18 AÑOS

CONEVAL
Consejo Nacional de Evaluación de la Política de Desarrollo Social

Metodología de la tipología municipal

Abril de 2023

Metodología de la tipología municipal

La tipología municipal es un ejercicio de agrupación de las demarcaciones y municipios del país con base en algunas características demográficas, económicas, geográficas y de infraestructura. Esta agrupación se realizó con el objetivo de analizar la situación de pobreza en el ámbito rural, así como identificar el contexto territorial de los municipios considerados rurales[1]. Este indicador forma parte de la Plataforma para el Análisis Territorial de la Pobreza.

De acuerdo con los criterios que definen a lo rural del capítulo uno del documento *Pobreza rural en México*, la clasificación de los municipios tenía que integrar variables de diversas dimensiones como la demográfica, la económica, la geográfica y de infraestructura:

- **Demográfica**: variables asociadas con las características del tamaño y distribución de la población
- **Económica**: variables asociadas con las características económicas de la población
- **Geográfica**: variables asociadas con los espacios y la conectividad de los municipios
- **Infraestructura**: características asociadas con las viviendas

De entre 30 indicadores y variables, se seleccionaron aquellos más representativos de acuerdo con las consideraciones operativas y mediante el análisis estadístico de *valores medios entre grupos y variables*, de los cuales resultaron 12 (ver cuadro 1):

[1] El análisis sobre la pobreza en los municipios rurales del país se puede consultar en el documento Pobreza rural en México. Para más información consultar: https://www.coneval.org.mx/Medicion/MP/Documents/PATP/Pobreza_rural.pdf

Cuadro 1. Listado de variables seleccionadas

Dimensión	Indicador/variable
Demográfica	Población total
	Población en localidades con menos de 2500 habitantes (%)
	Promedio de años de escolaridad (15 o más años)
	Densidad de población
Económica	Tasa de participación económica
	Personal ocupado entre el número de unidades económicas
	Cociente de localización: primario, secundario y terciario
Geográfica	Índice de conectividad
Infraestructura	Porcentaje de viviendas que no cuentan con servicio de drenaje o el desagüe tiene conexión con una tubería que va a dar a un río, lago, mar, barranca o grieta
	Porcentaje de viviendas sin internet
	Porcentaje de viviendas con pisos de tierra
	Porcentaje de viviendas donde usan leña o carbón para cocinar y no cuentan con chimenea

Fuente: elaboración del CONEVAL.

Posterior a la selección de las variables, se utilizó el método *k-means* para agruparlas. Esta técnica consiste en dividir el conjunto de observaciones (a partir de un conjunto de variables) en un número fijo de grupos, especificado previamente y, con base en la revisión operativa del capítulo uno del documento Pobreza rural en México, se determinaron 3 categorías. La idea principal se basa en minimizar la variación de las observaciones dentro de cada grupo. Se busca obtener grupos en donde la suma de las varianzas internas sea mínima mediante el uso de medidas de distancia entre observaciones.

De manera general el algoritmo consiste en lo siguiente:

1. Se seleccionan *k* observaciones del conjunto inicial (donde k es el número de grupos preestablecido). Estas observaciones son conocidas como centroides y alrededor de éstas, se formarán los grupos.

2. Cada observación es asignada al grupo cuyo centroide es el más cercano a partir de la distancia (euclidiana) entre dicha observación y cada centroide.

3. Se recalcula un nuevo centroide obteniendo el vector de medias de todos los integrantes asignados a cada grupo.

4. Con el nuevo centroide se repiten el paso 2 y 3 hasta que el algoritmo converge (los centroides no cambian y los grupos se mantienen iguales).

Adicionalmente se exploraron situaciones que se podrían considerar adversas al elegir este método, no obstante, también se buscaron las soluciones para justificar el uso de la técnica ante cada una de las contingencias (ver cuadro 2).

Cuadro 2. Justificación del uso de la técnica de K-means

Número	Situación	Estrategia utilizada
1	Para utilizar este tipo de análisis es necesario que se indique el número de grupos que se desean formar.	La propuesta no busca agrupar en un número indefinido de grupos, más bien, dada una estructura definida en tres categorías, se busca agrupar a los municipios con base en las dimensiones elegidas. A pesar de lo anterior, se llevaron a cabo diversos métodos para identificar el número óptimo de grupos, el cual mayoritariamente coincidió en tres.
2	Si se emplea la distancia euclidiana es necesario que las variables empleadas sean de tipo continuo, ya	En el análisis se propone emplear únicamente variables de tipo continuo.

Número	Situación	Estrategia utilizada
	que se debe poder calcular la media de cada una de ellas (existe una versión de k-means para variables discretas).	
3	Los resultados del método pueden variar dependiendo de la asignación inicial de los centroides. Puede ocurrir que con diferentes centroides los óptimos locales obtenidos sean distintos y por lo tanto las clasificaciones obtenidas también lo sean. El riesgo de esta desventaja puede ser disminuido al repetir el proceso un número alto de veces con distintos centroides, y seleccionar el resultado que tenga menor suma total de varianza interna.	Se realizaron pruebas en las que se puede observar que los resultados obtenidos tienen una alta probabilidad de ser las agrupaciones derivadas de los óptimos globales (óptimos en cuanto a la similitud dentro de grupos).
4	Si se realizan inicios aleatorios de los centroides solo se puede garantizar la replicabilidad de los resultados si se emplean semillas (puede existir replicabilidad aun cuando no se usen las mismas semillas dependiendo del comportamiento de los datos).	El método fue inicializado fijando los centroides por lo tanto es replicable siempre y cuando se usen las mismas observaciones como centroides.
5	El método presenta problemas de robustez en presencia de outliers. Lo anterior afecta los resultados ya que el promedio del grupo al que sea asignado el outlier se moverá más bruscamente que los demás y por lo tanto puede crear un centroide en un óptimo local. Como alternativa a este problema existen dos opciones, la primera es excluir los outliers del análisis; en caso de que no sea posible la alternativa más viable es cambiar el análisis a uno basado en medoides.	Dado que no es posible excluir ninguna observación de la base de datos usada para la tipología, se procedió con el análisis con la información completa. No hay evidencia de falta de robustez en el método (al menos ninguna que no se haya subsanado en los puntos tres y cuatro de esta tabla).
6	Este tipo de análisis no buscan una relación causal sino puramente descriptiva, basada en similitudes. Por lo anterior, los agrupamientos obtenidos no garantizan tener algún significado o utilidad más allá de la descripción a través de las variables	Los grupos obtenidos guardan concordancia con las características esperadas incluso en variables que no fueron incluidas en el análisis. Por lo anterior se concluye que los resultados son útiles para el objetivo de la

Número	Situación	Estrategia utilizada
	usadas en el análisis.	propuesta.
7	Dado el carácter descriptivo del método, los resultados obtenidos dependen completamente de las variables que se introduzcan al algoritmo y estos pueden variar incluso con solo agregar una variable más.	Las variables incluidas fueron compiladas de acuerdo con su relevancia con el tema. Además, se realizaron análisis de correlaciones para evitar incluir variables que pudieran afectar a los resultados.

Fuente: elaboración del CONEVAL.

Consideraciones

1. Debido al método seleccionado los resultados entre un año y otro pueden variar a pesar de incluir los mismos indicadores, por tal motivo, la tipología solo permite explicar el contexto de los municipios en un momento determinado y no es comparable.
2. Los tres valores que se manejan como semillas son los centroides con los que empieza a iterar el algoritmo, estos centroides son vectores de dimensión 12 cada uno. Se utilizaron los municipios 09007 (valor máximo), 22003 (valor mediano) y 20047 (valor mínimo) considerando la variable población.
3. Los algoritmos de optimización son Lloyd y Forgy (que básicamente son dos diferentes etiquetas para el mismo algoritmo según la descripción de la función en el programa R).
4. Para replicar los resultados es necesario estandarizar (se usó la función *scale* en el programa R), si no se realiza dicha estandarización (ya sea con la función *scale* u otra) no se obtiene la misma clasificación.
5. No importa la semilla inicial si se fijan los centroides, lo anterior por la construcción del método. Sin embargo, se realizaron 10, 000

réplicas con distintas semillas (1 a la 10,000 usando el comando *set.seed*) para corroborar lo anterior. El resultado se replica.

6. Se usaron 100 iteraciones máximas. En el programa R se pueden definir más iteraciones, pero es evidente que los datos manejados no requieren más de 100 iteraciones para converger a óptimos locales.

Bibliografía

Bradley, P. S., y Fayyad, U. M. (1998). *Refining initial points for k-means clustering*. ICML,98. 91-99.

González, S. y Larralde, A. (2013). Conceptualización y medición de lo rural. Una propuesta para clasificar el espacio rural en México. La situación demográfica de México 2013. Consejo Nacional de Población. México.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Johnson, R. A., y Wichern, D. W. (2002). *Applied multivariate statistical analysis* Upper Saddle River. Prentice hall.

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw Hill.