobres y ricos, los beneficios económicos de crímenes como robos o secuestros –que muchas veces terminan en homicidios- tienden a ser mayores" (Winkler, 2014).
https://www.bancomundial.org/es/news/feature/2014/09/03/latinoamerica-menos-desigualdad-se-reduce-el-crimen
https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-57952014000100010

> *Está en el norte un puerto muy importante de Guatemala en Puerto Barrios que da al Atlántico y luego tenemos en el sur el segundo puerto más grande de Guatemala que está en el departamento de Escuintla, y si uno traza una línea de norte a sur o de sur a norte, se concentra la mayoría de la violencia en los departamentos que uno atraviesa de puerto a puerto. Entonces, muchas de las personas que han tenido la oportunidad de investigar sobre esto, una de las principales hipótesis que manejan es que hay una estrecha relación con el tema de drogas, tráfico de armas y tráfico de personas.*
>
> *(Corzo, 2017)*

De cualquier manera, cabe resaltar la utilidad de la metodología para dividir el territorio según las características priorizadas para tener una comprensión inicial. Al permitir reconocer cuáles y cómo son los grupos de municipios con semejantes tasas de homicidio y características socioeconómicas entre sí, ella puede motivar hipótesis subsiguientes y el diseño de teorías de cambio y estrategias de intervención.

## 4.3. Análisis territorial de la participación política considerando la vulnerabilidad socioeconómica y participación en las elecciones

La participación en (y el reconocimiento de los resultados de) elecciones libres es uno de los pilares de la democracia. A través de ese deber y derecho, los ciudadanos se ven representados en la toma de decisiones públicas y en formación de las leyes y políticas que les gobiernan. A medida que más personas se unan al padrón electoral y voten, es más probable que las acciones de entidades públicas se alineen a sus necesidades e intereses, se promueva el cambio social y respeto a los derechos y libertades y se fiscalice a funcionarios públicos.

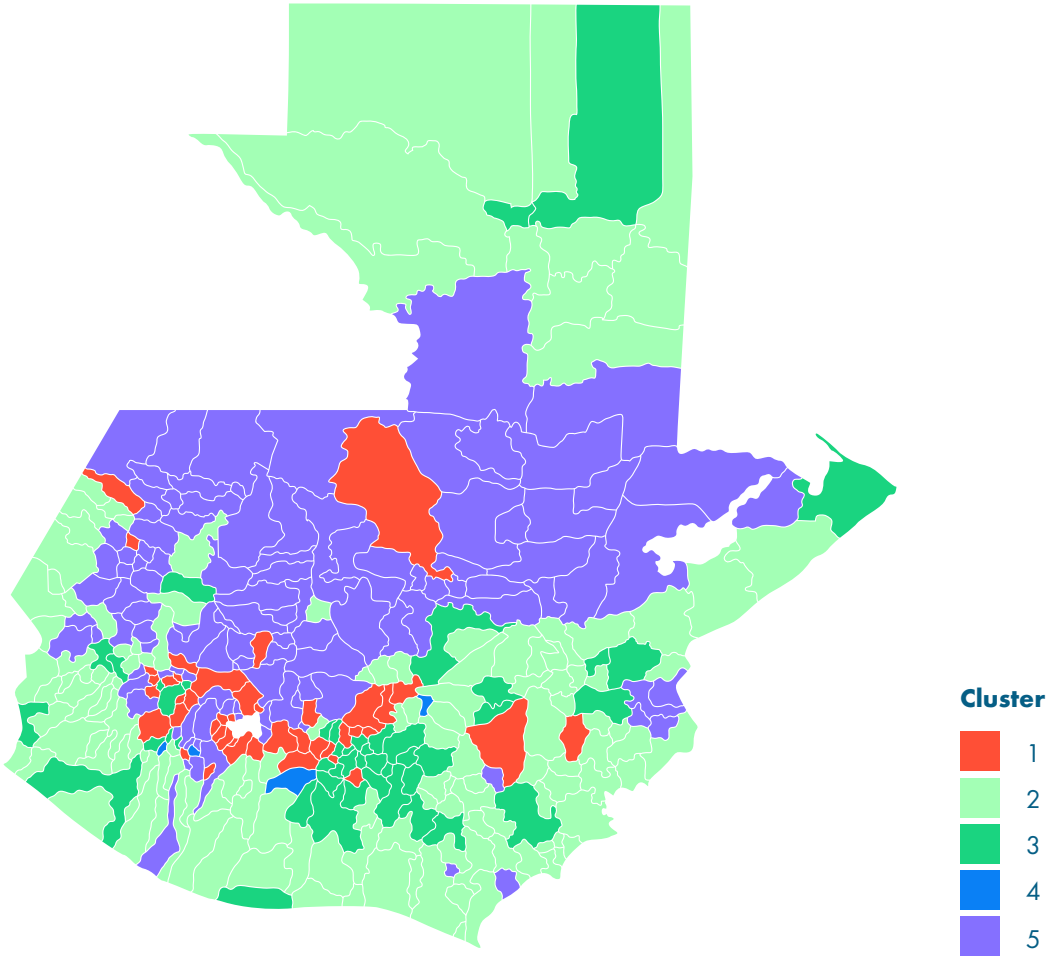
El domingo 25 de junio de 2023 se celebraron elecciones generales y de diputados al Parlamento Centroamericano para el período de 2024 a 2028. Según el portal de resultados preliminares del Tribunal Supremo Electoral (TSE)[26] al 29 de junio, del padrón electoral de 9.27 millones de personas, 5.54 (60.6%) de ellas asistieron a las urnas y 3.60 (39.4%) se abstuvieron de participar. Si bien este porcentaje de participación está en línea con el promedio mundial[27], la cantidad de personas que no asistieron es altamente significativa. Como punto ilustrativo, en las elecciones a diputados por listado nacional, los 6 partidos con más votos acumularon 2.55 millones y 24 diputaciones[28] del total de 32. Hipotéticamente, si 3.60 millones de personas adicionales hubiesen votado válidamente por otro partido, estos 2.55 millones de votos solo se habrían traducido en 8 diputaciones[29], transformando radicalmente el panorama del Congreso. Similarmente, los partidos de UNE y Semilla pasaron a segunda vuelta de las elecciones a presidente y vicepresidente habiendo

26. https://www.trep.gt/
27. El promedio mundial de participación es de 63.8% para elecciones presidenciales y de 60.9% para elecciones parlamentarias, según Wisevoter (2023). https://wisevoter.com/country-rankings/voter-turnout-by-country/ .
28. Vamos, con 0.63 millones (6 diputados); UNE, con 0.54 millones (5 diputados); Semilla, con 0.49 millones (5 diputados); Cabal, con 0.37 millones (3 diputados); VIVA, con 0.29 millones (3 diputados) y Coalición Valor Unionista, con 0.23 (2 diputados). Estos datos fueron recopilados del portal de resultados preliminares del Tribunal Supremo Electoral, disponible en: https://www.trep.gt/#!/tc2/ENT
29. Estos 2.55 millones de votos pasarían de ser el 60% de todos los votos válidos emitidos (2.55 de 4.17), a 33% (2.55 de 7.77). 7.77 millones de votos / 32 diputaciones = 0.24 millones de votos / diputación. VAMOS, UNE y Semilla habrían conseguido 2 diputados cada uno y CABAL y Viva habrían conseguido 1 cada uno. La coalición VALOR-Unionista se habría quedado sin diputados por listado nacional.

conseguido respectivamente 0.88 y 0.65 millones de votos. 3.60 millones de votos válidos adicionales para otros partidos podrían dejar a esos dos partidos en quinto y sexto lugar en vez de primero y segundo[30]. En este ejercicio se desarrollan clusters con la variable de interés de porcentaje de participación en las elecciones generales de 2023 por municipio[31], con miras a que ello contribuya al análisis y diseño de estrategias para promover una mayor participación libre, consciente e informada en futuros procesos democráticos en el país. La fórmula empleada para calcular este porcentaje se dividió la cantidad total de votos emitidos, tanto válidos como inválidos, dentro del padrón electoral, por cada municipio. La fórmula empleada para calcular este porcentaje se dividió la cantidad total de votos emitidos, tanto válidos como inválidos, dentro del padrón electoral, por cada municipio. Al igual que en los casos anteriores, este ejercicio se realizó para 5 clusters e incluye a las variables asociadas a vulnerabilidad socioeconómica de porcentaje municipal de población en condiciones de pobreza, de población que reside en áreas rurales y de población que se autoidentifica como ladina. Los resultados se ilustran mediante un mapa y diagramas de caja de las variables.

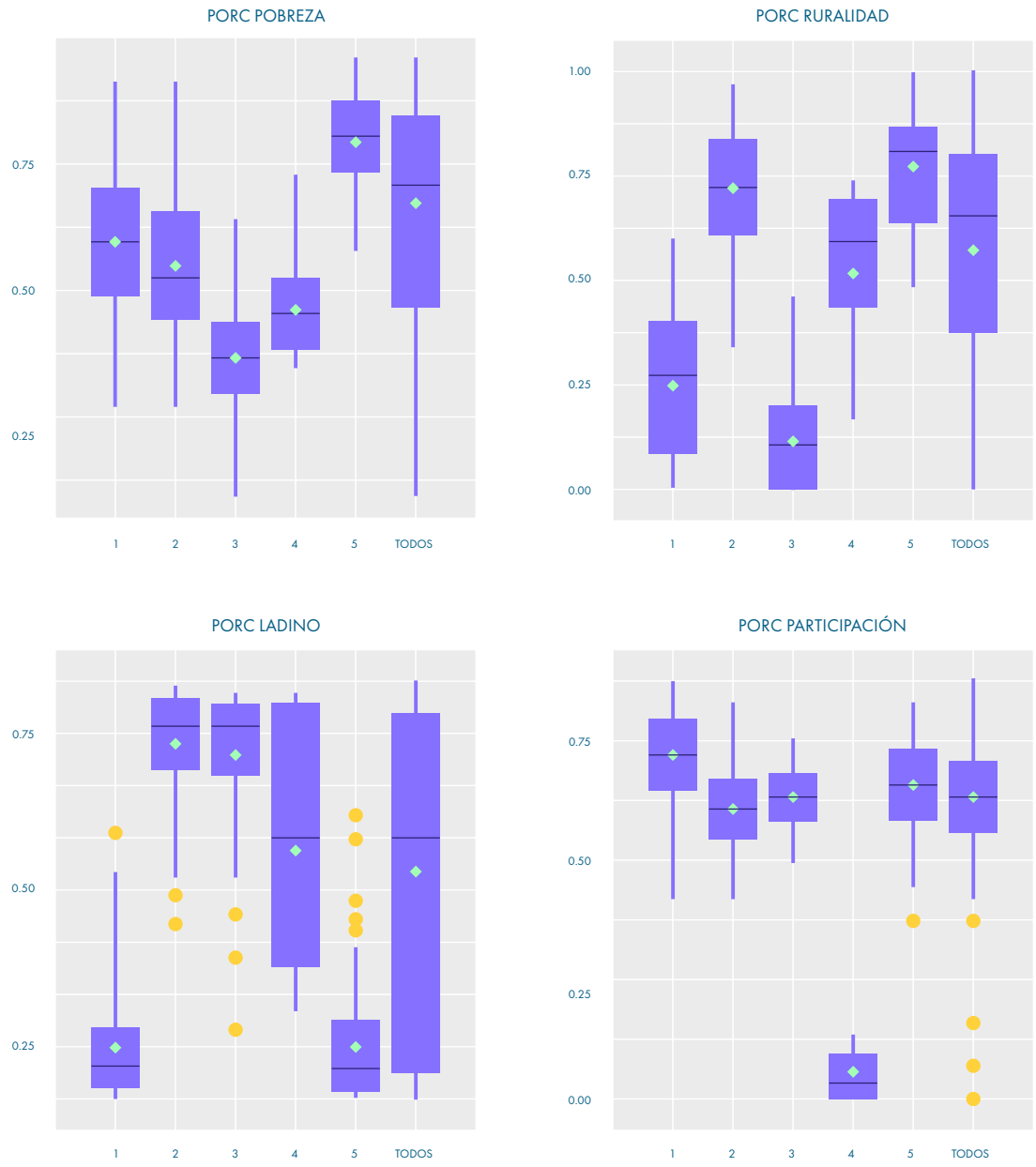
30. Por ejemplo, si los 3.60 millones de votos se repartieran uniformemente (0.90 millones) para cuatro partidos que no recibieron absolutamente ningún otro voto, estos cuatro habrían encabezado la contienda. Incluso, alternativamente, estos 3.60 millones de votos habrían logrado que Semilla ganase en primera vuelta con 0.65+3.60=4.65 millones de votos (59.8% de los potenciales 7.77 millones de votos válidos emitidos).
31. Calculado con la base de datos de resultados del TSE, extraídos el 27 de junio a las 11pm de: https://www.trep.gt/#!/tc1/db/

**Mapa M3:** Municipios de Guatemala según su segmentación en clusters a partir de variables de vulnerabilidad y porcentaje de participación en las elecciones generales y de diputados al Parlamento Centroamericano 2023




**Fuente:** elaboración propia con datos del TSE (2023), INE (2018) y ENCOVI (2014)

**Gráfica G4:** Diagramas de caja[32] y bigote por cluster de cada variable utilizada para generar la categorización: porcentaje de pobreza, porcentaje de ruralidad, porcentaje de población ladina y porcentaje de participación en las Elecciones Generales y de Diputados al Parlamento Centroamericano




**Fuente:** elaboración propia con datos de PNC (2019-2022), INE (2018) y ENCOVI (2014)

32. Nota: el rombo rojo representa el promedio de las distribuciones, mientras que la línea horizontal en cada caja ilustra la mediana. Las cajas encierran al 50% de los datos de cada distribución (aquellos comprendidos entre el quintil 1 y 3) mientras que los puntos representan datos atípicos. En los anexos pueden encontrarse tablas con las magnitudes numéricas de las distribuciones.

Los clusters resultantes de este caso resultan con perfiles similares a los de los casos anteriores, pero con una clara diferencia: en lugar de tener dos clusters para municipios rurales ladinos, se tiene solo uno ("2"); y hay un cluster ("4") que solo agrupa municipios con valores atípicos en su variable de interés (porcentaje de participación en las elecciones). De esa manera, los clusters resultantes son:

Cluster "1": 47 municipios urbanos e indígenas (en promedio, 24.9% de habitantes en áreas rurales y autoidentificación ladina de 12.3%) con el mayor porcentaje promedio de participación en las elecciones de 2023 (71.5%). La mayoría de los municipios de este cluster se ubican en el altiplano guatemalteco y sus municipios con mayor padrón son San Juan Sacatepéquez, Cobán, Jalapa, Totonicapán y Sololá. En total, en este cluster 0.68 millones de personas votaron en las elecciones generales de 2023, y 0.37 millones de personas no asistieron a las urnas.

Cluster "2": agrupa a 132 municipios con población predominantemente ladina y rural[33], tal como a Coatepeque, Malacatán, Morales, Santa Lucía Cotzumalguapa y Chiantla. En promedio, los municipios de este cluster presentan 56.0% de habitantes bajo condiciones de pobreza, 69.6% de habitantes que residen en áreas rurales y 87.3% de habitantes que se autoidentifican como ladinos. Los municipios de este cluster tuvieron una participación promedio de 61.0% en las elecciones y en total sumaron 1.44 millones de votos emitidos y 1.04 millones de personas que no participaron.

33. En casos anteriores, muchos de estos 132 de clusters se repartían en dos clusters distintos de municipios ladinos y rurales.

Cluster "3": los municipios urbanos, ladinos y poco pobres que corresponden a cabeceras departamentales y municipios de la región central del país, tales como Guatemala, Mixco, Villa Nueva, Quetzaltenango y Escuintla. Estos municipios tienen un porcentaje promedio de pobreza 30.1%, 11.4% de población rural, 83.7% de población ladina y 63.2% de participación en las elecciones. Es el cluster con el mayor número de empadronados, de los cuáles 1.92 millones de personas votaron y 1.29 no lo hicieron.

Cluster "5": en contraposición al cluster "3", este posee a 105 municipios rurales, indígenas y muy pobres, que están ubicados principalmente en Alta Verapaz, Quiché y Huehuetenango. Estos poseen porcentaje promedio de población en condiciones de pobreza, que reside en áreas rurales y que se auto identifica como ladina de 79.9%, 78.5% y 11.5%, respectivamente. De este cluster, se presentaron 1.58 millones de personas a votar y 0.92 millones de personas se abstuvieron, y los municipios tienen un promedio de participación de 65.4%.

El cluster "4" consiste de los 4 municipios que aparentemente tienen una participación atípicamente baja. Estos municipios son San José del Golfo del departamento de Guatemala; San Pedro Yepocapa, de Chimaltenango; San Pablo Jocopilas, de Suchitepéquez y San Martín Zapotitlán, de Retalhuleu. En cada uno de ellos se suspendieron las elecciones por disturbios o se declararon nulos sus resultados debido a irregulares, por lo que aparece en el portal de resultados que solo existían 3,601 votos de un padrón de 52 mil[34].

34. Fuente: https://www.prensalibre.com/guatemala/elecciones-generales-guatemala-2023/municipios-en-los-que-se-repetiran-las-elecciones-que-ocurrio-en-los-lugares-donde-el-20-de-agosto-se-votara-de-nuevo-por-alcalde/. San Bartolomé Jocotenango es el quinto municipio en el que se repetirán elecciones debido a disturbios en el proceso. Este municipio registró una participación del 37% y se agrupó bajo el cluster "5"(es el dato atípico que este cluster tiene en el diagrama de caja y bigote de la variable de porcentaje de participación).

Otra diferencia que estos resultados tienen con los ejercicios anteriores de clusters con las variables de incidencia de COVID-19 y tasa de homicidios es que la expresión de la problemática en cuestión (porcentaje de participación en las elecciones) no es significativamente diferente entre cada grupo de municipios. Con excepción del cluster atípico "4", la distribución municipal del porcentaje de participación de los otros clusters están superpuestas una encima de los otras. El único cluster con más participación fue el "1", compuesto por municipios urbanos e indígenas. Una posible hipótesis para explicar este fenómeno es que en esos municipios hay más organización popular y, por ende, una mayor asistencia a las urnas. No obstante, las diferencias en participación entre esos municipios y el resto del país no están tan marcadas, sugiriendo que las variables de nivel de pobreza, área geográfica y auto-identificación étnica de la población no son tan diferenciadoras en la decisión de votar o no hacerlo. Por tal razón, el diseño de una estrategia para promover el empadronamiento no se podría priorizar únicamente a raíz de los resultados de este ejercicio, sino deberían realizarse acciones para atender a los municipios con menos participación a nivel nacional siguiendo directrices especializadas y pertinentes al cluster al que corresponda cada uno.

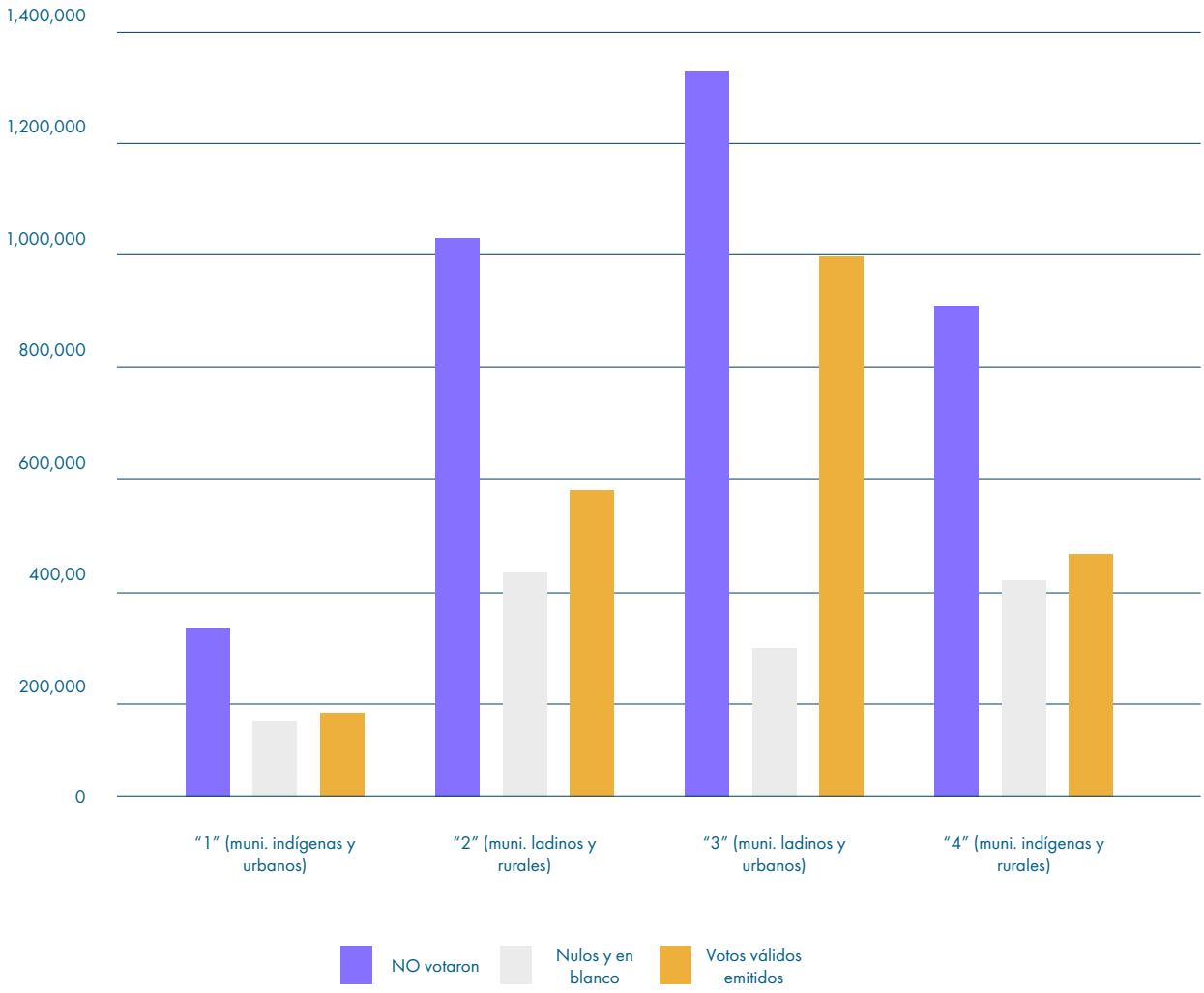
Para llegar a evaluar los factores asociados a la participación cívica que se traduce en el porcentaje de ciudadanía que asistió a las urnas, sería esencial "considerar múltiples factores, incluidos el clima político general, la accesibilidad a las elecciones, la educación de los votantes y la presencia de instituciones democráticas robustas" (Wisevoter, 2023)[35]. En particular, podría buscarse la relación (por cluster) entre el porcentaje de participación en las elecciones y variables que reflejen el ambiente político a nivel municipal: qué tan disputadas se perciben las elecciones a alcalde o qué tan incierto se percibe su resultado; cuánta desigualdad socioeconómica y/o cuánta conflictividad social hay en el municipio; cuáles son los partidos con más afiliados en los municipios (e.g., algunos partidos podrían ser más efectivos para movilizar sus bases locales); cuánto desgaste ha sufrido la imagen el actual alcalde o alcaldesa en el último período de 4 años; si el alcalde o alcaldesa actual es del mismo partido que la dupla presidencial; cuánto ha invertido el TSE y otras entidades estatales en campañas de sensibilización sobre importancia del voto y en

35. Recopilado de htts://wisevoter.com/country-rankings/voter-turnout-by-country/

fiscalización de partidos, entre otras. Otro punto que puede ser estudiado son las diferentes influencias de medios de comunicación sobre la participación electoral: por ejemplo, puede suponerse que municipios rurales (clusters "2" y "5") reciben mayor influencia de medios de comunicación tradicionales como televisión por cable y radio, mientras que municipios urbanos (clusters "2" y "5") tienen mayor acceso a teléfonos inteligentes e internet por lo que son más influidos por redes sociales como Twitter, Facebook y TikTok, por lo que podrían hacerse análisis de sentimientos hacia las elecciones por cada medio distinto por cluster y estudiar si ello se relaciona con los resultados de participación electoral. De cualquier manera, la generación de clusters suscita subsiguientes preguntas de investigación que a su vez pueden llevar a una mejor comprensión de los fenómenos en cuestión.

Por otro lado, la generación de estos clusters por variables socioeconómicas y participación puede aprovecharse para otros fines, como para profundizar sobre el análisis de los resultados de las elecciones. Por ejemplo, ayudan a retratar en qué perfiles de municipios hay potencialmente más desvinculación entre la ciudadanía y su actuar cívico y/o procesos democráticos - o desde la perspectiva de partidos políticos, dónde deben redoblar esfuerzos para llegar a e incluir en su propuesta política a distintos grupos de personas. La gráfica G5 muestra la cantidad bruta de personas que no participaron en las Elecciones Generales y de Diputados al Parlamento Centroamericano de 2023 comparado con la cantidad de votos inválidos (nulos y en blanco) y votos válidos emitidos para la elección de dupla presidencial, y en ella se puede observar que para todos los clusters hubo más personas que se abstuvieron de votar que votos válidos. También se puede observar que, para los clusters con población predominantemente indígena, la cantidad de votos nulos y blancos es casi igual en magnitud a la de los votos válidos, ilustrando el gran impacto que estos tuvieron sobre los resultados. La decisión de votar nulo o en blanco y la decisión de no votar podrían estar vinculadas, y podrían darse por un conjunto de factores distinto para cada cluster. Una hipótesis es que para municipios urbanos e indígenas (cluster "1"), el voto nulo fue impulsado por la descalificación de la dupla presidencial del partido Movimiento de Liberación de los Pueblos (MLP), mientras que para municipios rurales y ladinos (cluster "2") el voto nulo fue impulsado por la descalificación de la dupla del partido Prosperidad Ciudadana.

**Gráfica G5:** Cantidad de personas empadronadas que no asistieron a las urnas, cantidad de votos nulos y blancos y cantidad de votos válidos emitidos por cluster*, elecciones presidenciales del 25 de julio de 2023.

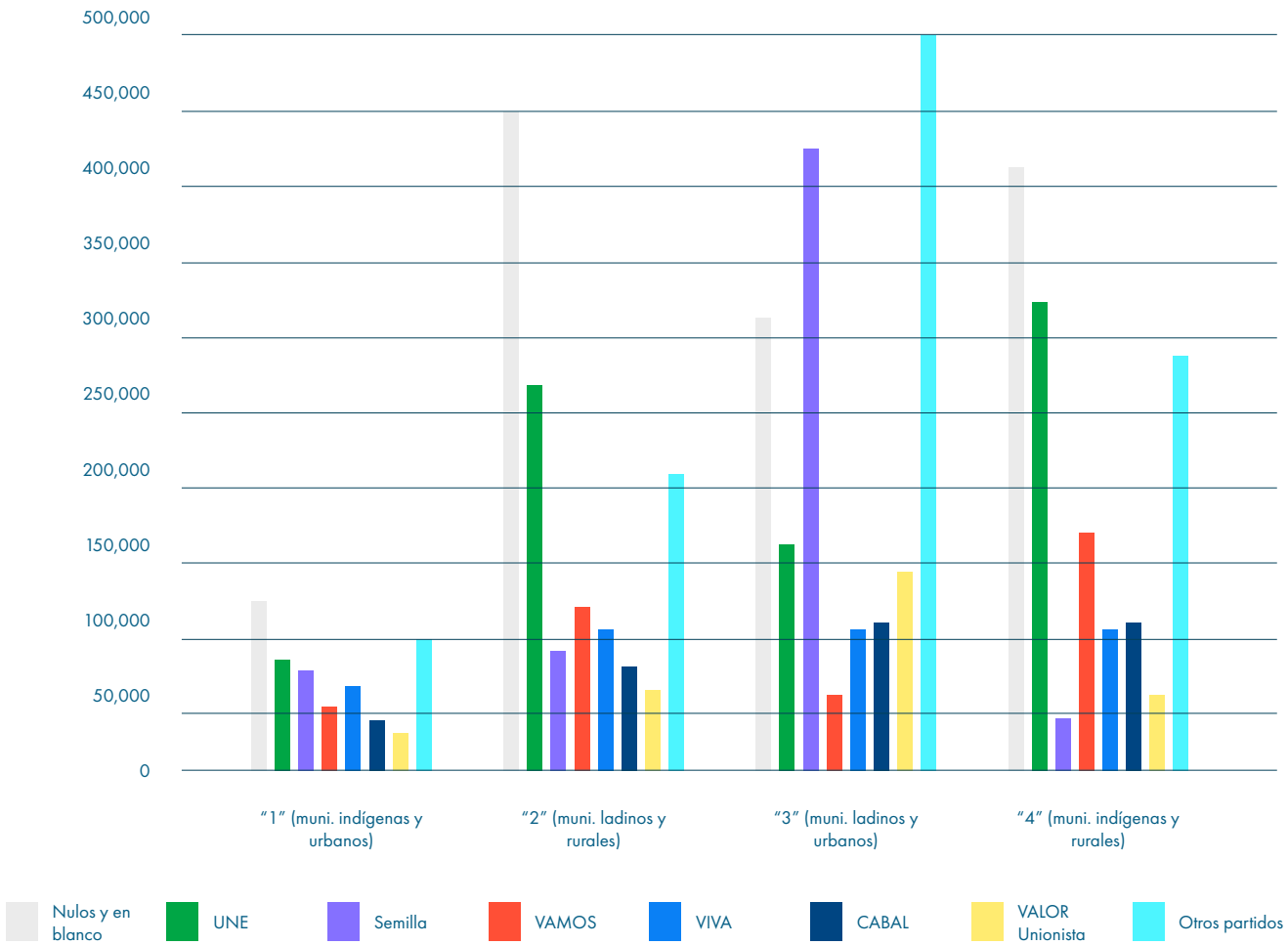



**Fuente:** elaboración propia

*Nota: se excluye el cluster "4" debido a los resultados suspendidos o anulados en sus municipios.

La gráfica 6, por su parte, muestra la cantidad de votos válidos que recibió por cluster cada agrupación política, permitiendo otra lectura a los resultados electorales. De ella, puede observarse como el partido Semilla está fuertemente representado en municipios urbanos y ladinos (siendo el único partido que superó la cantidad de votos nulos y en blanco en ese cluster), pero muy subrepresentado en municipios indígenas y rurales (lo mismo sucede en el caso de la coalición Valor-Unionista). Contrariamente, los partidos UNE y VAMOS está subrepresentados en municipios urbanos, pero compensan con una alta recepción de votos en municipios rurales, mientras que los partidos VIVA y CABAL muestran un desempeño relativamente estable en cada cluster.

**Gráfica G6:** Resultado de las elecciones presidenciales del 25 de julio de 2023, cantidad de votos válidos por cluster*




**Fuente:** elaboración propia

*Nota: se excluye el cluster "4" debido a los resultados suspendidos o anulados en sus municipios.

Esta lectura de resultados podría ser útil para los partidos de UNE y Semilla en el contexto de la segunda vuelta. Comprendiendo cuáles fueron las tendencias en los clusters para la primera vuelta, pueden orientar esfuerzos más efectivos para llegar a acuerdos y dar a conocer sus propuestas a territorios clave para aumentar sus probabilidades de ser electos. Los clusters "2" y "5" (municipios rurales tanto indígenas como ladinos) recibieron mayor porcentaje de votos nulos: suponiendo que conlleva menos trabajo transformar votos nulos en votos válidos que transformar no-asistencias en votos válidos, estos deberían ser los municipios rurales en los que Semilla y UNE deberían dedicar esfuerzos para la segunda vuelta. El partido VAMOS también consiguió una relativa alta cantidad de votos en estos municipios rurales debido a la afiliación de varias alcaldías a ese partido[36], por lo que las alianzas que se formen con este partido también tienen el potencial de impactar las elecciones presidenciales vía corporaciones municipales electas.

## 5. Discusión de resultados

Al construir los clusters de municipios empleando el algoritmo de k-means con k = 5 y utilizando las tres variables de vulnerabilidad (porcentaje de población en condiciones de pobreza, porcentaje de población que se autoidentifica como ladina y porcentaje de población que reside en áreas rurales) junto a otra variable de interés (incidencia de COVID-19, tasa de homicidios o porcentaje de participación en elecciones), las agrupaciones generadas resultaron seguir un patrón similar, principalmente atribuible a las primeras tres variables (ya que no existe correlación entre las variables de interés). Aunque existen ciertas variaciones a lo interno de los cluster para cada caso, de manera general, las agrupaciones resultantes poseen ciertas características socioeconómicas similares, que se les ha designado como perfiles:

- Un cluster compuesto por municipios donde la mayoría se autoidentifica como indígena y reside en áreas urbanas, con niveles de pobreza medios.
- Un cluster de municipios donde la mayoría de la población se autoidentifica como ladina y reside en áreas urbanas, con niveles de pobreza bajos.

36. Ver Kestler, González, Morales Y Sanz (2023). Así construyó Giammattei su ejército de alcaldes para las elecciones 2023. No Ficción. Nota periodística publicada el 15 de junio de 2023 en: https://www.no-ficcion.com/projects/giammattei-estrategia-alcaldes-elecciones2023

- Un cluster donde la mayoría de la población es indígena y reside en áreas rurales, y donde se registran los niveles de pobreza más elevados.
- Uno (en el caso de participación en elecciones) o dos (en el caso de incidencia de Covid-19 y tasa de homicidios) clusters de población mayoritariamente ladina y que reside en áreas rurales, con niveles de pobreza medios. Para los casos de dos clusters, estos estarían diferenciados entre sí esencialmente por la variable de interés (por ejemplo, en el caso de tasa de homicidios, un cluster de municipios rurales y ladinos con pocos o ningún homicidio y otro cluster de municipios rurales y ladinos con una tasa de homicidios elevada).

A lo largo de los tres ejercicios, 306 municipios (equivalente al 90% de todos ellos) fueron agrupados bajo el mismo perfil. En particular, los municipios que menos variaciones tuvieron en su agrupación fueron los rurales, indígenas y muy pobres que yacen principalmente en la franja transversal del norte del país (93.5% se clasificaron en el perfil a lo largo de los 3 ejercicios). Esto es probable que se deba a que las variaciones a lo largo de la variable de interés (incidencia de COVID-19, tasa de homicidios y porcentaje de participación en elecciones) entre los municipios de este cluster sean mínimas en relación a las diferencias en las variables socioeconómicas que estos municipios tienen con los municipios de otros perfiles. Por el otro lado, los municipios del perfil urbano, ladino y poco pobre que consisten en cabeceras departamentales y la región central solo se mantuvieron constante el 73.6% del tiempo, posiblemente indicando que estos tuvieron mayores diferencias entre sí en la expresión de las variables de interés. La tabla T1 muestra un comparativo de las características de los perfiles resultantes en los tres ejercicios, así como cuántos municipios fueron clasificados bajo el mismo perfil a lo largo de ellos.

**Tabla 1:** omparación de los clusters construidos a través de las tres variables de vulnerabilidad socioeconómicas y variables de interés: incidencia de casos COVID-19 (número "1" en la tabla), tasa de homicidios ("2"), y porcentaje de participación ("3")

| Perfil | Cluster Equivalente | | | Cantidad de municipios | | | Promedio % Rural | | | Promedio % Pobreza | | | Promedio % Población Ladina | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Rural - Indigena | "5" | "4" | "5" | 107 | 105 | 105 | 76.9 | 77.2 | 78.5 | 80 | 79.8 | 79.9 | 10.2 | 10.3 | 11.5 |
| Rural - Ladino - Pobre | "3" "4" | "2" "3" | "2" | 145 | 138 | 132 | 66.9 | 69.5 | 69.6 | 55.3 | 56.8 | 56.0 | 87.2 | 85.1 | 87.3 |
| Urbano - Indígena - Pobre | "1" | "1" | "1" | 46 | 45 | 46 | 21.5 | 21.0 | 24.9 | 56.1 | 55.8 | 58.2 | 17.2 | 15.5 | 12.3 |
| Urbano - Ladino - Poco Pobre | "2" | "5" | "3" | 42 | 52 | 53 | 11.0 | 14.3 | 11.4 | 26.1 | 29.4 | 30.1 | 82.4 | 85.7 | 83.7 |
| Todos | TO DOS | TO DOS | sin "4" | 340 | 340 | 336 | 57.0 | 57.0 | 57.1 | 59.6 | 59.6 | 59.7 | 52.9 | 52.9 | 52.8 |

| Perfil | Promedio variable de interés* | | | Municipios en el mismo perfil a lo largo de los tres ejercicios | % de municipios en el mismo perfil a lo largo de los tres ejercicios** |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | | |
| Rural - Indígena - Muy Pobre | 0.55 | 1.14 | 63.3 | 100 | 93.5 |
| Rural - Ladino - Pobre | 1.28 | 6.18 | 60.1 | 127 | 87.6 |
| Urbano - Indígena - Pobre | 1.11 | 1.33 | 69.6 | 40 | 87.0 |
| Urbano - Ladino - Poco Pobre | 4.00 | 6.50 | 61.6 | 39 | 73.6 |
| Todos | 1.36 | 4.03 | 62.6 | 306 | 90.0 |

**Fuente:** elaboración propia

*Para el ejercicio 1 se muestra el promedio de la tasa de incidencia de Covid-19; para el ejercicio 2, el promedio de la tasa de homicidios y para el caso 3, el promedio del porcentaje de población.

**Este porcentaje tomó como denominador a la cantidad de municipios más grande a lo largo de los tres ejercicios para dar las estimaciones más conservadoras.

Aunque perfil rural-indígena-muy pobre es el que está sujeto a la mayor vulnerabilidad bajo el presente marco teórico y el perfil urbano-ladino-poco pobre a la menor, cada uno de estos perfiles posee distintos desafíos y diferentes expresiones de las problemáticas que sus poblaciones enfrentan. La interrelación entre los distintos perfiles de vulnerabilidad socioeconómica y las problemáticas son no-triviales. Por ejemplo, el perfil rural-ladino-pobre tiene la mayor incidencia de COVID-19 y el perfil rural-indígena-pobre extremo tiene la menor, pero cada territorio se enfrenta a barreras de acceso a la salud muy diferentes (logísticas, económicas, comunicacionales, etc.). En cualquier caso, el desarrollo de clusters particulares para diversas problemáticas permite hacer un primer acercamiento al estudio de estas diversas problemáticas, la generación de estrategias para afrontarlas, el desarrollo de acciones específicas para ciertos contextos y la mitigación de vulnerabilidad en todo el territorio nacional.

> *« La desigualdad en la cobertura vacunal para la COVID-19 replica la desigualdad estructural del país, y que ésta se media, muy probablemente, por los déficits de acceso a los servicios de salud y a condiciones dignas de vida en estas poblaciones, que no facilitan el acceso físico, económica y cultural a la vacunación. La desigualdad estructural, de corte socioeconómico, se intersecta con el racismo y discriminación, también estructural para generar las desigualdades en la vacunación. »*
>
> *(Slowing, K. y Chávez, O. 2022)*

Al no reconocer las diferentes expresiones de vulnerabilidad y la heterogeneidad del territorio a la hora de abordar un problema, puede resultar en enfoques mono-causales (tales como, por ejemplo, solo priorizar municipios urbanos-ladinos-poco pobres en el caso del COVID-19 porque ahí es donde se concentran los datos oficiales de casos confirmados) que profundizan las desigualdades estructurales. Lo mismo podría suceder al no adoptar un enfoque interseccional para abordar la interrelación entre factores de exclusión.  Ello es una advertencia de la necesidad de incluir la lectura de la vulnerabilidad multidimensional en el análisis de estas problemáticas y de no atribuir una causalidad entre dos variables sin el debido rigor.

Otra potencial aplicación de los clusters se ilustra en el ejercicio 4.3 de participación en las elecciones. En ella, se generaron los clusters según el porcentaje de personas en torno a variables socioeconómicas y al porcentaje de personas que asistieron a las urnas, pero posteriormente se analizaron internamente para hacer una lectura de los resultados electorales dentro de cada uno, logrando identificar tendencias para los distintos grupos de municipios y suscitando próximas preguntas de investigación. Ello ilustra la capacidad que tienen los algoritmos de aprendizaje de máquina no supervisado de realizar análisis exploratorios que puedan dar con hallazgos más sofisticados que simples medidas de tendencia central de la estadística descriptiva tradicional.

En este esfuerzo de agrupación de municipios se decidió priorizar las variables de vulnerabilidad socioeconómica, con el fin de generar grupos de municipios congruentes a lo interno en torno a estas variables para facilitar el diseño de intervenciones. La variable de porcentaje de pobreza utilizada es una aproximación a la posesión de activos, capacidad de consumo y capacidad de responder a shocks de los hogares en cada municipio, que además es proporcional a la escolaridad de la población. La variable de porcentaje de población rural fue seleccionada porque las variables de interés (en particular, incidencia de COVID-19 y tasas de homicidio) se supone a priori que tienen expresiones muy distintas dependiendo de la densidad poblacional y acceso a infraestructura, pero también fue seleccionada bajo el supuesto que áreas rurales reciben menos provisión de bienes y servicios públicos por parte de entidades estatales.  La variable de porcentaje de población ladina, si bien estaría suponiendo una categorización sobre simplificada de la adscripción étnica en el territorio nacional (en esencia, supone una dicotomía entre identidad ladina e indígena), fue incluida ya que ella es un acercamiento a la pertinencia social y cultural que debe abordarse para la lectura de estas problemáticas y propuesta de soluciones. Aunque estas tres variables se usaron como un primer acercamiento para comprender las vulnerabilidades desiguales en el territorio nacional a priori, cabe una discusión más a profundidad de si ellas son las más apropiadas para clasificar a los municipios en diferentes grupos.

Incluso, podría considerarse en próximos ejercicios utilizar estas variables más otras adicionales para robustecer la lectura de vulnerabilidades (como, por ejemplo,

índices de brecha de género o de presencia gubernamental), pero debe prestarse atención a que la inclusión de más variables reduciría el peso relativo que cada una (incluyendo la variable de interés de la problemática) tiene sobre la clasificación final de los municipios. Como alternativa para ello, podrían utilizarse índices compuestos de varias variables juntas, donde cada variable posee un determinado peso, para la generación de clasificaciones. No obstante, una característica de los algoritmos no supervisados como el de k-means es que no hay una única forma objetiva de determinar cuáles son las variables y pesos óptimos para realizar los ejercicios, por lo que hay lugar para experimentar con distintas combinaciones hasta llegar a los resultados más útiles para el análisis de la problemática.

Otro punto a considerar es que estos clusters se construyeron a un nivel de detalle municipal. Pero así como existen variaciones a nivel nacional, regional y/o departamental, que tratamos de superar con este esfuerzo, pueden existir a lo interno de cada municipio una multitud de desigualdades y diferencias en la expresión de problemáticas sociales. A modo de ejemplo, el 71.3% de los habitantes del municipio de Cobán se encuentran en situación de pobreza[37], pero al introducir grupo étnico, de la población que se autoidentifica como ladina (que reside primordialmente en el casco urbano de Cobán) solo el 14.6% se encuentra en esta situación, comparado con el 80.6% de la población que se autoidentifica como indígena que está en ella. Sin embargo, la presente metodología podría aplicarse al análisis de un problema a nivel de lugar poblado dado el caso que se cuente con la información. A mayor desglose de los datos a nivel territorial (e.g., departamento – municipio - lugar poblado), mayor será el nivel de análisis de clusters y las subsiguientes estrategias lograrán acciones más pertinentes a los desafíos y fortalezas en el territorio local. Del mismo modo, mientras más actualizada y frecuente sea la información con la que se realizan estos estudios, se desarrollará un entendimiento que mejor refleje los problemas socioeconómicos y el impacto que el contexto y las políticas públicas tienen sobre ellos.

37. Según el indicador de Figueroa et al. (2020) que se ha utilizado en este estudio.

# 6. Conclusión

- Guatemala es un país con profundas desigualdades que se traducen a distintos niveles de exposición y vulnerabilidades en cuanto a problemáticas socioeconómicas que a su vez retroalimentan las desigualdades existentes. Es necesario realizar los análisis de estas problemáticas considerando esta vulnerabilidad heterogénea a lo largo de todo el territorio. A primera instancia, la inclusión de la situación de pobreza, el grupo étnico y el área geográfica de la población al estudio de las problemáticas en los distintos municipios podría lograr una comprensión más integral e interseccional de las mismas.
- Se utilizó un algoritmo de clustering para segmentar a los municipios según su porcentaje de pobreza, población ladina y población en áreas rurales y otra variable asociada a un problema socioeconómico: incidencia de COVID-19, tasa de homicidios y porcentaje de población que participó en las elecciones. Con ello se logró crear agrupaciones de municipios con similares expresiones de vulnerabilidad y de esa problemática. Estas agrupaciones son un insumo para orientar el desarrollo de estrategias para mitigar estas problemáticas al facilitar el desarrollo de directrices de acción diferenciadas según las características de los municipios, al contribuir al planteamiento de subsiguientes preguntas de investigación y favorecer una mirada multidimensional de dinámicas y fenómenos observados.A lo largo de la creación de clusters con las tres variables asociadas a las problemáticas, se encontró que se repitieron 4 tipos de agrupaciones: municipios urbanos, indígenas y pobres; municipios urbanos, ladinos y poco pobres; municipios rurales, indígenas y muy pobres; y, por último, una o dos agrupaciones distintas de municipios rurales y ladinos. De ello se observan 4 perfiles de vulnerabilidad entre los municipios del territorio nacional según su promedio de pobreza, adscripción étnica y áreas geográficas para el análisis de la incidencia de COVID-19, la tasa de homicidios y el porcentaje de participación en elecciones.
- Es importante incorporar el uso de herramientas de investigación innovadoras tal como algoritmos de aprendizaje de máquina (entre ellos los no supervisados como el clustering con k-means), aprovechando la creciente oferta tecnológica y de datos, para ir mejorando la comprensión territorial de las problemáticas que la población guatemalteca enfrenta y afrontar la vulnerabilidad heterogénea inherente a estos desafíos.

# 7. Bibliografía

Figueroa Chávez, W.; Peñate, M.; Marsicovetere, P. Estimación de Pobreza a Nivel Municipal en Guatemala Mediante la Utilización de Machine Learning. 2020. Available online: https://www.researchgate.net/publication/343678849_Estimacion_de_pobreza_a_nivel_municipal_en_Guatemala_mediante_la_utilizacion_de_machine_learning (accessed on 5 December 2022).

Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21 (3): 768–769. JSTOR 2528559.

Gu, X., Angelov, P. P., Kangin, D., & Principe, J. C. (2017). A new type of distance metric and its use for clustering. En Evolving Systems (Vol. 8, Issue 3, pp. 167–177). Springer Science and Business Media LLC. https://doi.org/10.1007/s12530-017-9195-7

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means Clustering Algorithm. En Applied Statistics (Vol. 28, Issue 1, p. 100). JSTOR. https://doi.org/10.2307/2346830

Lloyd, Stuart P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, Stuart P. (1982). "Least squares quantization in PCM" (PDF). IEEE Transactions on Information Theory. 28 (2): 129–137. CiteSeerX 10.1.1.131.1338. doi:10.1109/TIT.1982.1056489. S2CID 10833328. Retrieved 2009-04-15.

Pizarro, R. (2001). "La vulnerabilidad social y sus desafíos, una mirada desde América Latina", CEPAL, División de Estadística y Proyecciones Económicas (LC/L. 1490-P). Santiago de Chile, Chile.

Ruiz Rivera, N. (2012) "La definición y medición de la vulnerabilidad social. Un enfoque normativo". En: Investigaciones Geográficas, boletín del Instituto de Geografía, Universidad Nacional Autónoma de México, México.

Slowing, K. y Chávez, O. (2022). "Vacunación COVID-19 y poblaciones vulnerables: Desigualdad y barreras institucionales (MSPAS) de acceso a la vacunación" https://labdedatosgt.com/estudio-guatemala-desigualdad-vacuna-covid/

Slowing, K., Chávez, O. (2022). Diagnóstico: Distribución desigual de vacunas y tratamientos del COVID-19 en poblaciones vulnerables. OXFAM-Laboratorio De Datos GT, Guatemala.

Slowing, K., Chávez, O., Maldonado, E., & García, A. L. (2021). Propuesta para fortalecer la aplicación equitativa de la vacuna contra COVID-19. Guatemala.

Virmani, D., Taneja, S., & Malhotra, G. (2015). "Normalization based K means Clustering Algorithm (Versión 1). https://doi.org/10.48550/ARXIV.1503.00900

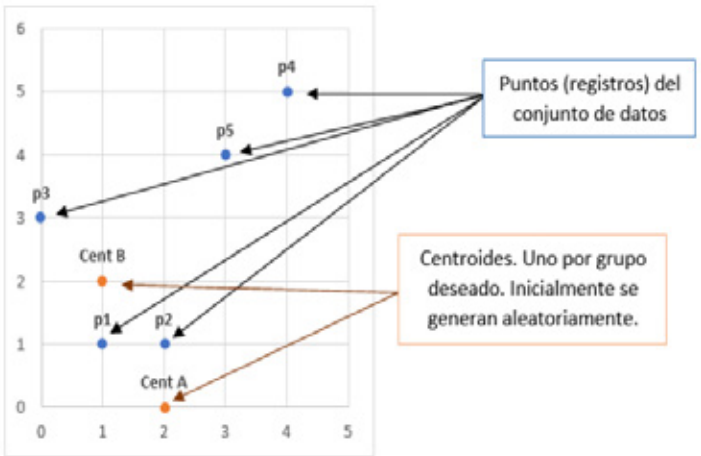
# 8. Anexos

**Figura A1.** Captura de pantalla de un extracto del conjunto datos utilizado para generar clusters

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | coddepartamento | departamento | codmunicipio | municipio | porc_pobreza | porc_ruralidad | porc_ladino | incidencia_casos | tasa_homicidios | porc_participacion |
| 2 | 01 | Guatemala | 101 | Guatemala | 0.1076 | 0.0000 | 0.9132 | 7.0972 | 11.9566 | 0.5745 |
| 3 | 01 | Guatemala | 102 | Santa Catarir | 0.1472 | 0.1186 | 0.9451 | 5.5617 | 12.5445 | 0.6415 |
| 4 | 01 | Guatemala | 103 | San José Pin | 0.2203 | 0.1551 | 0.9497 | 4.7076 | 8.8695 | 0.6280 |
| 5 | 01 | Guatemala | 104 | San José del | 0.2312 | 0.6775 | 0.9775 | 2.2845 | 0.0000 | 0.0000 |
| 6 | 01 | Guatemala | 105 | Palencia | 0.3981 | 0.5518 | 0.9779 | 2.9903 | 9.3190 | 0.5312 |
| 7 | 01 | Guatemala | 106 | Chinautla | 0.2215 | 0.0853 | 0.8177 | 3.4799 | 5.5883 | 0.4884 |
| 8 | 01 | Guatemala | 107 | San Pedro Ay | 0.3436 | 0.1687 | 0.7436 | 2.2803 | 13.0679 | 0.5555 |
| 9 | 01 | Guatemala | 108 | Mixco | 0.1148 | 0.0059 | 0.8935 | 4.4926 | 8.7292 | 0.6262 |
| 10 | 01 | Guatemala | 109 | San Pedro Sa | 0.4025 | 0.2800 | 0.2347 | 2.7278 | 0.0000 | 0.7762 |
| 11 | 01 | Guatemala | 110 | San Juan Sac | 0.4852 | 0.2851 | 0.3627 | 1.2891 | 5.9078 | 0.5695 |
| 12 | 01 | Guatemala | 111 | San Raymun | 0.4839 | 0.5112 | 0.2846 | 1.3621 | 0.0000 | 0.6340 |
| 13 | 01 | Guatemala | 112 | Chuarrancho | 0.6624 | 0.3959 | 0.1359 | 0.8274 | 0.0000 | 0.6591 |
| 14 | 01 | Guatemala | 113 | Fraijanes | 0.2143 | 0.1889 | 0.9308 | 3.7762 | 10.8606 | 0.7029 |
| 15 | 01 | Guatemala | 114 | Amatitlán | 0.2177 | 0.1589 | 0.9580 | 1.7915 | 5.8409 | 0.5599 |
| 16 | 01 | Guatemala | 115 | Villa Nueva | 0.1371 | 0.0171 | 0.9410 | 2.9520 | 9.1508 | 0.5788 |
| 17 | 01 | Guatemala | 116 | Villa Canales | 0.2874 | 0.1978 | 0.9627 | 2.4169 | 7.1171 | 0.5947 |
| 18 | 01 | Guatemala | 117 | San Miguel P | 0.1077 | 0.0467 | 0.9342 | 3.3480 | 2.6231 | 0.6094 |
| 19 | 02 | El Progreso | 201 | Guastatoya | 0.2511 | 0.0000 | 0.9701 | 4.3649 | 14.4975 | 0.6510 |
| 20 | 02 | El Progreso | 202 | Morazán | 0.4634 | 0.7880 | 0.9845 | 0.8633 | 7.8740 | 0.5969 |

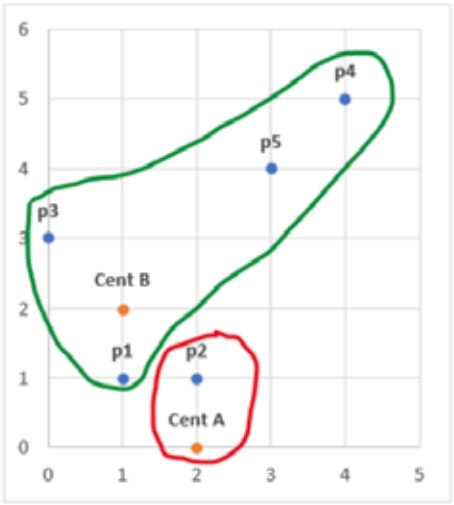
En las gráficas A2, A3 y A4 se ejemplifica la progresión de k-means en un conjunto de 5 datos, con dos clusters y distancia euclidiana.

**Gráfica A2.** Ilustración gráfica de algoritmo de k-means para hacer agrupaciones con un conjunto de 5 datos de 2 variables con k=2, empleando la distancia euclidiana.




**Fuente:** elaboración propia

**Gráfica A3.** Ilustración gráfica del algoritmo de k-means, pasos 3, 4 y 5: se mide la distancia de cada punto a los centroides para determinar a cuál se asocia cada punto y se calcula la posición de los nuevos centroides.

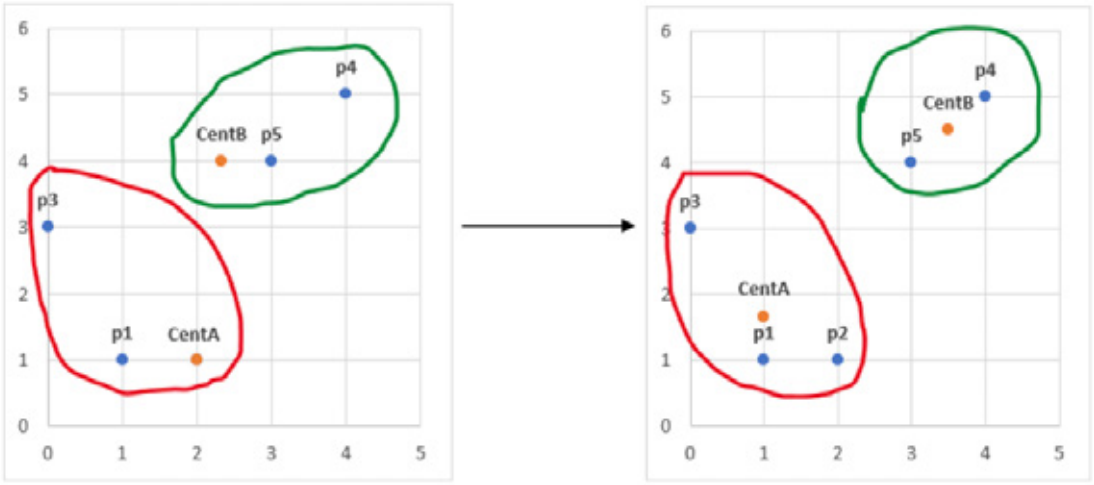



| Distancia a CentA | Distancia a CentB | Centroide más cercano |
|---|---|---|
| 1.41 | 1.00 | B |
| 1.00 | 1.41 | A |
| 3.61 | 1.41 | B |
| 5.39 | 4.24 | B |
| 4.12 | 2.83 | B |

Nuevos centroides
A: Promedio(p2) = (2,1)
B: Promedio(p1, p3, p4, p5) = (2.33, 4)

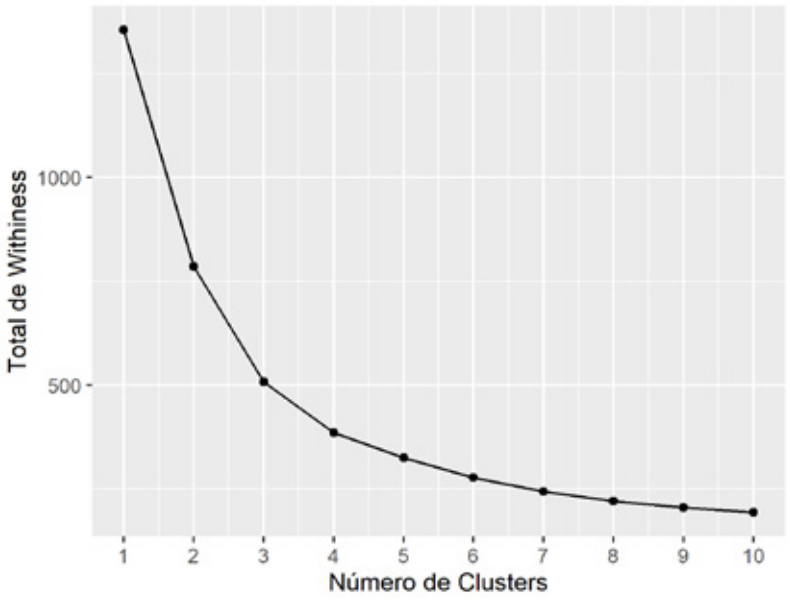
**Fuente:** elaboración propia

**Gráfica A4.** Ilustración gráfica de algoritmo de k-means. Se reitera el procedimiento utilizando los nuevos centroides calculados hasta llegar a agrupaciones que ya no varíen entre una iteración y otra.



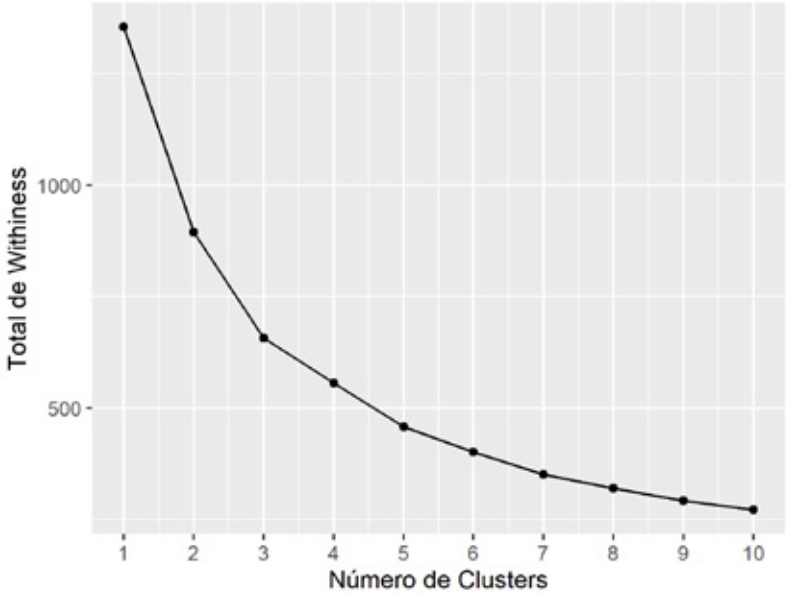

**Fuente:** elaboración propia

**Gráfica A6.** Diagrama de codo para la generación de clusters con k-means de los municipios a partir de incidencia de casos de COVID-19, porcentaje de población ladina, porcentaje de ruralidad y porcentaje de pobreza.



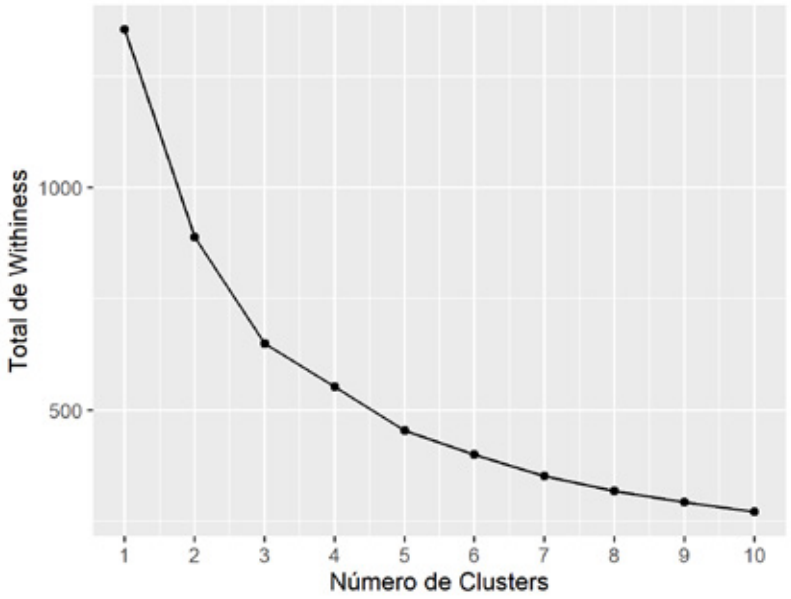

**Fuente:** elaboración propia

**Gráfica A7.** Diagrama de codo para la generación de clusters con k-means de los municipios a partir de tasa de homicidios, porcentaje de población ladina, porcentaje de ruralidad y porcentaje de pobreza.




**Fuente:** elaboración propia

**Gráfica A8.** Diagrama de codo para la generación de clusters con k-means de los municipios a partir de porcentaje de participación en elecciones, porcentaje de población ladina, porcentaje de ruralidad y porcentaje de pobreza.




**Fuente:** elaboración propia

**Tabla A9a.** Medidas por cluster de la variable porcentaje de pobreza, clusters de Incidencia de COVID-19

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 56.1% | 14.7% | 27.2% | 48.0% | 56.2% | 68.2% | 90.1% |
| 2 | 42 | 26.1% | 9.0% | 10.6% | 21.0% | 27.5% | 32.0% | 42.9% |
| 3 | 69 | 66.1% | 8.9% | 41.6% | 60.3% | 66.0% | 71.7% | 87.8% |
| 4 | 76 | 45.4% | 8.0% | 23.1% | 41.4% | 44.9% | 50.2% | 70.4% |
| 5 | 107 | 80.0% | 9.3% | 58.5% | 72.5% | 81.5% | 87.5% | 94.6% |
| Todos | 340 | 59.6% | 20.5% | 10.6% | 44.0% | 62.0% | 75.8% | 94.6% |

**Fuente:** elaboración propia

**Tabla A9b.** Medidas por cluster de la variable de porcentaje de ruralidad, clusters de Incidencia de COVID-19

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 21.5% | 18.4% | 0.0% | 0.3% | 22.0% | 34.2% | 61.0% |
| 2 | 42 | 11.0% | 12.0% | 0.0% | 0.0% | 9.9% | 18.6% | 46.8% |
| 3 | 69 | 77.3% | 14.5% | 34.8% | 69.4% | 79.6% | 88.6% | 97.5% |
| 4 | 76 | 57.6% | 19.3% | 0.0% | 46.9% | 62.4% | 69.5% | 91.9% |
| 5 | 107 | 76.9% | 14.7% | 39.0% | 65.4% | 81.2% | 88.1% | 98.8% |
| Todos | 340 | 57.0% | 29.9% | 0.0% | 37.1% | 64.8% | 82.7% | 98.8% |

**Fuente:** elaboración propia

**Tabla A9c.** Medidas por cluster de la variable porcentaje de población ladina, clusters de Incidencia de COVID-19

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 17.2% | 19.5% | 0.1% | 2.0% | 9.6% | 27.0% | 77.2% |
| 2 | 42 | 82.4% | 20.6% | 17.0% | 76.2% | 91.9% | 94.9% | 98.4% |
| 3 | 69 | 83.8% | 13.9% | 41.5% | 75.4% | 88.6% | 95.9% | 99.5% |
| 4 | 76 | 90.2% | 13.6% | 41.3% | 91.0% | 96.2% | 97.9% | 99.3% |
| 5 | 107 | 10.2% | 11.8% | 0.1% | 1.6% | 5.9% | 15.1% | 48.8% |
| Todos | 340 | 52.9% | 39.8% | 0.1% | 8.2% | 63.7% | 93.5% | 99.5% |

**Fuente:** elaboración propia

**Tabla A9d.** Medidas por cluster de la variable incidencia de casos de COVID-19, clusters de Incidencia de COVID-19

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 1.11 | 0.72 | 0.02 | 0.45 | 1.18 | 1.51 | 2.98 |
| 2 | 42 | 4.00 | 1.42 | 1.79 | 2.97 | 3.57 | 4.68 | 7.43 |
| 3 | 69 | 0.83 | 0.39 | 0.21 | 0.58 | 0.76 | 1.01 | 2.13 |
| 4 | 76 | 1.68 | 0.66 | 0.42 | 1.26 | 1.63 | 2.02 | 3.72 |
| 5 | 107 | 0.55 | 0.47 | 0.00 | 0.22 | 0.41 | 0.74 | 2.83 |
| Todos | 340 | 1.36 | 1.29 | 0.00 | 0.49 | 0.98 | 1.75 | 7.43 |

**Fuente:** elaboración propia

**Tabla A10a.** Medidas por cluster de la variable porcentaje de pobreza, clusters de tasa de homicidios

| ID cluster | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 102 | 58.1% | 13.2% | 23.1% | 49.1% | 58.5% | 69.2% | 87.8% |
| 2 | 42 | 50.9% | 13.6% | 31.1% | 42.8% | 47.1% | 57.4% | 89.5% |
| 3 | 45 | 56.6% | 15.2% | 27.2% | 48.0% | 56.8% | 68.9% | 90.1% |
| 4 | 105 | 79.8% | 9.4% | 58.5% | 72.0% | 80.7% | 87.3% | 94.6% |
| 5 | 46 | 27.6% | 10.3% | 10.6% | 21.5% | 28.0% | 32.6% | 49.7% |
| Todos | 340 | 59.6% | 20.5% | 10.6% | 44.0% | 62.0% | 75.8% | 94.6% |

**Fuente:** elaboración propia

**Tabla A10b.** Medidas por cluster de la variable porcentaje de ruralidad, clusters de tasa de homicidio

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 102 | 71.8% | 15.7% | 34.8% | 61.7% | 72.1% | 85.1% | 97.5% |
| 2 | 42 | 59.7% | 22.0% | 0.0% | 46.8% | 62.5% | 73.0% | 98.8% |
| 3 | 45 | 21.7% | 18.0% | 0.0% | 0.7% | 22.3% | 34.4% | 61.0% |
| 4 | 105 | 76.9% | 14.3% | 39.0% | 65.6% | 80.9% | 87.7% | 98.6% |
| 5 | 46 | 10.9% | 11.1% | 0.0% | 0.0% | 10.8% | 18.9% | 37.4% |
| Todos | 340 | 57.0% | 29.9% | 0.0% | 37.1% | 64.8% | 82.7% | 98.8% |

**Fuente:** elaboración propia

**Tabla A10c.** Medidas por cluster de la variable porcentaje de población ladina, clusters de tasa de homicidios

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 102 | 83.9% | 15.4% | 34.6% | 73.7% | 89.8% | 96.4% | 99.5% |
| 2 | 42 | 92.2% | 13.2% | 35.3% | 95.7% | 97.1% | 98.6% | 99.2% |
| 3 | 45 | 14.5% | 16.3% | 0.1% | 2.0% | 9.2% | 22.2% | 64.0% |
| 4 | 105 | 10.0% | 11.5% | 0.1% | 1.5% | 5.9% | 14.3% | 48.8% |
| 5 | 46 | 83.7% | 18.2% | 17.0% | 79.0% | 92.1% | 94.4% | 98.4% |
| Todos | 340 | 52.9% | 39.8% | 0.1% | 8.2% | 63.7% | 93.5% | 99.5% |

**Fuente:** elaboración propia

**Tabla A10d.** Medidas por cluster de la variable tasa de homicidios por cada 100,000 habitantes entre 2019 y 2022, clusters de tasa de homicidios

| ID *cluster* | cantidad municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 102 | 44.4 | 31.6 | 0.0 | 16.0 | 44.8 | 70.3 | 120.7 |
| 2 | 42 | 162.6 | 55.5 | 92.8 | 122.9 | 152.5 | 185.3 | 331.8 |
| 3 | 45 | 20.2 | 23.5 | 0.0 | 4.3 | 12.4 | 27.8 | 116.3 |
| 4 | 105 | 12.1 | 18.2 | 0.0 | 1.5 | 7.0 | 13.7 | 118.6 |
| 5 | 46 | 79.7 | 51.0 | 6.8 | 34.8 | 79.6 | 116.7 | 197.6 |
| Todos | 340 | 50.6 | 58.7 | 0.0 | 7.5 | 25.9 | 79.3 | 331.8 |

**Fuente:** elaboración propia

**Tabla A11a.** Medidas por cluster de la variable porcentaje de pobreza, clusters de porcentaje de empadronamiento

| ID *cluster* | cant. municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 58.2% | 16.1% | 27.2% | 48.0% | 58.3% | 70.3% | 92.0% |
| 2 | 132 | 56.0% | 12.8% | 31.1% | 45.2% | 54.4% | 65.8% | 87.8% |
| 3 | 53 | 30.1% | 12.1% | 10.6% | 22.0% | 30.1% | 38.6% | 63.1% |
| 4 | 4 | 46.9% | 19.9% | 23.1% | 39.1% | 46.4% | 54.2% | 71.7% |
| 5 | 105 | 79.9% | 9.3% | 58.5% | 72.6% | 81.5% | 87.3% | 94.6% |
| Todos | 340 | 59.6% | 20.5% | 10.6% | 44.0% | 62.0% | 75.8% | 94.6% |

**Fuente:** elaboración propia

**Tabla A11b.** Medidas por cluster de la variable porcentaje de ruralidad, clusters de porcentaje de empadronamiento

| ID *cluster* | cant. municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 24.9% | 18.6% | 0.0% | 8.8% | 26.5% | 39.5% | 61.0% |
| 2 | 132 | 69.6% | 15.9% | 34.8% | 58.5% | 69.8% | 82.2% | 97.5% |
| 3 | 53 | 11.4% | 12.5% | 0.0% | 0.0% | 10.7% | 18.9% | 46.8% |
| 4 | 4 | 51.4% | 24.0% | 17.4% | 42.8% | 59.5% | 68.1% | 69.0% |
| 5 | 105 | 78.5% | 13.3% | 48.0% | 69.5% | 82.9% | 88.8% | 98.8% |
| Todos | 340 | 57.0% | 29.9% | 0.0% | 37.1% | 64.8% | 82.7% | 98.8% |

**Fuente:** elaboración propia

**Tabla A11c.** Medidas por cluster de la variable porcentaje de población ladina, clusters de porcentaje de empadronamiento

| ID *cluster* | cant. municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 12.3% | 13.5% | 0.1% | 1.6% | 7.5% | 17.8% | 49.4% |
| 2 | 132 | 87.3% | 13.7% | 41.3% | 80.0% | 92.4% | 97.8% | 99.5% |
| 3 | 53 | 83.7% | 18.1% | 17.0% | 77.2% | 92.2% | 95.7% | 98.8% |
| 4 | 4 | 62.3% | 39.5% | 21.5% | 31.9% | 64.9% | 95.2% | 97.7% |
| 5 | 105 | 11.5% | 14.2% | 0.1% | 1.8% | 5.9% | 16.4% | 67.6% |
| Todos | 340 | 52.9% | 39.8% | 0.1% | 8.2% | 63.7% | 93.5% | 99.5% |

**Fuente:** elaboración propia

**Tabla A11d.** Medidas por cluster de la variable porcentaje de población empadronada, clusters de porcentaje de empadronamiento

| ID *cluster* | cant. municipios | promedio | desviación estándar | mínimo (Q0) | Q1 | mediana (Q2) | Q3 | máximo (Q4) |
|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 71.5% | 10.0% | 52.4% | 63.3% | 71.3% | 80.3% | 87.5% |
| 2 | 132 | 61.0% | 8.2% | 40.7% | 55.1% | 61.3% | 67.9% | 85.2% |
| 3 | 53 | 63.2% | 7.3% | 49.3% | 57.3% | 63.1% | 68.4% | 76.1% |
| 4 | 4 | 5.2% | 7.0% | 0.0% | 0.0% | 3.0% | 8.2% | 14.8% |
| 5 | 105 | 65.4% | 9.5% | 37.0% | 58.6% | 65.3% | 72.3% | 83.3% |
| Todos | 340 | 63.5% | 11.3% | 0.0% | 57.6% | 63.7% | 69.8% | 87.5% |

**Fuente:** elaboración propia

# COVID-19

*Análisis territorial de grandes problemas de Desarrollo en Guatemala con una mirada de vulnerabilidad socioeconómica utilizando aprendizaje de máquina no supervisado*

Laboratorio de Datos GT

OXFAM

VACUNAS PARA LA GENTE LATINOAMÉRICA

AGENDA REGIONAL PARA UNA VACUNA UNIVERSAL CONTRA LA COVID-19