Análisis de conglomerados: una aplicación a la educación

Lucero Martínez Bonilla, Fernando Velasco Luna, Francisco Solano Tajonar Sanabria
Facultad de Ciencias Físico-Matemáticas, BUAP
lucero.martinezb@alumno.buap.mx
Septiembre 2025

Resumen

El presente estudio analiza la esperanza de escolaridad en México a través de la aplicación de técnicas de análisis de conglomerados utilizando datos del INEGI correspondientes al periodo 2015–2023. El objetivo principal fue identificar patrones regionales y agrupar a las entidades federativas de acuerdo con su desempeño educativo. La metodología incluyó el escalamiento de datos, análisis de componentes principales (PCA), método K-medias, método del codo e índice de silueta, lo que permitió determinar tres conglomerados diferenciados: un grupo de estados con desempeño medio-alto, un conglomerado con rezagos educativos significativos y un caso atípico representado por la Ciudad de México.

Palabras clave: esperanza de escolaridad, análisis de conglomerados, *k*-medias, educación en México.

Introducción

La educación es un pilar fundamental para el desarrollo social, económico y humano. Un indicador clave en este ámbito es la esperanza de escolaridad, que mide los años promedio que un niño permanecerá en el sistema educativo. En México, este valor se ha mantenido entre 13 y 14 años en la última década, aunque existen marcadas desigualdades entre estados. Mientras entidades como Ciudad de México, Nuevo León y Querétaro superan el promedio nacional, regiones como Oaxaca, Chiapas y Guerrero presentan rezagos significativos que reflejan brechas estructurales y sociales.
Desde la segunda mitad del siglo XX, las técnicas de análisis multivariado han ofrecido herramientas para identificar patrones ocultos en datos sociales y económicos. Dentro de estas técnicas, el análisis de conglomerados (cluster analysis) se consolidó como un método de clasificación no supervisada que permite agrupar elementos en función de su similitud. Autores como Everitt (2011) y Johnson & Wichern (2007) han señalado que este enfoque no solo se aplica en estadística matemática, sino también en áreas tan diversas como la biología, la psicología, la mercadotecnia y, de manera creciente, en la educación.
En el ámbito educativo, el análisis de conglomerados se ha empleado para detectar desigualdades regionales, comparar sistemas escolares y segmentar poblaciones estudiantiles de acuerdo con características socioeconómicas o niveles de rendimiento. Estas aplicaciones permiten a las autoridades diseñar políticas focalizadas y asignar recursos con mayor equidad. En el caso mexicano, este tipo de estudios cobra especial relevancia dado el contraste entre entidades con alta cobertura y permanencia escolar, frente a otras donde persisten retos significativos relacionados con pobreza, infraestructura deficiente o limitada presencia de programas de apoyo.
En este trabajo se utiliza el análisis de conglomerados para agrupar las entidades federativas de México según su desempeño en esperanza de escolaridad durante el periodo 2015–2023. Se espera que los resultados permitan a especialistas y responsables de política educativa identificar patrones regionales y proponer estrategias diferenciadas que contribuyan a reducir las desigualdades y avanzar hacia una educación más inclusiva y equitativa.

Metodología

Datos: INEGI, esperanza de escolaridad por entidad federativa (ciclos 2015–2016 a 2022–2023).

Cuadro 1: Variables de estudio utilizadas

Tipo de variable	Variable	Valor
Variable de identificación	Entidad federativa	Aguascalientes, Baja California, Baja California Sur, ..., Veracruz de Ignacio de la Llave, Yucatán, Zacatecas.
Variable de identificación	Ciclo escolar	2015–2016, 2016–2017, 2017–2018, ..., 2020–2021, 2021–2022, 2022–2023
Variable activa (numérica)	Esperanza de escolaridad	11.9, 12.0, 12.2, ..., 18.8, 19.0, 19.1

Metodología

Método del codo: Para cada número de conglomerados k, se calcula la Suma de Errores Cuadráticos Internos (Within-Cluster Sum of Squares, WCSS):

$$WCSS(k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

donde C_i representa el i-ésimo conglomerado y μ_i es el vector centroide asociado a dicho conglomerado.
A medida que k aumenta, el valor de $WCSS(k)$ disminuye, ya que los grupos se ajustan mejor a los datos. Sin embargo, esta reducción presenta un punto de rendimiento decreciente: a partir de cierto valor de k, las mejoras adicionales son marginales.
El número óptimo de conglomerados se elige en el punto donde la gráfica de $WCSS(k)$ contra k muestra un cambio notable de pendiente, semejante a un "codo". Este valor indica un balance adecuado entre complejidad del modelo y capacidad explicativa de los datos.

```
# Método del codo
jpeg("/Users/luceromartinezbonilla/Desktop/TESIS/Metodo_del_codo.jpeg")
fviz_nbclust(data_scaled, kmeans, method = "wss") +
  labs(title = "Método del codo")
dev.off()
```

Figura 1: Método del codo para determinar conglomerados en R

k-medias: Intenta minimizar la suma de las distancias cuadráticas dentro de cada conglomerado, utilizando la fórmula:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$$

donde:

- K es el número de conglomerados,
- x_i es el vector de datos de la *í-ésima* observación,
- μ_k es el centroide del conglomerado C_k,
- $\|\cdot\|$ denota la norma euclidiana.

Primero, se realiza la inicialización de los centroides (μ_k). Luego, se asigna cada observación al conglomerado más cercano:

$$C_k = \{x_i : \|x_i - \mu_k\| \leq \|x_i - \mu_j\|, \forall j \neq k\}$$

Luego, se actualizan los centroides:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

Se repite hasta la convergencia.
Como resultado, se obtiene una matriz de asignaciones de conglomerado y los centroides optimizados para los datos.

```
# Conglomerados K-medias
set.seed(123)
k <- 3  # Cambiar a otro valor si es necesario
km_res <- kmeans(data_scaled, centers = k, nstart = 25)

jpeg("/Users/luceromartinezbonilla/Desktop/TESIS/Kmeans_clusters_named.jpeg", width = 1000, height = 1000, res = 150)
fviz_cluster(km_res, data = data_scaled,
             ellipse.type = "convex",
             geom = "point",
             show.clust.cent = TRUE,
             palette = "jco") +  # Mejora los colores
  geom_text(aes(label = data$Entidad),
            vjust = 1.5, hjust = 1.5, size = 3, check_overlap = TRUE) +  # Evita superposición de etiquetas
  theme_minimal(base_size = 12) +  # Usa un tema limpio y ajusta el tamaño base
  labs(title = "Conglomerados K-medias con nombres de regiones",
       x = "Dim1", y = "Dim2") +
  theme(plot.title = element_text(hjust = 0.5))  # Centra el título
dev.off()
```

Figura 2: Agrupación por K-medias en R

Agrupamiento jerárquico (dendrograma): El procedimiento comienza calculando una matriz de distancias $D = (d_{ij})$ entre todas las observaciones, donde una medida común es la distancia euclidiana:

$$d_{ij} = \|x_i - x_j\|^2 = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}.$$

A partir de D, se sigue un proceso iterativo:

1. Cada observación se considera inicialmente como un conglomerado individual.
2. En cada paso, se fusionan los dos conglomerados más cercanos según un criterio de enlace:

$$d(A, B) = \begin{cases} \min\{d_{ij} : x_i \in A, x_j \in B\}, & \text{enlace simple,} \\ \max\{d_{ij} : x_i \in A, x_j \in B\}, & \text{enlace completo,} \\ \dfrac{1}{|A||B|} \sum_{x_i \in A} \sum_{x_j \in B} d_{ij}, & \text{enlace promedio.} \end{cases}$$

3. El proceso continúa hasta que todos los elementos se agrupan en un único conglomerado.

```
# Agregar resultados de cluster al dataset original
data$cluster <- as.factor(km_res$cluster)

# Matriz de distancias y clustering jerárquico
dist_matrix <- dist(data_scaled, method = "euclidean")
hc <- hclust(dist_matrix, method = "complete")

jpeg("/Users/luceromartinezbonilla/Desktop/TESIS/Dendrograma_named.jpeg")
plot(hc, labels = data$Entidad, main = "Dendrograma jerárquico con nombres", sub = "", xlab = "", cex = 0.8)
rect.hclust(hc, k = k, border = 2:4)
dev.off()
```

Figura 3: Agrupación jerárquica y dendrograma en R

Resultados principales

Método del codo:
Al aumentar k, el valor de *WCSS* disminuye porque los grupos se ajustan mejor a los datos. El "codo" de la gráfica se observa alrededor de $k = 3$, lo que indica que el número óptimo de conglomerados es 3, ya que en ese punto se alcanza un equilibrio entre simplicidad del modelo y capacidad de explicar la variabilidad de los datos.

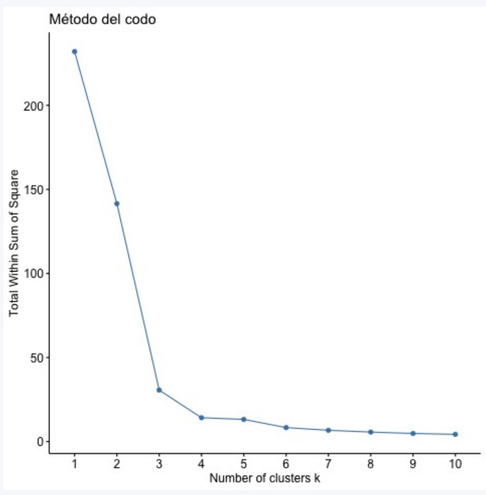

Figura 4: Método del codo

k-medias:

- **Conglomerado 1:** Baja California Sur, Coahuila de Zaragoza, Nuevo León, entre otras.
- **Conglomerado 2:** Ciudad de México, que destaca como un caso atípico con características únicas que la separan del resto.
- **Conglomerado 3:** Campeche, Oaxaca, Chiapas, entre otras.

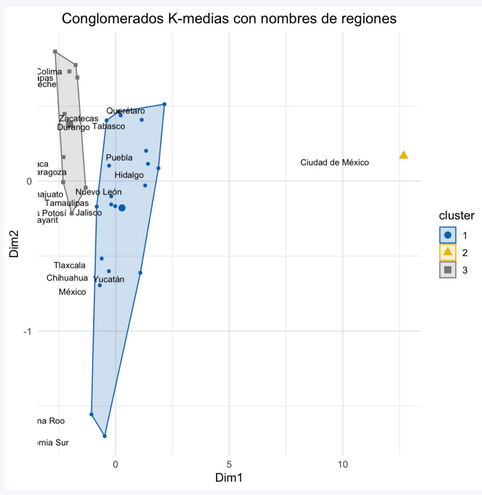

Figura 5: Conglomerados por K-medias

Dendrograma jerárquico:
El eje vertical ("Height") indica la distancia o disimilitud entre los grupos: cuanto más alto se unen dos ramas, más diferentes son entre sí. Confirma la clasificación previa hecha con k-medias, pero además ofrece una visión jerárquica de cómo se van fusionando los estados en niveles crecientes de similitud.

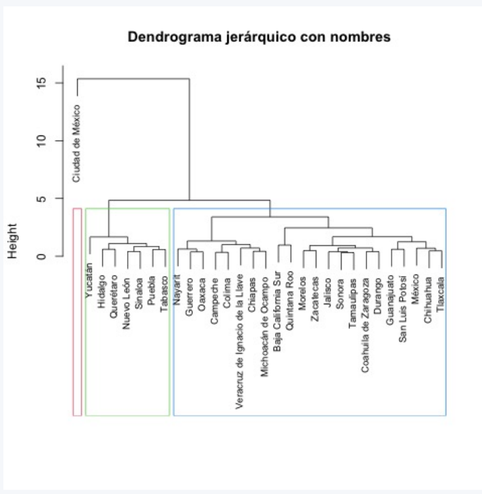

Figura 6: Dendrograma jerárquico con nombres

Conclusión

Los gráficos generados confirman la robustez de los conglomerados formados, proporcionando confianza en la utilidad del análisis para la planificación educativa. Estos resultados pueden ser utilizados por investigadores y tomadores de decisiones para diseñar políticas focalizadas y evaluar su impacto a nivel estatal o nacional.

Referencias

- Everitt, B. et al. (2011). *Cluster Analysis*. Wiley.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.
- INEGI (2024). *Esperanza de escolaridad por entidad federativa*. https://www.inegi.org.mx
- SEP (2023). *Panorama educativo mexicano*.
- González, A. et al. (2021). *Análisis de conglomerados aplicados al desempeño educativo en México*.