Clustering Spanish Provinces

Making easier region choosing within a country

Alejandro González Casal

March 9, 2021

1. The problem of choosing a region

- Nowadays, people need to move to other countries due to work or other circumstances.
- Even if the country is clear, the specific region is not.
- The large number of options may be intractable.
- A clustering algorithm could reduce the alternatives to evaluate and facilitate the search process.

2.1. Data collection

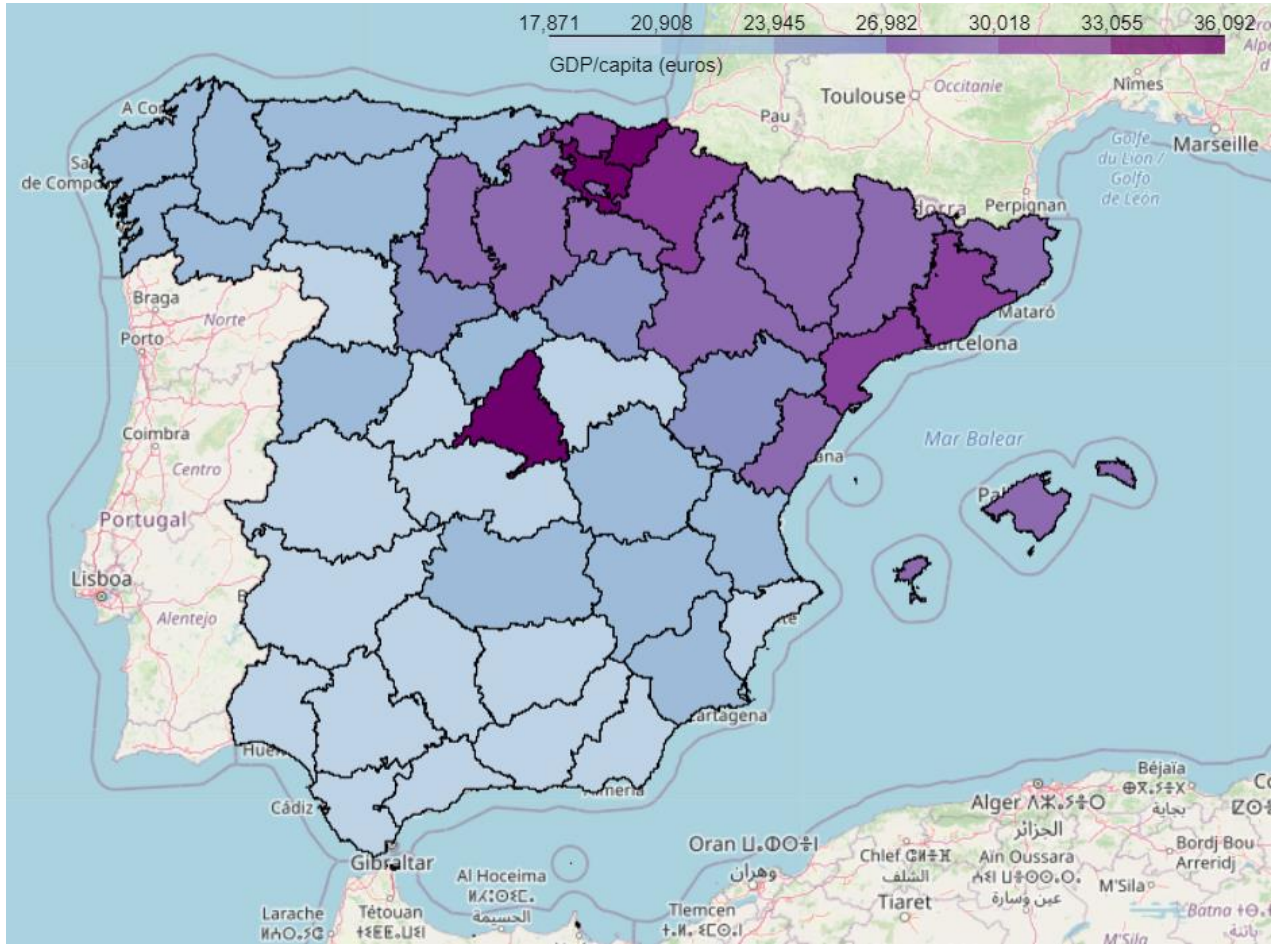
- Province global GDP and per capita GDP from Wikipedia scrapped from its page.
- Province population and area from Wikipedia scrapped from its page.
- Industrial States and Power Plant of each province obtained from Foursquare API.
- Some additional data such as autonomous community of each province, geographical coordinates from Geopy and equivalent radius of the province.

2.2. Data cleaning

- Adapting province names to match between datasets and also with geojson file used for Choropleth maps.
- Missing values detection and replacement with the actual values from the internet.
- Data type selection according to each feature necessities.
- NaN replacing with 0 in venue's data.

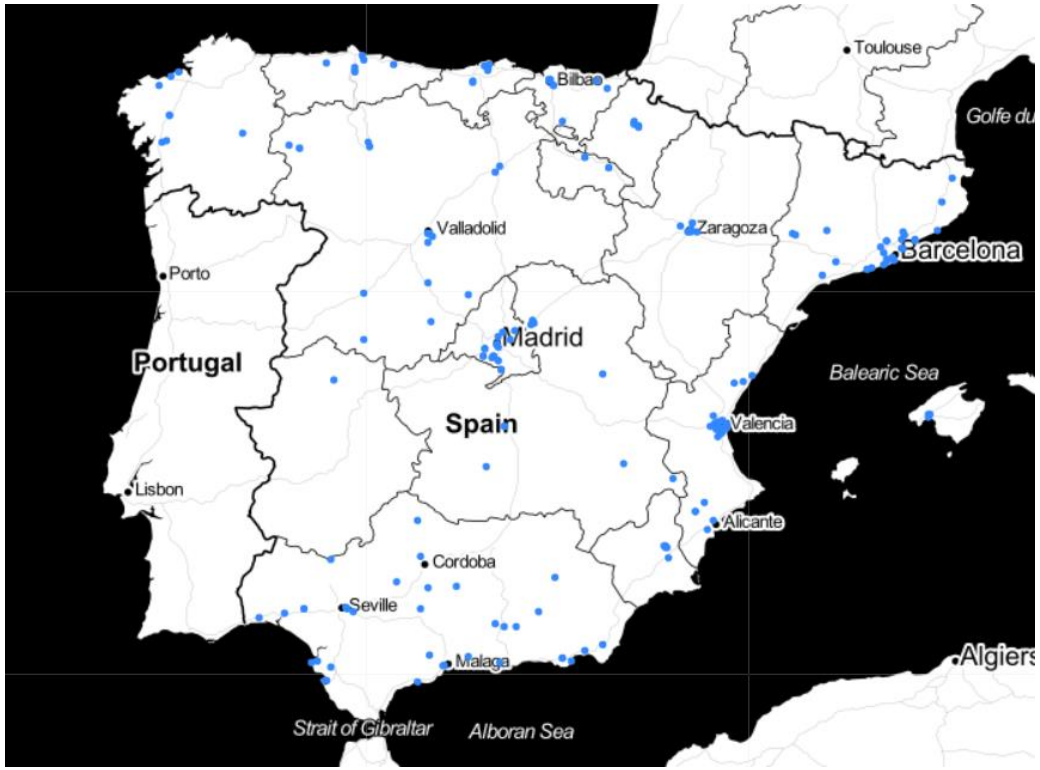
3.1 Geograhical distribution of GDP/capita

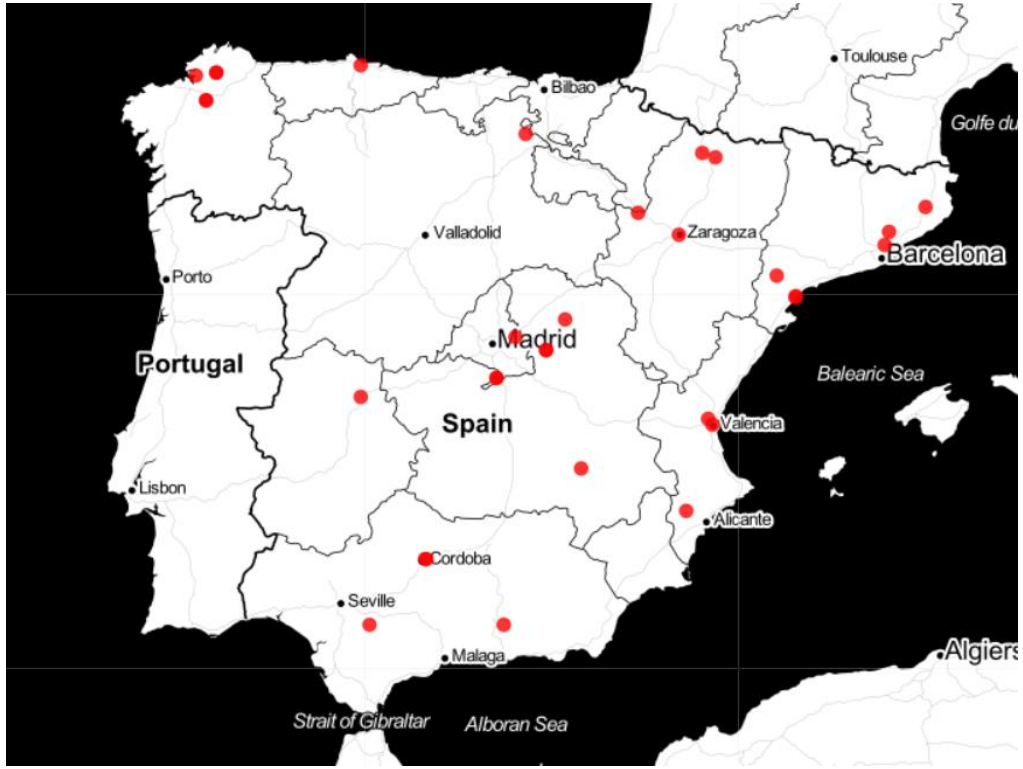
- Higher values are grouped in northeast corner and Madrid.



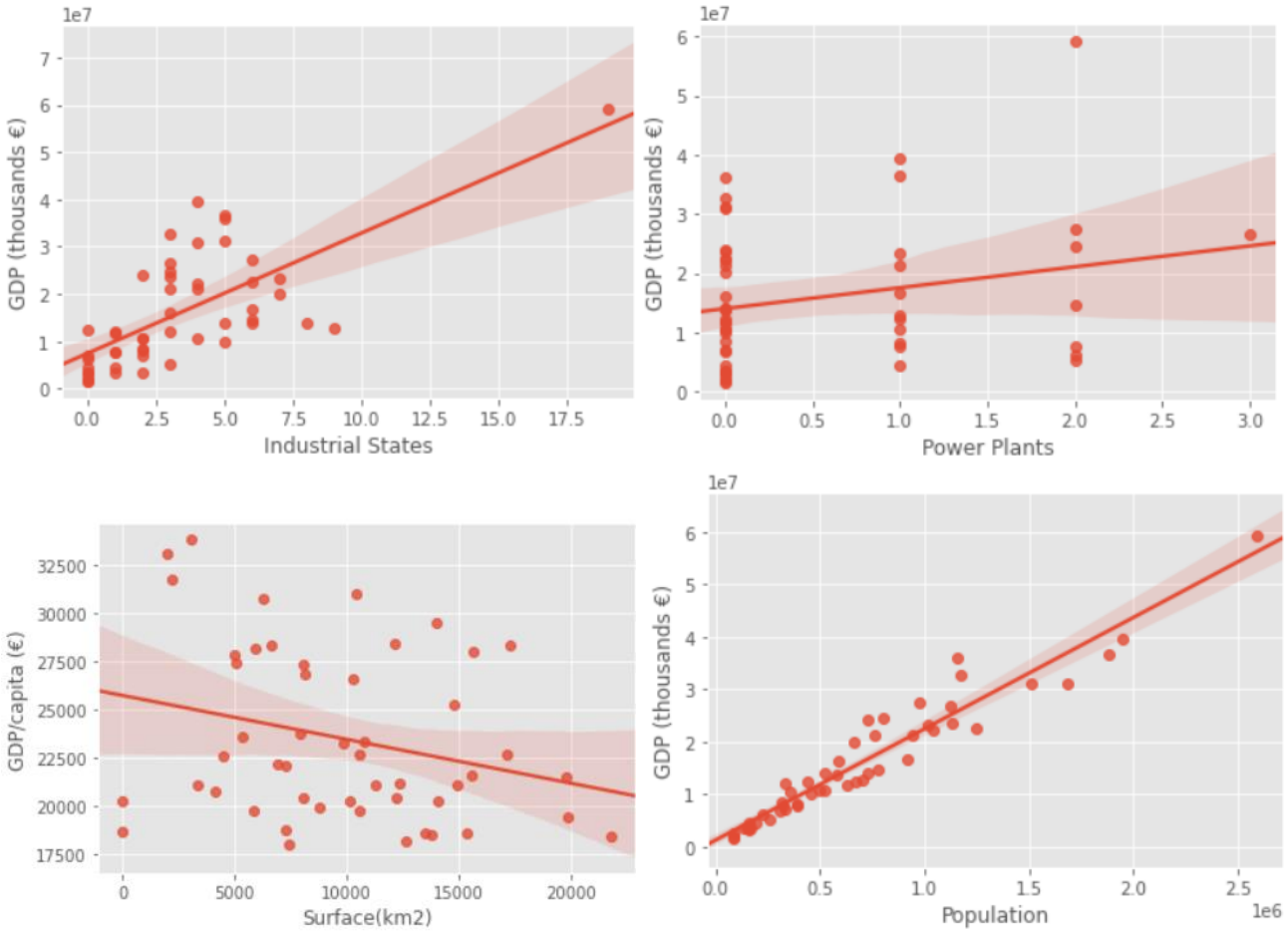
3.2. Industrial States and Power Plants location

Industrial States are placed near the significant cities, similar happens with Power Plants but in that case A Coruña (northwest corner) also gather several plants.





3.3 – Regression Analysis

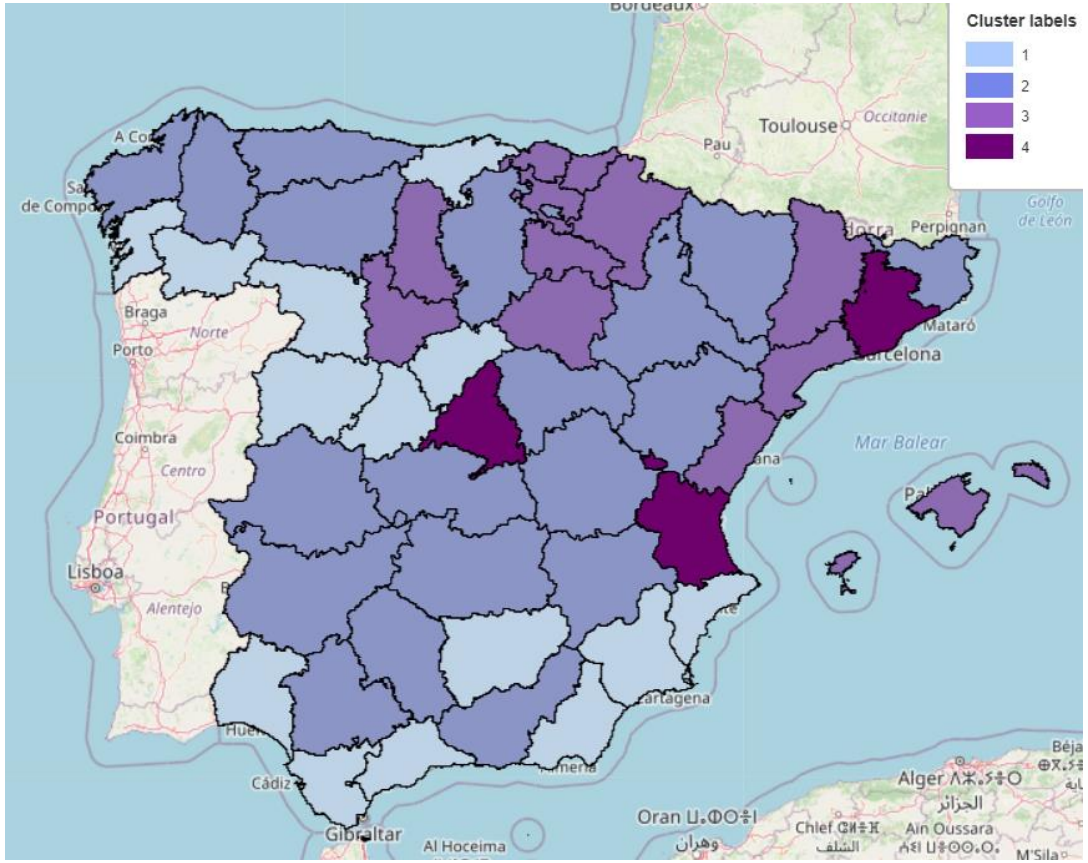


- The strongest correlation are between global GDP and Population.
- Global GDP is also highly related with the number of Industrial States, but not clearly with number of Power plants.
- Finally, as the area of a province increases its GDP per capita tend to decrease.

4.1. Clustering implementation

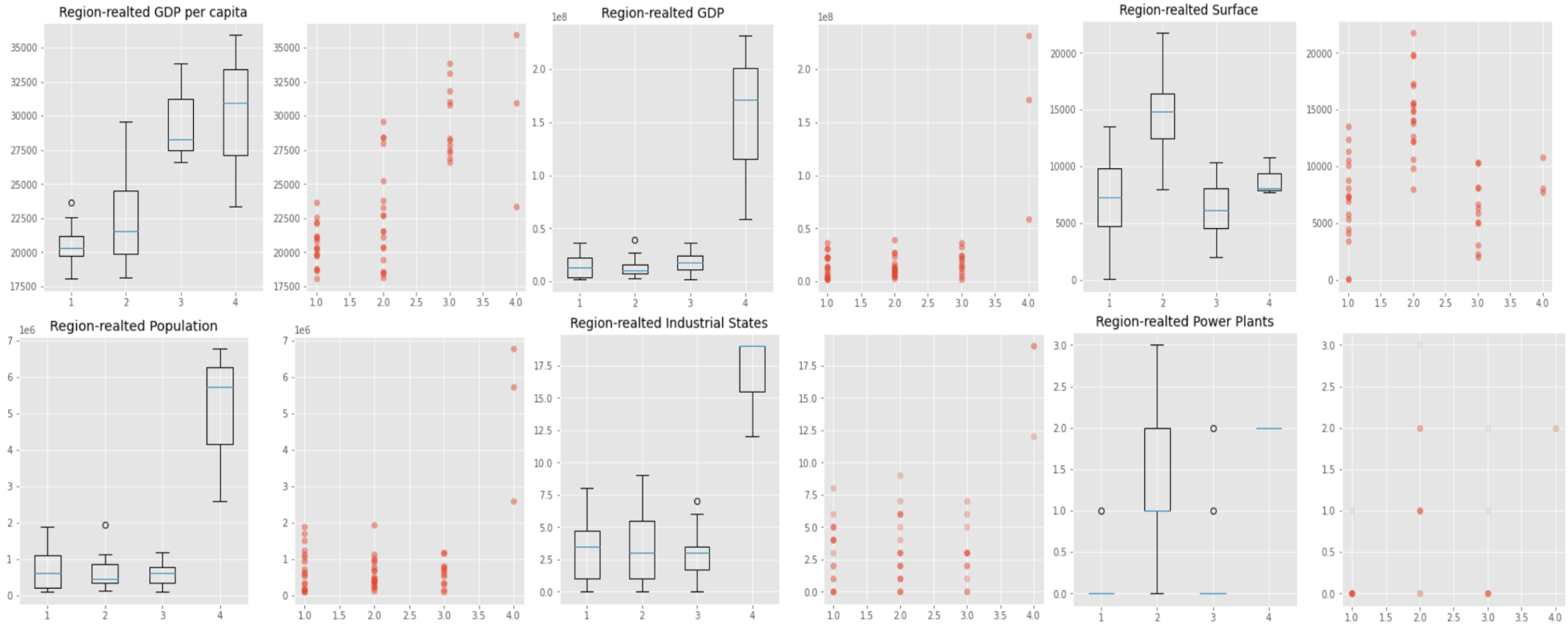
- The selected features were: GDP per capita, Global GDP, Population, Area, Nº Industrial States and Nº Power Plants.
- Features were scaled to avoid biases between them.
- To perform the clustering Kmeans, an unsupervised classification method, was used.
- Number of cluster was set at four after some tests.

5.1. Clustering analysis – Choropleth map



- Cluster 4 is composed of three provinces Madrid, Barcelona and Valencia.
- Cluster 3 includes provinces with the highest GDP per capita excluding, admittedly, the one includes in cluster 4.
- Cluster 1 and 2 englobe the remaining provinces, differences between both Will be showed further on.

5.2. Cluster-related feature distribution (1/2)



5.2. Cluster-related feature distribution (2/2)

As was previously noted these graphs show that cluster 4 and 3 have clearly the highest values of GDP per capita, the difference between this two clusters is the global GDP where cluster 4 overtake the other clusters significantly. A similar pattern appears in terms of population and industrial states number.

As for cluster 1 and 2, again is shown their lower values in terms of GDP per capita. They also have similar values to cluster 3 in population, global GDP and industrial states numbers. The difference between these two similar clusters is in terms of area and power plants where cluster 2 has higher values.

5.3. Cluster labeling

According to the previous visualizations these could be suitable labels for each cluster:

1 → Low GDP provinces with low area and few power plants.

2 → Low GDP provinces with high area and several power plants.

3 → High GDP per capita provinces with low global GDP, population and few industrial states.

4 → High global and per capita GDP with high population and much industrial states.

6. Conclusion and future directions

- The new province classification will make easier the region choosing within Spain.
- Similar analysis could be applied to other countries or regions.
- More data and other classification algorithm could be used for further analysis.