UNIVERSIDAD DE BUENOS AIRES – FCE

TALLER DE PROGRAMACIÓN

TP N°2 HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

PROFESORA: María Noelia Romero

Alumnos: Bergant, Agustín Rivas
Cabrera, Juan
Leiva, Francisco

Link a GitHub

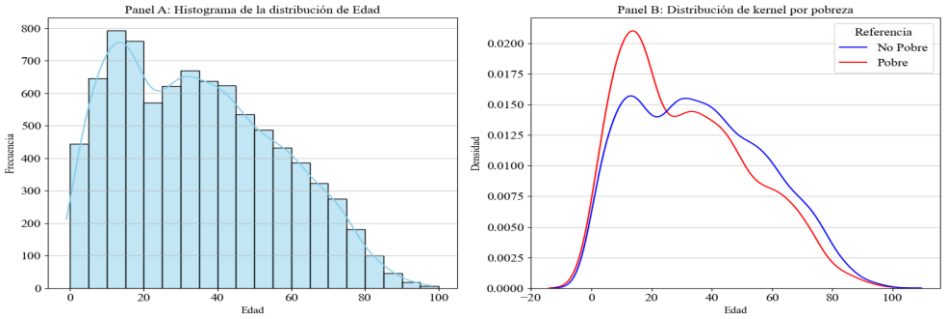
Introducción

El presente trabajo práctico tiene como objetivo profundizar en el análisis de la base de datos de la Encuesta Permanente de Hogares[1] (EPH) para la región Patagónica, aplicando herramientas de estadística descriptiva y métodos no supervisados. A partir de la base construida en el TP 1, se indagan las principales características sociodemográficas y laborales de la población –como edad, educación, ingresos y horas trabajadas– mediante histogramas, distribuciones kernel y técnicas de agrupamiento de datos (PCA y cluster). El análisis busca identificar diferencias en la distribución de estas variables entre personas pobres y no pobres, así como posibles patrones dentro de la estructura socioeconómica regional.

Parte I

1) En el Panel A de la Figura 1, se observa que la distribución de edades presenta una mayor concentración en los grupos jóvenes, con una frecuencia que disminuye progresivamente a medida que aumenta la edad. En el Panel B, las curvas de densidad kernel muestran que las personas pobres tienden a concentrarse en edades más tempranas, mientras que las no pobres presentan una distribución más extendida hacia edades adultas. Esto sugiere que la pobreza se asocia con una estructura etaria más joven. Algunas de las causas que podrían explicar esta situación son la baja inserción laboral de los jóvenes en los segmentos formales del mercado laboral, o bien el mayor alcance y cobertura del sistema jubilatorio.

Figura 1 | Histograma de la distribución de la edad y distribución de kernel por pobreza.



2) En la Tabla 1, se pueden observar las estadísticas descriptivas de la variable *educ*, que mide la cantidad de años de educación formal. En primer lugar, la variable presenta un promedio de 10,1 años y una mediana de 11, lo que indica que la mayoría de las personas alcanzan, en promedio, un nivel cercano a la educación secundaria completa. La desviación estándar de aproximadamente 5 años refleja una importante heterogeneidad educativa dentro de la población. El valor mínimo de 0 evidencia la presencia de individuos sin educación formal, mientras que el máximo de 23 años corresponde a personas con estudios universitarios o de posgrado. En conjunto, los resultados muestran una estructura educativa relativamente amplia, con una mayoría que accede al nivel medio y una menor proporción con educación superior.

Tabla 1 | Estadísticas descriptivas de los años de educación.

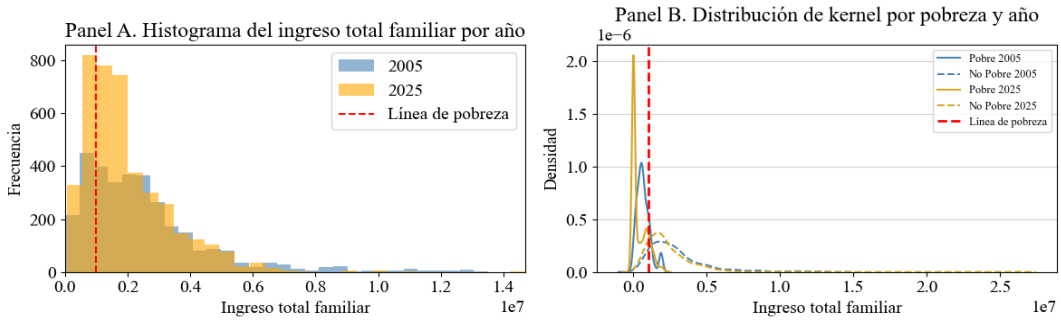
Media	10,1 horas
Desvío estándar	5,1 horas
Mínimo	0,0 horas
Mediana	11,0 horas
Máximo	23,0 horas

3) En el Panel A de la Figura 2 se aprecia que la distribución del ingreso total familiar en ambos años (2005 y 2025) presenta una fuerte asimetría hacia la derecha, con la mayoría de los hogares concentrados en los tramos de ingresos bajos y una caída progresiva a medida que aumentan los ingresos. Al comparar ambos años, la distribución de 2025 (en amarillo) muestra un leve desplazamiento hacia la derecha respecto de

[1] De lo anterior se desprende que todas las figuras y tablas desarrolladas en el presente trabajo proviene de dicha fuente, eliminando en consecuencia la referencia en cada una de ellas.

2005, lo que da cuenta de una peor situación en términos generales. En el Panel B, la densidad de los hogares pobres se concentra claramente por debajo de la línea de pobreza, –siendo más aguda la situación en el caso de hogares pobres de 2025–, mientras que los no pobres muestran una distribución más extendida y heterogénea. En conjunto, las curvas evidencian una marcada desigualdad en la distribución del ingreso y una brecha persistente entre los hogares pobres y no pobres en la región patagónica.

Figura 2 | Distribución del Ingreso total familiar y distribución según pobreza por año.



4) La variable *horastrab*, que mide el total de horas trabajadas semanalmente por el jefe o jefa del hogar, presenta un promedio de 43,5 horas semanales y una mediana de 40,0, lo que indica que la mayoría de los trabajadores cumple una jornada cercana a la de tiempo completo (Tabla 2). La desviación estándar de 16,2 horas refleja una heterogeneidad significativa en las cargas horarias, probablemente asociada a la presencia de empleos informales, subocupación o múltiples ocupaciones. El mínimo es de 2,0 horas y el máximo de 126,0, lo cual en principio es un valor fuera de lo común (suponiendo que trabaja los 7 días de la semana, da un promedio de 18 horas diarias). En conjunto, los resultados sugieren una estructura laboral diversa.

Tabla 2 | Estadísticas descriptivas de las horas trabajadas.

Media	43,5
Desvío estándar	16,2
Mínimo	2,0
Mediana	40,0
Máximo	126,0

5) La base final para la región Patagónica cuenta con un total de 8.588 observaciones tal como se observa en la Tabla 3, integrando los primeros trimestres de 2005 (3229 observaciones) y 2025 (5359). En ambos años no se registran valores faltantes en la variable Pobre, lo que indica una correcta unificación y limpieza de la base. Se observa un aumento considerable en la cantidad de hogares pobres, que pasan de 623 en 2005 a 2.136 en 2025, reflejando un deterioro relativo en las condiciones económicas de la población. En total, la base cuenta con 43 variables homogéneas entre ambos períodos, lo que garantiza la comparabilidad de los análisis posteriores.

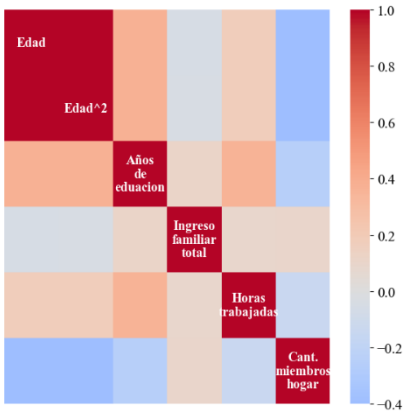
Tabla | Resumen de la base final para la región patagónica.

	2005	2025	Total
Cantidad de observaciones	3.229	5.359	8.588
Cantidad de observaciones con NAs en la variable “Pobre”	0	0	0
Cantidad de pobres	623	2.136	2.759
Cantidad de no pobres	2.606	3.223	5.829
Cantidad de variables limpias y homogeneizadas	43	43	43

Parte II

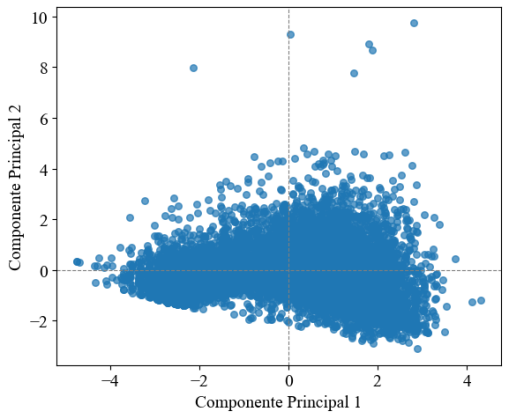
1) Tal como se ve en la Figura 3, la matriz de correlaciones muestra una relación muy fuerte y positiva entre edad y edad^2, como era esperable por construcción. También se observa una correlación positiva –aunque moderada– entre los años de educación, el ingreso familiar total y las horas trabajadas, lo que sugiere que mayores niveles educativos y mayor tiempo de trabajo se asocian con mayores ingresos. En cambio, la cantidad de miembros del hogar presenta correlaciones negativas y leves con las demás variables, lo que indica que los hogares más numerosos no necesariamente poseen mayores niveles de ingreso o educación.

Figura 3 | Matriz de correlaciones de los predictores de la región patagónica



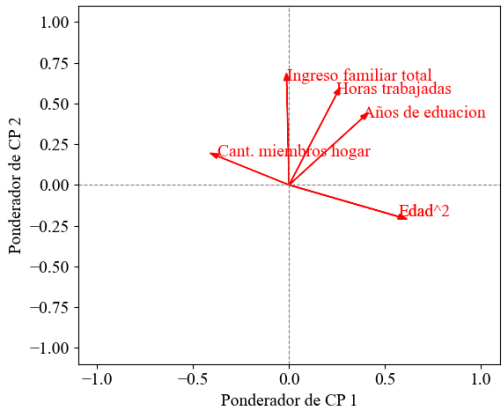
2) El gráfico de dispersión de los dos primeros componentes principales muestra una nube de puntos concentrada alrededor del origen, con una mayor dispersión sobre el eje del primer componente. Esto sugiere que el primer componente captura la mayor parte de la variabilidad de las variables originales, mientras que el segundo agrega información complementaria, aunque de menor peso en términos comparativos. Adicionalmente, en la Figura 4, no se observan agrupamientos claramente definidos, lo que indica que las dimensiones reducidas del PCA no separan de manera evidente a los individuos según sus características socioeconómicas.

Figura 4 | Gráfico de dispersión de los componentes principales.



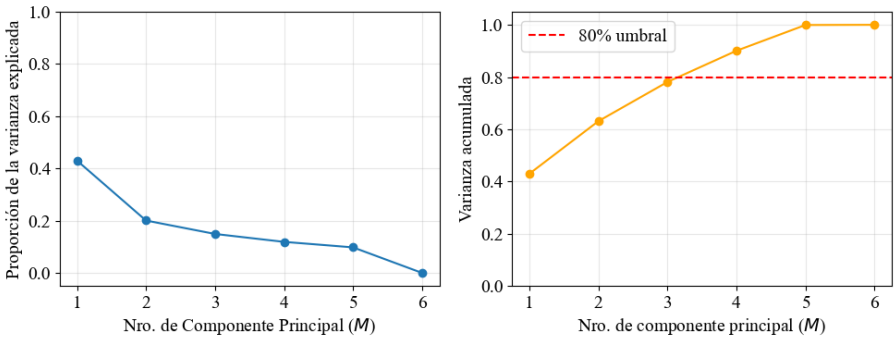
3) En la Figura 5, se ve el gráfico de ponderadores. Lo que se observa es que el primer componente principal (que explica el 43% de la varianza) se encuentra fuertemente influenciado por las variables *Edad*^{2} y, en menor medida, por los años de *Educación* y la *Cantidad de miembros del hogar*. En cambio, el segundo componente (que explica el 20% de la varianza), está más vinculado con las variables *Ingreso total familiar* y *Horas trabajadas*. La dirección y longitud de las flechas indican la importancia relativa de cada variable en la construcción de los componentes (cuanto más larga, mayor es el peso en la explicación de la varianza). En conjunto, sintetizan la información socioeconómica del conjunto de datos en 2 dimensiones principales.

Figura 5 | Gráfico de ponderadores de los componentes principales.



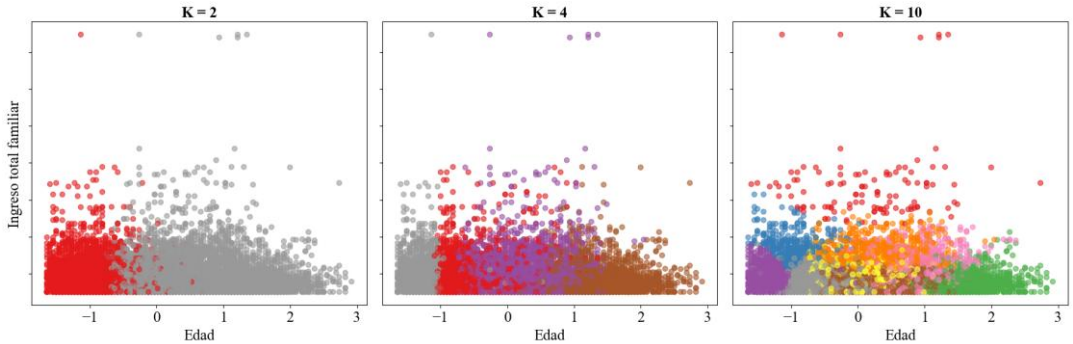
4) La Figura 6 A muestra la proporción de la varianza que explican cada uno de los componentes principales. El primer componente explica el 42% de la varianza total de los datos, mientras que el segundo componente agrega un 20%, alcanzando entre ambos cerca de dos tercios de la información contenida en las seis variables originales. A partir del tercer componente, el aporte marginal a la varianza explicada disminuye (un 14% de aporte), evidenciando rendimientos menores. Esto sugiere que los dos primeros componentes capturan la estructura esencial de los datos, reduciendo la dimensionalidad sin perder información relevante. La Figura 6 B, exhibe el acumulado de la varianza explicado por cada uno de los componentes.

Figura 6 | Proporción explicada de la varianza por cada uno de los componentes.



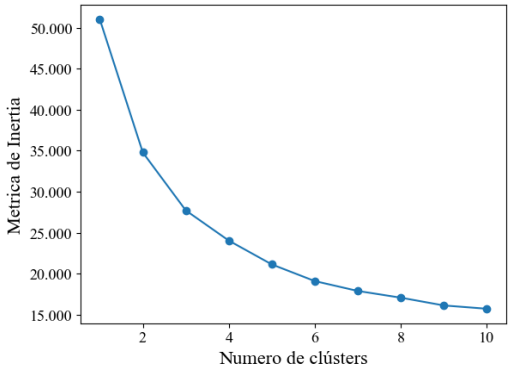
5) A) Los resultados del algoritmo *k-medias* se ven en la Figura 7. Con *k=2*, los individuos se dividen en dos grupos determinados principalmente por el nivel de ingreso familiar, aunque con una superposición considerable de los clusters. Esto sugiere que el modelo logra captar parcialmente la diferencia entre pobres y no pobres, pero sin ser nítida debido a la continuidad de la variable ingreso y la influencia de la edad. Al aumentar el número de grupos, *k=4* y *k=10*, los clusters se vuelven más fragmentados, aunque sin mejoras sustanciales en la identificación del binomio pobres-no pobres.

Figura 7 | Cluster de k-medias para las variables Ingreso familiar total y Edad.



B) El gráfico de la métrica de inercia muestra una caída pronunciada entre *k=1* y *k=3*, a partir de la cual la pendiente comienza a aplanarse progresivamente. Esto sugiere que el número óptimo de clusters se encuentra alrededor de *k=2* o *k=3*, ya que agregar más grupos no mejora sustancialmente la capacidad explicativa del modelo, como puede verse en la Figura 8. La cantidad óptima de grupos (entre dos y tres) no permite distinguir de manera estricta entre pobres y no pobres, ya que ambos grupos presentan cierta superposición dentro de los clusters. Esta segmentación parece estar asociada mejor a las diferencias en la estructura socioeconómica, identificando hogares de ingresos bajos, medios y altos. Por lo tanto, el algoritmo refleja una clasificación más asociada a distintos estratos de bienestar económico que a la dualidad pobre-no pobre.

Figura 8 | Métrica de inercia e identificación de Elbow.



6) El dendrograma del clustering jerárquico muestra cómo las observaciones se agrupan progresivamente según su nivel de disimilitud, representado en el eje vertical. Las ramas más bajas reflejan un alto grado de similitud entre individuos o hogares, mientras que las fusiones en niveles superiores agrupan conjuntos cada vez más heterogéneos. Un dendrograma es una representación gráfica en forma de árbol que permite observar la estructura jerárquica de los datos y decidir visualmente el número de clusters cortando el árbol a una determinada altura. En este caso, la Figura 9 sugiere la presencia de tres o cuatro grandes conglomerados socioeconómicos, más asociados a diferentes niveles de ingreso y composición del hogar que a una separación estricta entre pobres y no pobres.

Figura 9 | Dendrograma del clustering jerárquico.