DSI – Trabajo de Clustering

Luis Cabañero Gómez

Problema

- Buscar patrones en los datos de la renta por municipios de 2007
- Obligatorio: Mapas de Kohonen

Preprocesado

- Python
  - Numpy
  - Pandas
  - Scikit-learn
  - Matplotlib
  - Seaborn

Limpieza de datos

- Campos innecesarios:
  - Código INE municipio (5 dígitos)
  - Código INE Provincia
  - Código INE Comunidad Autónoma
- Filas vacías: 8

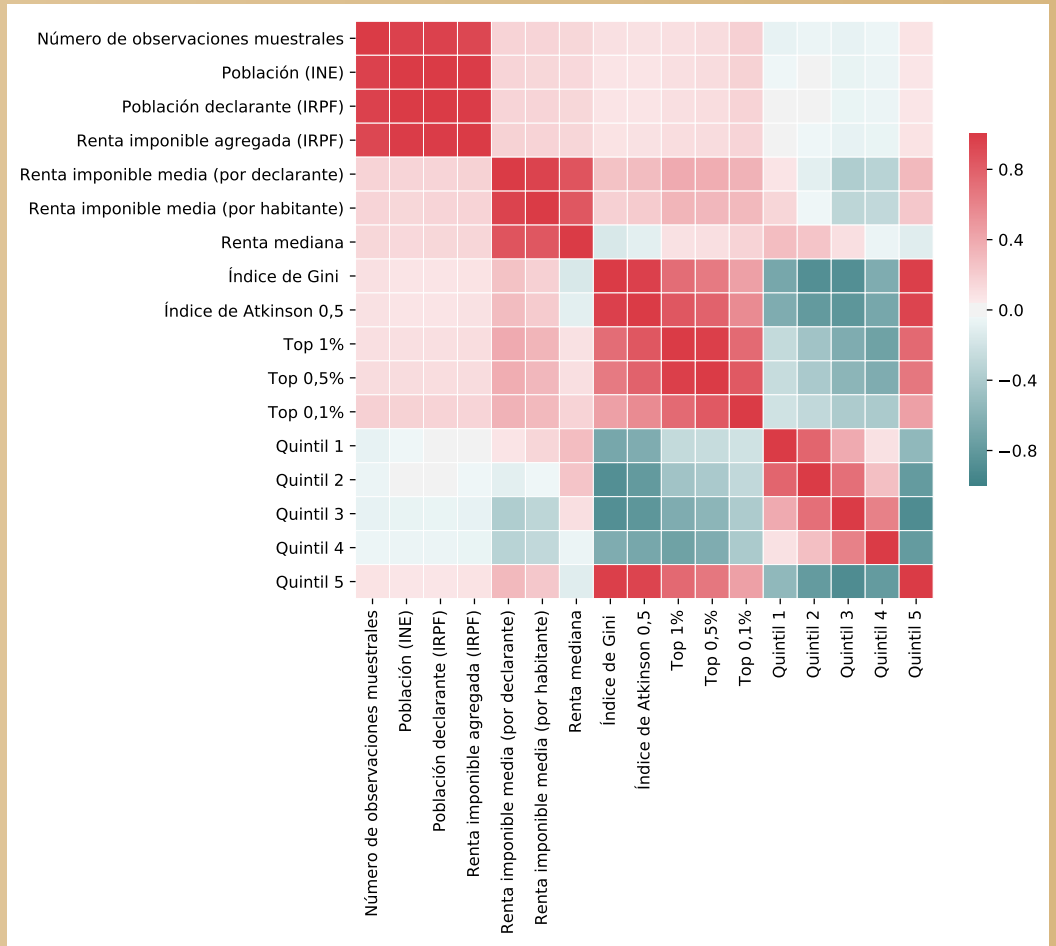
Añadir coordenadas geográficas

- Set de datos con código INE y coordenadas
- Join entre ese set y los datos
- Eliminación del campo “Código INE municipio”

Selección de campos

- Limitar los campos que se usarán para el clustering
- En función de la semántica y de las correlaciones

Selección de campos

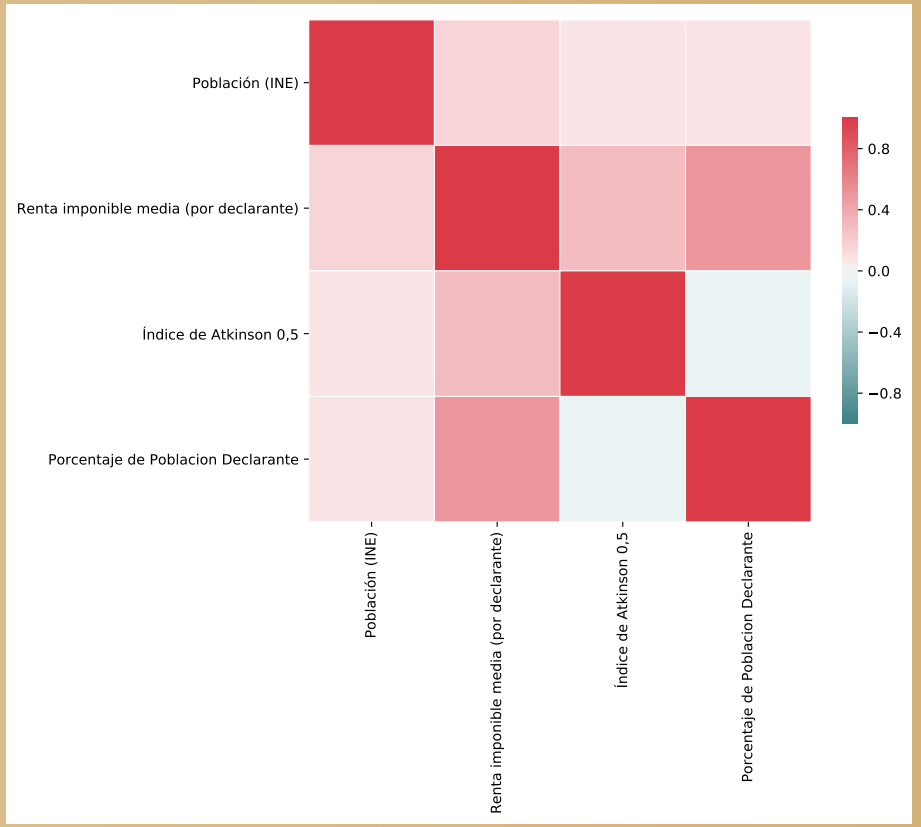


Selección de campos

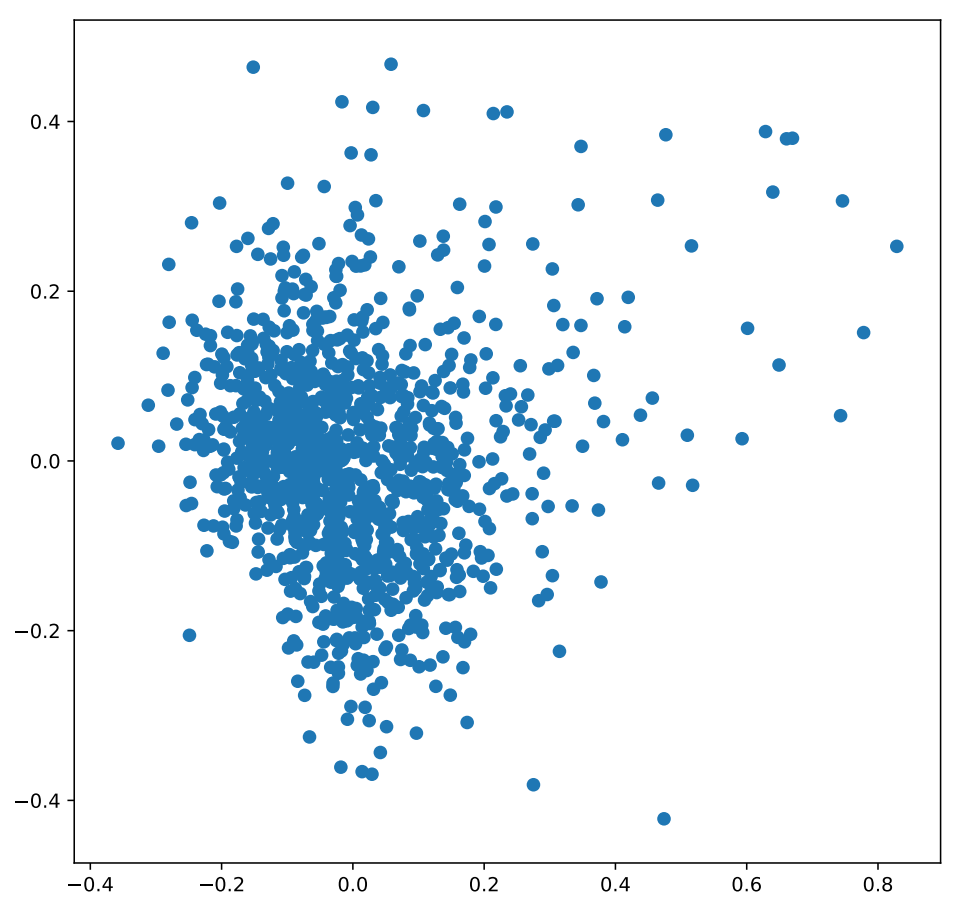
- Número de observaciones descartado
- Gini descartado
- Quintiles y top descartados
- Porcentaje de declarantes creado
- Renta media por declarante

Campos seleccionados

- Población INE
- Renta imponible media por declarante
- Índice de Atkinson 0.5
- Porcentaje de Poblacion Declarante



PCA



Clustering

- R
  - Kohonen

Clustering: preparación de datos

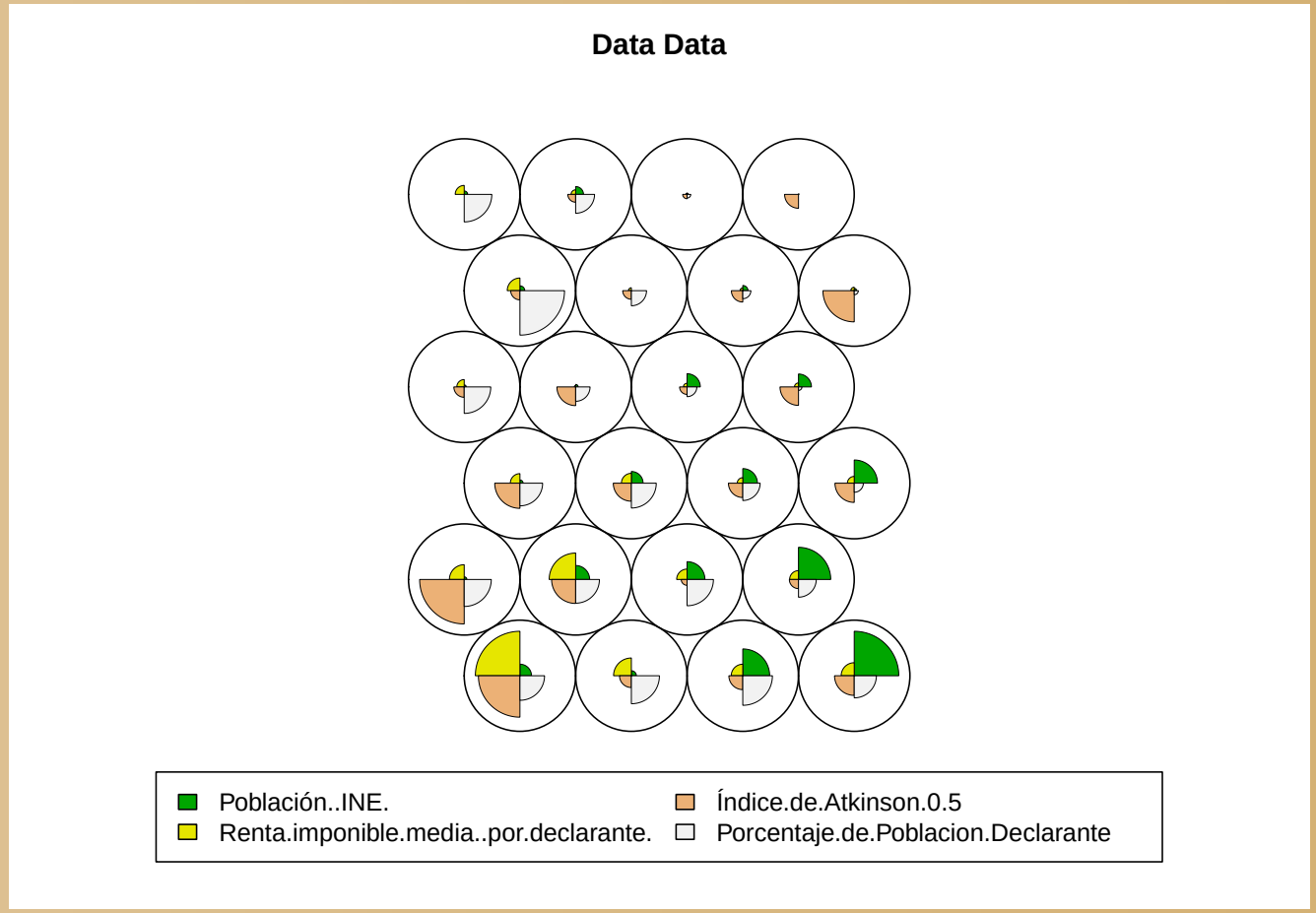
- Logaritmo en base 10 sobre Población
- Normalización min-max

Clustering: Parametrización

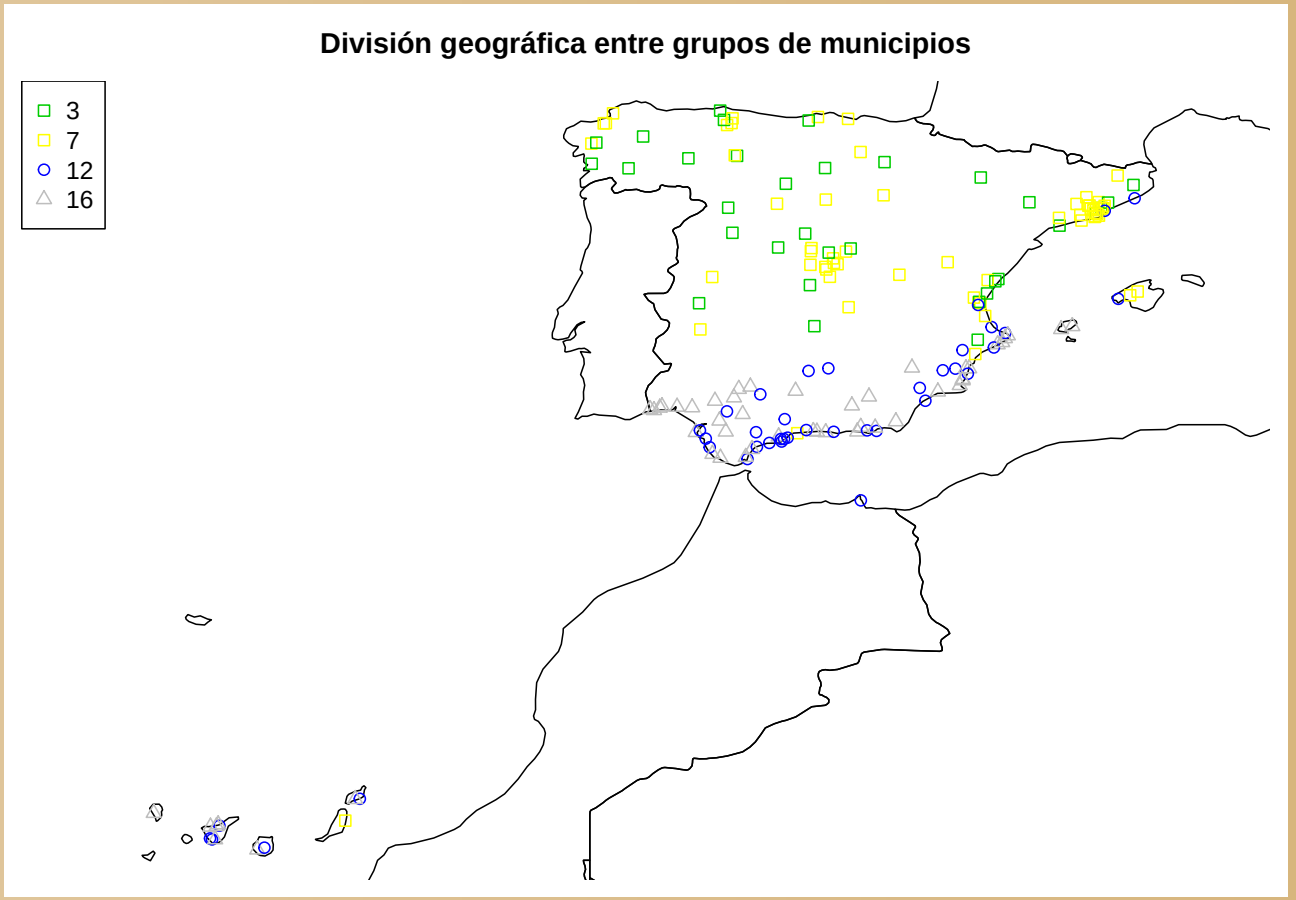
- 6x4 neuronas
- Hexagonal
- No toroidal

Clustering

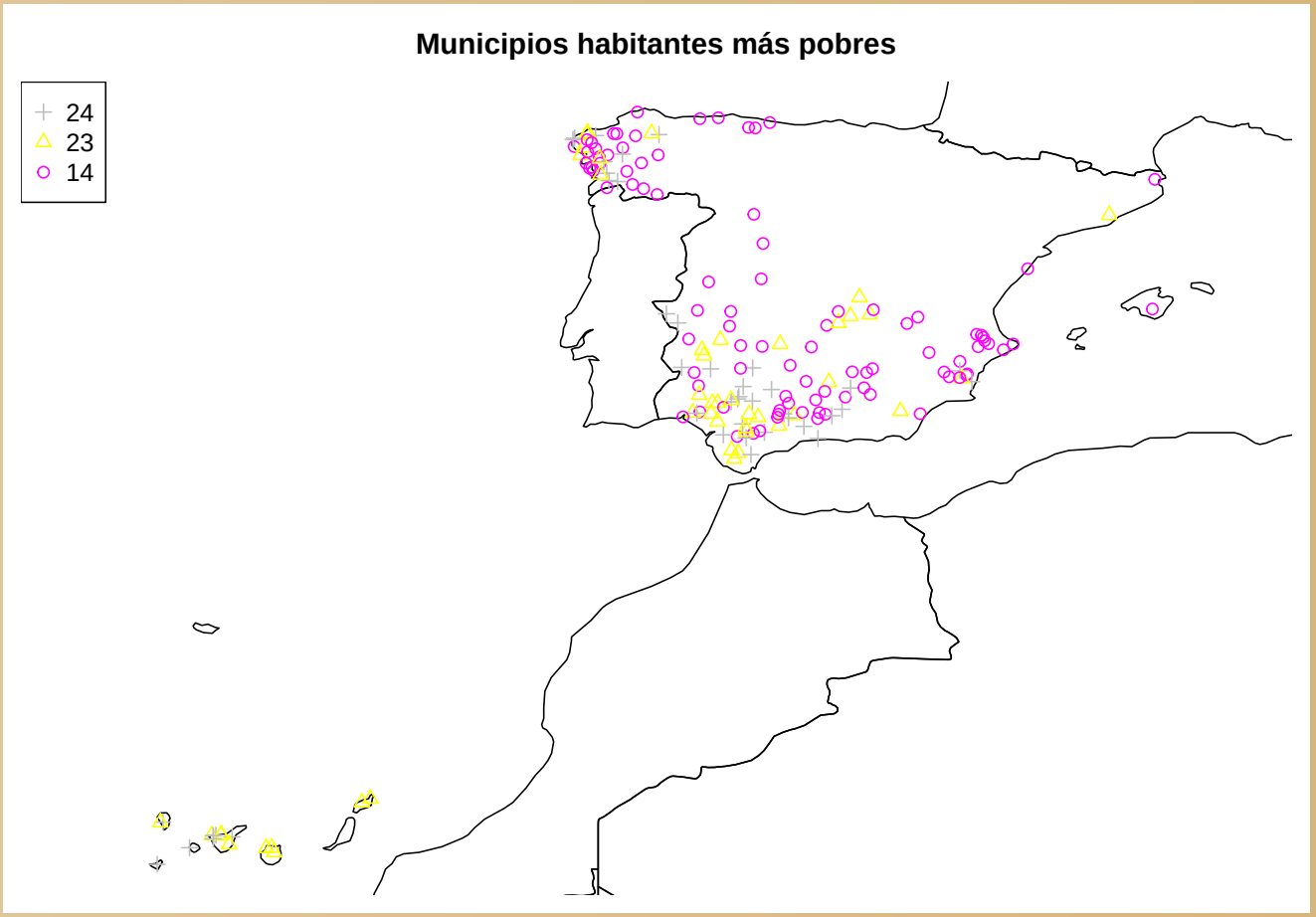
- Grupo con más población
- Mayor renta y mayor desigualdad
- Grupo con mayor desigualdad y poca población
- Grupo con mucha población declarante



Clustering



Clustering



Clustering