

Macrolocalizacion de regiones con K-means

Índice

Carga de datos	1
Normalización de la base de datos	1
Número de clusters óptimo	1
Cómputo de las K-means	2
Flujos migratorios entre clusters	3
Definición de migrante según cluster de origen	5
Factores determinantes de expulsión de los migrantes	5

Carga de datos

Para comenzar, cargamos los datos de las variables sociales y economicas que definen a las distintas provincias de la Argentina.

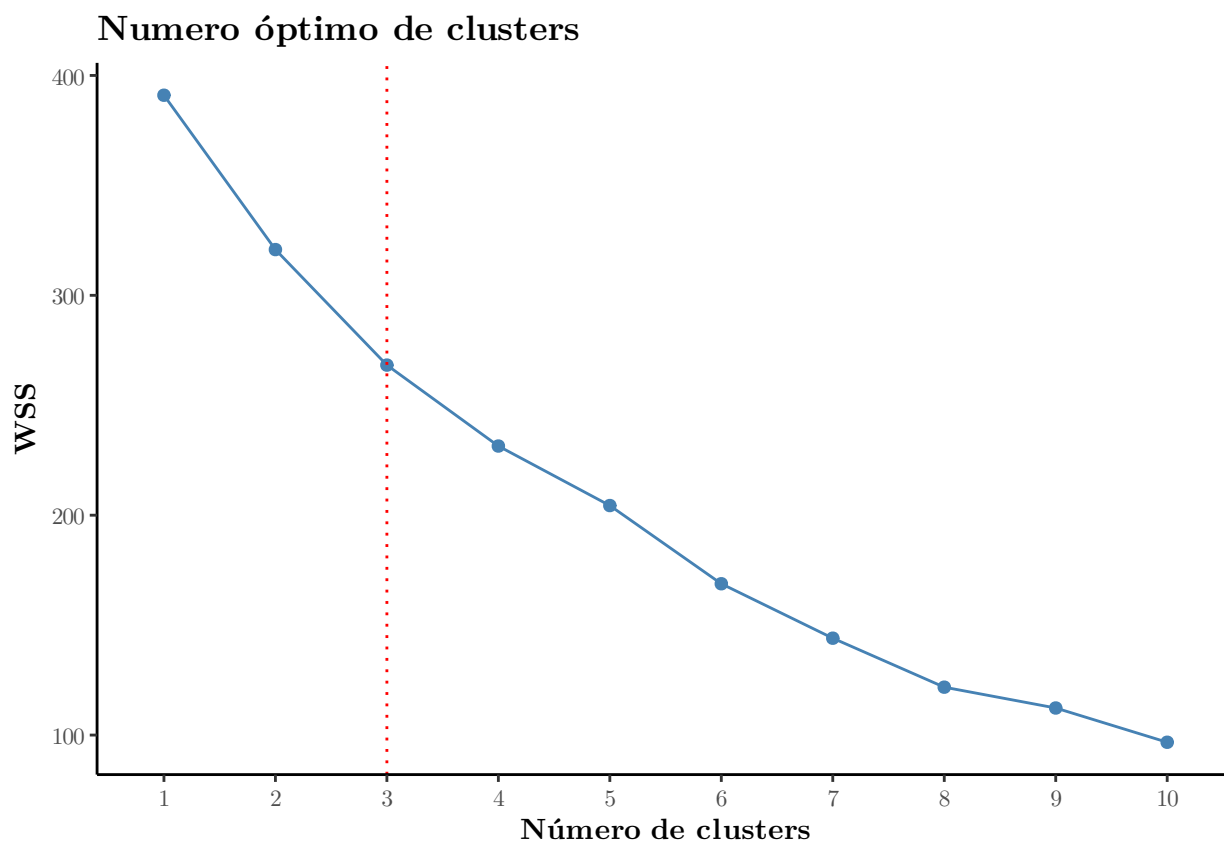
Normalización de la base de datos

Ahora normalizamos la base de datos para lograr coherencia en la aplicación de K-means, debido a que trabajamos con distintas escalas de datos.

Número de clusters óptimo

Verificamos el numero de clusters optimos a través de el Within-cluster sum of square, el cual es un indicador que mide la suma de las distancias entre las variables dentro de los cluster y sus centroides, la idea es minimizar esta discrepancia dentro de cada grupo, teniendo en cuenta el trade-off que implica cuando este es mínimo, el cual es el caso en que el numero de clusters es igual al número de variables a clusterizar, en donde si tenemos j observaciones, el numero de clusters sería tal que $j=k$, y no se estaría dando ninguna información relevante a los efectos de poder resumir características comunes entre los grupos.

$$WCSS = \sum_{i=1}^{N_C} \sum_{\mathbf{x} \in C_i} distancia(\mathbf{x}, \bar{\mathbf{x}}_{C_i})$$



Obtenemos que el numero óptimo de clusters se encuentra en torno al **número 3**, debido a que la contribución marginal de aumentar el numero de clusters a 4, no aportaría una reducción muy elevada al **WSS**, y seguiríamos perdiendo generalidad en la clusterización de los grupos sin una ganancia significativa de similitud dentro de los grupos.

Cómputo de las K-means

Setearemos una semilla para el cálculo de los centroides iniciales del algoritmo de K-means y procederemos a realizar el cálculo de los clusters a través de 1000 iteraciones, eligiendo la que de menor numero de WSS.

Podemos ver la media de los parámetros para cada uno de los clusters

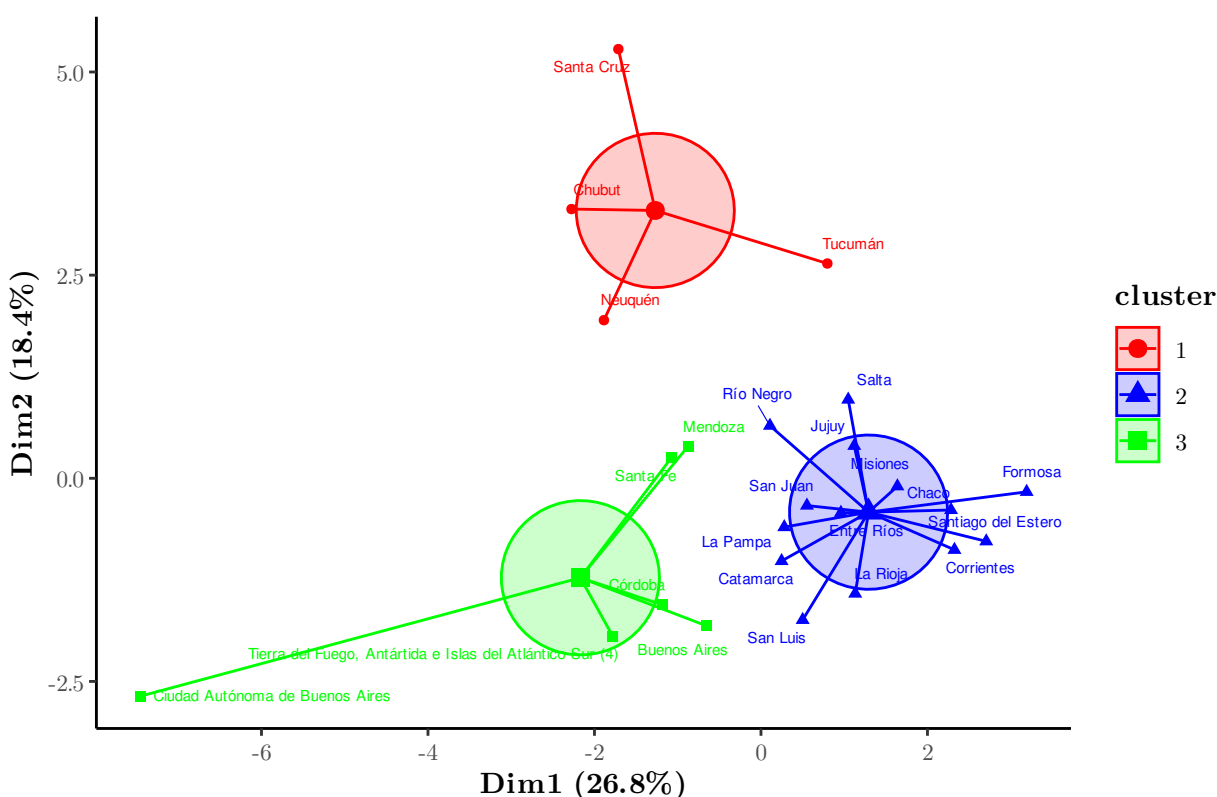
Cuadro 1: Resumen de indicadores por cluster

Indicadores	Cluster 1	Cluster 2	Cluster 3
Habitantes (2010)	530,187.00	677,181.00	3,042,344.00
Tasa promocion efectiva secundaria (2017)	76.25	78.60	82.05
Mortalidad infantil promedio 2016-2019	8.20	9.50	8.25
Homicidios dolosos C/ 100.000 hab. promedio (2016-2019)	7.25	3.75	5.30
Muertes en Accidentes Viales C/ 100.000 hab. promedio (2016-2019)	12.40	14.05	7.50
Robos (excluye los agravados) C/ 100.000 hab. Promedio (2016-2019)	973.40	654.85	1,566.95
Robos agravados C/ 100.000 hab. Promedio (2016-2019)	27.80	1.70	9.50
Violaciones C/ 100.000 hab. Promedio (2016-2019)	15.45	11.15	8.75
Exportaciones per-cápita en USD promedio (2016-2019)	2,857.20	794.20	1,420.55
Demanda de MWH energía electrica per cápita (2016)	3.80	2.20	3.35
Pobreza Promedio (2017-2019)	26.05	32.55	29.35
Tasa actividad Promedio (2017-2019)	44.55	42.50	46.00

Indicadores	Cluster 1	Cluster 2	Cluster 3
Cantidad empresas C/ 100.000 hab. (2016-2017)	1,569.62	846.96	1,771.95
Remuneracion real de trabajadores registrados del sector privado (2016-2019)	29,995.27	14,997.52	17,149.25
Porcentaje de empleados en Agricultura, ganadería, pesca y actividades extractivas (2016-2019)	0.21	0.13	0.05
Porcentaje de empleados en Comercio, servicios, electricidad, gas, agua y construccion (2016-2019)	0.67	0.69	0.71
Porcentaje de empleados en Industria (2016-2019)	0.10	0.15	0.22

agregamos a la base de datos original la denominación de cada cluster

Plot de clusters resultantes de la macrolocalización



Se puede ver claramente la diferencia en la similitud de los tres grupos, siendo solamente la Capital Federal la única que presenta mayor disimilitud con respecto a su cluster, lo cual indica que esta capital lleva una dinámica socio-económica muy peculiar con respecto al promedio de las provincias argentinas, inclusive considerablemente distinta que las provincias con la cual ostenta mayor similitud.

Flujos migratorios entre clusters

Dentro de los cluster, existen provincias de las cuales salen gran cantidad de inmigrantes, al igual que existen provincias que son receptoras de estos mismos, en primer lugar caracterizaremos cual es el cluster con mayor nivel de *expulsion* de inmigrantes interprovinciales.

Cuadro 2: Cluster de origen de los migrantes

Cluster	Porcentaje
1	10.41 %
2	61.67 %
3	27.92 %

Se puede notar que las provincias del **cluster N°2** son las que mayor nivel de expulsión poseen, seguidas por las provincias pertenecientes al cluster N°3, y por último se encuentran las pertenecientes al cluster N°1.

Cuadro 3: Cluster de destino de los migrantes

Cluster	Porcentaje
1	7.85 %
2	12.83 %
3	79.32 %

Sin embargo, viendo cuales son los cluster de destino con mayor porcentaje de migrantes, encontramos que el **cluster N° 3** es el que mayor nivel de *atracción* posee por elevada diferencia (79.32 %)

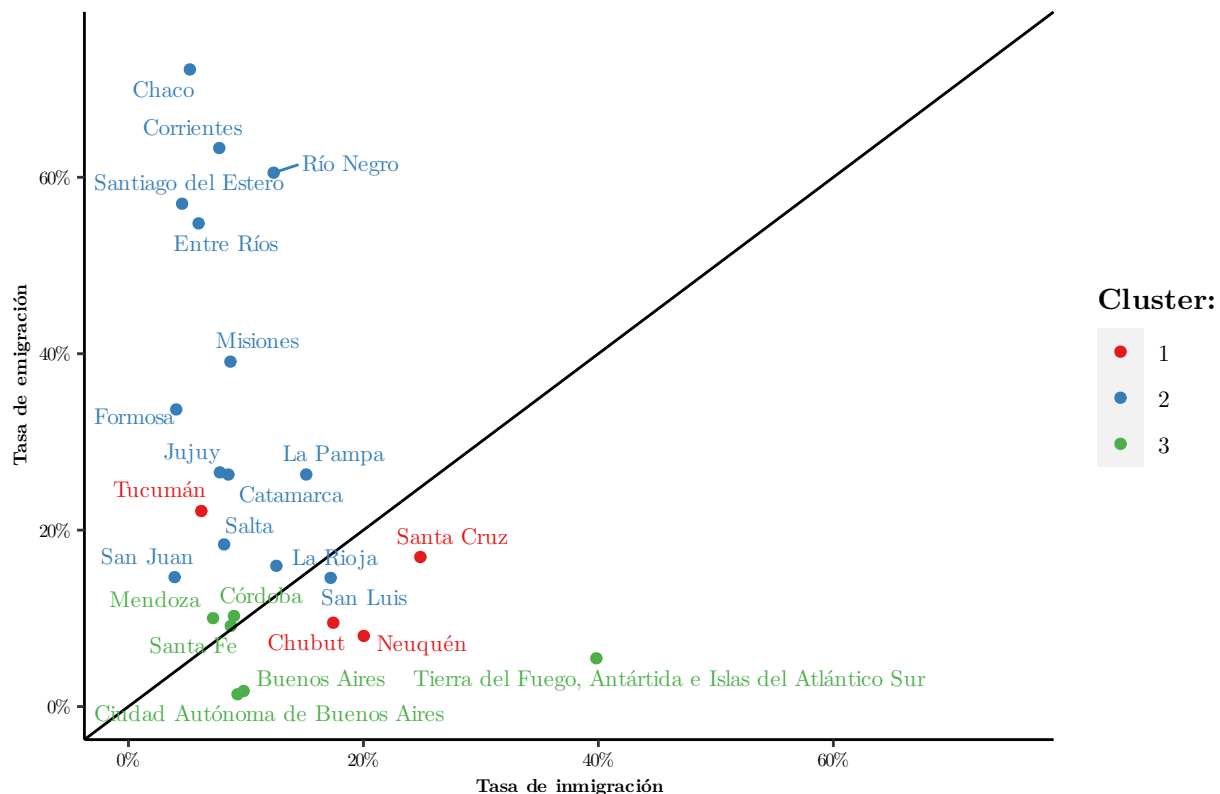
```
## Joining, by = "Provincia"
```

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "Provincia"
```

```
## Joining, by = "Provincia"
```

Tasa de inmigración y emigración por provincia



Definición de migrante según cluster de origen

En primer lugar se analizará si existen microdeterminantes dentro de las características de los migrantes que provoquen que ciertas personas con tengan una mayor propensión a migrar desde ciertas localizaciones, es decir, se analizaran los factores de expulsión de los distintos cluster; por otro lado, se realizará un análisis diferenciando el destino de los distintos migrantes, con el fin de establecer los determinantes de atracción de las distintas localizaciones.

Factores determinantes de expulsión de los migrantes

En el universo de habitantes de una determinada región/cluster, pueden existir características determinadas que diferencien de manera significativa entre las personas que nacieron y decidieron quedarse en la misma región, de las que si bien nacieron en la misma región, optaron por migrar hacia otra. A continuación se realizará un análisis diferenciando por cluster de características socioeconómicas de las personas que nacieron en un mismo cluster, con el fin de determinar si existen diferencias entre los migrantes y los no migrantes.

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "CH15_COD"
```