https://towardsdatascience.com/back-to-the-municipality-clusters-celestial-objects-and-maps-55c64e39f805

Fernando Barbalho

13/1/2021

Back to the municipality clusters: celestial objects and maps

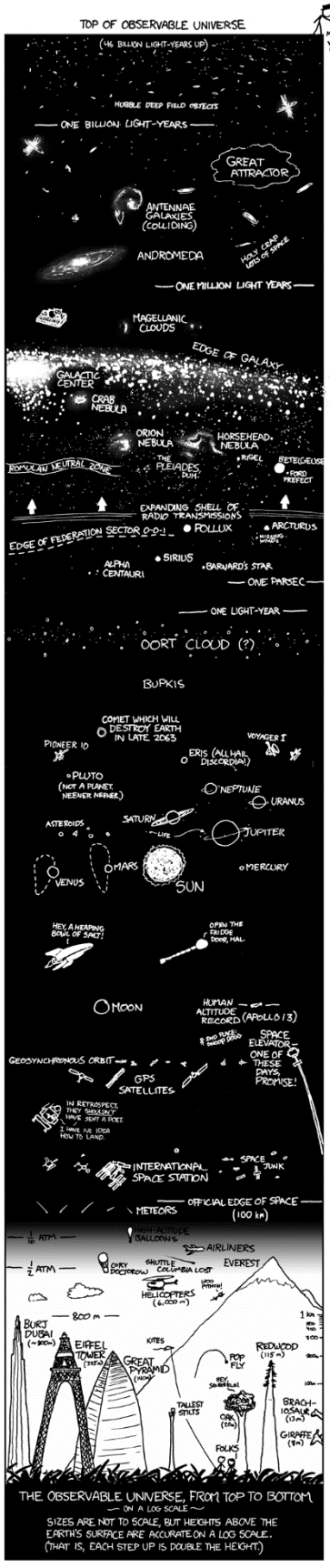
It was in August of the fateful year 2020 that I wrote a text right here in the Medium showing the findings of a model of clusters of municipalities from variables related to GDP. The time has come to visit that model again and make new experiments. Here the reports will focus on findings that are easily interpreted in visual mode.

The idea is to detail the clusters in distribution graphics and maps. With the help of the XKCD astronomical inspiration and its black-silver palette, we assembled the graphics, solving at least visually the distribution problems of the variables.

The galaxies, the nebulae and the logarithms

The journey begins with the stars. The data used here present a huge distribution problem. We compare cities as different as Xapuri and São Paulo. In order to have some success in the interpretation of the graphs, we borrowed from XKCD the idea of representing the distances of the various celestial bodies from the Earth. A huge graph in logarithmic scale allows to put in the same figure the height of the giraffe and the astronomical distances to the nebulae and galaxies far far away from here.

Gradually roll the screen and make the inverse path between the most distant objects observed by the Hubble telescope and the ground on which we step. Be careful not to fall.

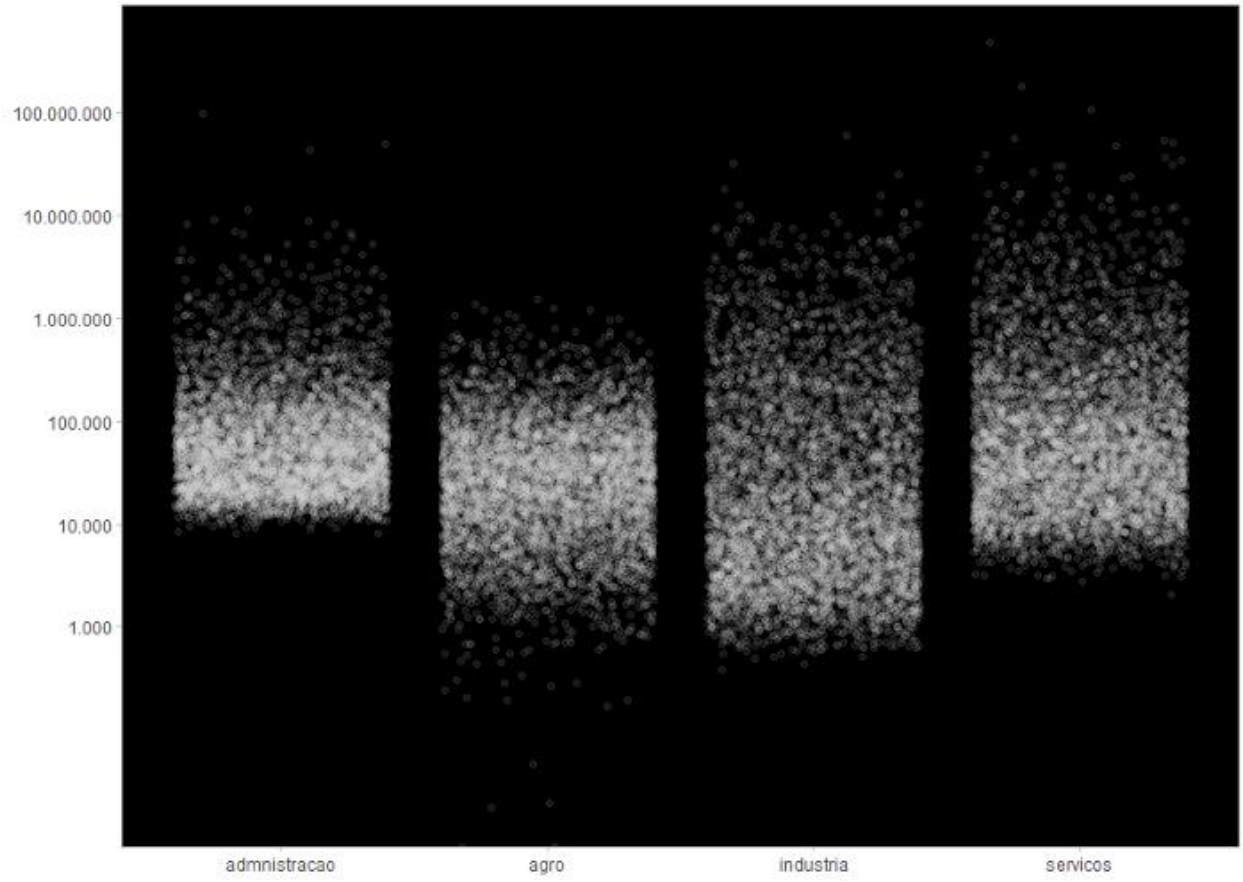

TOP OF OBSERVABLE UNIVERSE
(46 BILLION LIGHT-YEARS UP)
MADUK!
HUBBLE DEEP FIELD OBJECTS
ONE BILLION LIGHT-YEARS
GREAT ATTRACTOR
ANTENNAE GALAXIES (COLLIDING)
ANDROMEDA
HOLY GRAIL AREA OF SPACE
ONE MILLION LIGHT YEARS
MAGELLANIC CLOUDS
EDGE OF GALAXY
GALACTIC CENTER
CRAB NEBULA
ORION NEBULA
HORSEHEAD NEBULA
THE PLEIADES (M45)
RIGEL
BETELGEUSE
FORD PREFECT
ROMULAN NEUTRAL ZONE
EXPANDING SHELL OF RADIO TRANSMISSIONS
EDGE OF FEDERATION SECTOR 0-0-1
POLLUX
ARCTURUS
HOWARD WINDS
ALPHA CENTAURI
SIRIUS
BARNARD'S STAR
ONE PARSEC
ONE LIGHT-YEAR
OORT CLOUD (?)
BUPKIS
COMET WHICH WILL DESTROY EARTH IN LATE 2063
PIONEER 10
ERIS (ALL HAIL DISCORDIA!)
VOYAGER I
PLUTO (NOT A PLANET, NEENER NEENER)
NEPTUNE
URANUS
ASTEROIDS
SATURN
LIFE
JUPITER
MARS
MERCURY
VENUS
SUN
HEY, A HEAPING BOWL OF SPACE!
OPEN THE OUTSIDE DOOR, HAL.
MOON
HUMAN ALTITUDE RECORD (APOLLO 13)
SPACE ELEVATOR — ONE OF THESE DAYS, PROMISE!
GEOSYNCHRONOUS ORBIT
GPS SATELLITES
IN RETROSPECT, THEY SHOULDN'T HAVE SENT A POET.
I HAVE NO IDEA HOW TO LAND.
INTERNATIONAL SPACE STATION
SPACE JUNK
OFFICIAL EDGE OF SPACE (100 km)
METEORS
1/10 ATM
HIGH ALTITUDE BALLOONS
AIRLINERS
1/2 ATM
EVEREST
SHUTTLE COLUMBIA LOST
DAVY CROCKOW
HELICOPTERS (6,000 m)
LOCH NESS MONSTER
800 m
BURJ DUBAI (~800m)
EIFFEL TOWER (324m)
KITES
GREAT PYRAMID (140m)
POP FLY
REDWOOD (115 m)
KEY SUBMARINE
TALLEST STILTS
ORK (30m)
BRACHIOSAUR (13m)
GIRAFFE (4m)
FOLKS
THE OBSERVABLE UNIVERSE, FROM TOP TO BOTTOM
— ON A LOG SCALE —
SIZES ARE NOT TO SCALE, BUT HEIGHTS ABOVE THE EARTH'S SURFACE ARE ACCURATE ON A LOG SCALE. (THAT IS, EACH STEP UP IS DOUBLE THE HEIGHT.)

XKCD vision of distances in the universe: https://xkcd.com/482/

The nebulae of the GDP components and GDP per Capita

The variables evaluated here will always be the components of the sectors of the economy: agriculture, administration, industry and services. We also analyze the per capita GDP data. Remembering that all the data refer to variables of Brazilian municipalities.

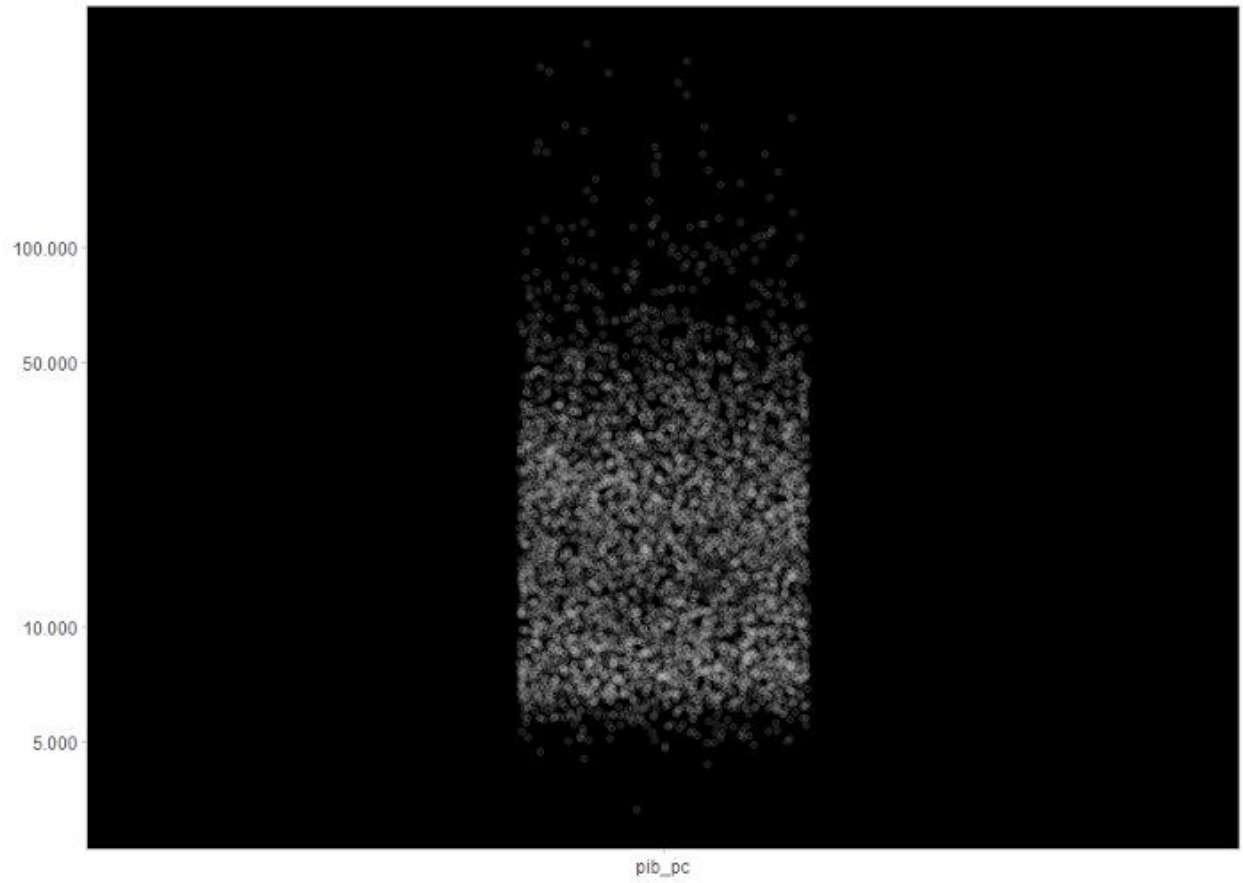
When the distribution of the variables in logarithmic scale is done, one can perceive the formation of images that recall the nebulae described in astronomy texts.



Distribution of GDP components by municipalities in logarithmic scale (values in R$ thousand). Image by Author

In the figure above it can be seen that:

- The administration variable presents a large concentration of points in the range between 10,000 and 100,000, and then there is a large dispersion until it is close to 100 Mi.
- The other variables also present concentration ranges, although less dense.
- The variable that is most dispersed is industry, the one that is responsible for the lowest values is agriculture and the one that reaches the highest values is services.



Distribution of GDP per capita by municipalities in logarithmic scale (values in R$ 1). Image by Author

The analysis of the GDP per capita is always done in a separate graph given the difference in scale, since it is a value that is a ratio between GDP and population. The figure above shows a concentration that starts a little above the R$ 5,000 mark and reaches the R$ 50,000 mark and then the points are dispersed.
These dispersion characteristics are fundamental to determine the six clusters formed. In the following section we see how we fill the nebula points with colors related to each of the clusters.

Colouring the nebulae from the clusters

The Brazilian inequality expressed in the values of the variables ends up being the determinant of the six clusters. In general, the lower values of the GDP and GDP per capita components are represented in clusters 1, 2 and 3, while the higher values appear in clusters 4, 5 and 6. With this characteristic it is always possible to make cluster analyses from two superclusters. This allows a cleaner visual analysis, which in turn helps to better identify the characteristics of the clusters.

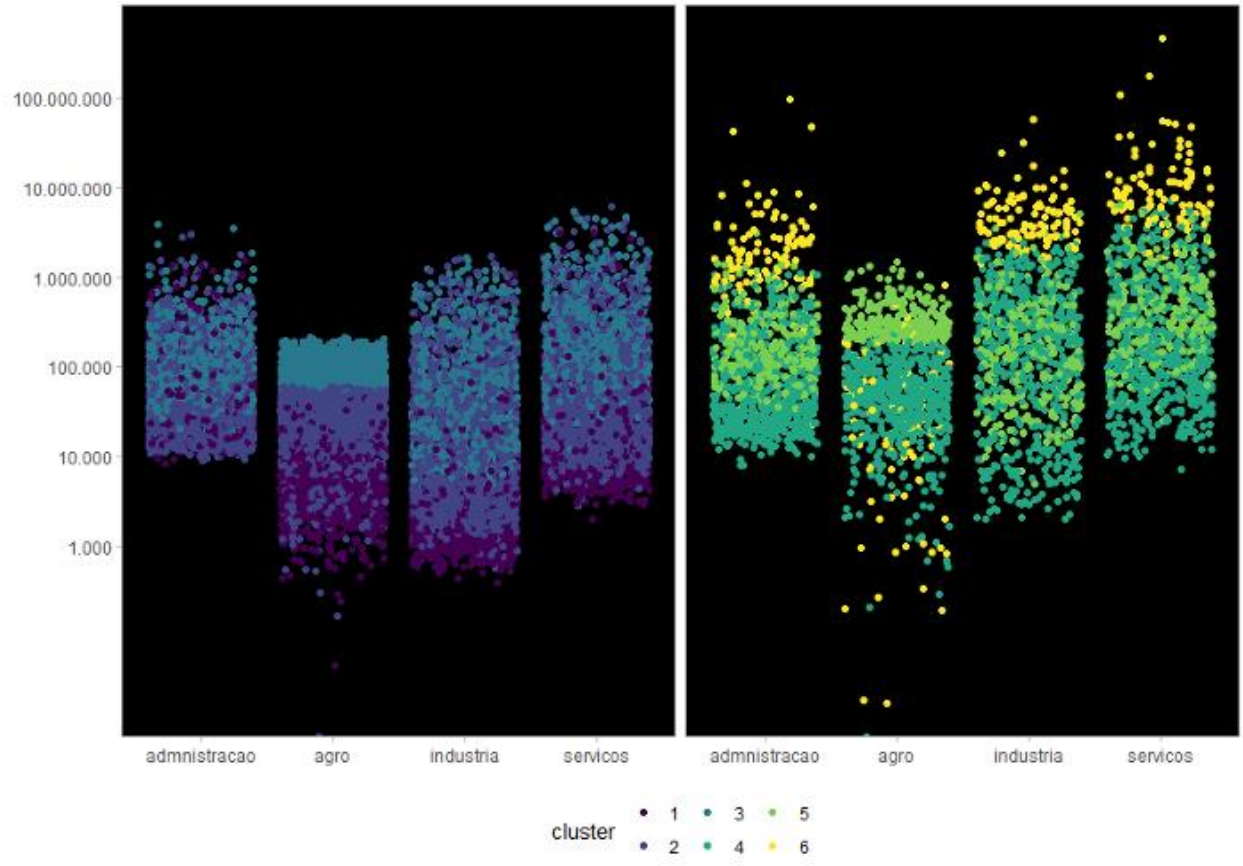


Image by Author

In the figure above it can be seen:

- Clusters 4, 5 and 6 are distributed in larger value ranges than the others
- Clusters 3 and 6 are concentrated in the agriculture sector
- Cluster 6 presents the highest values in sectors: administration, industry and services

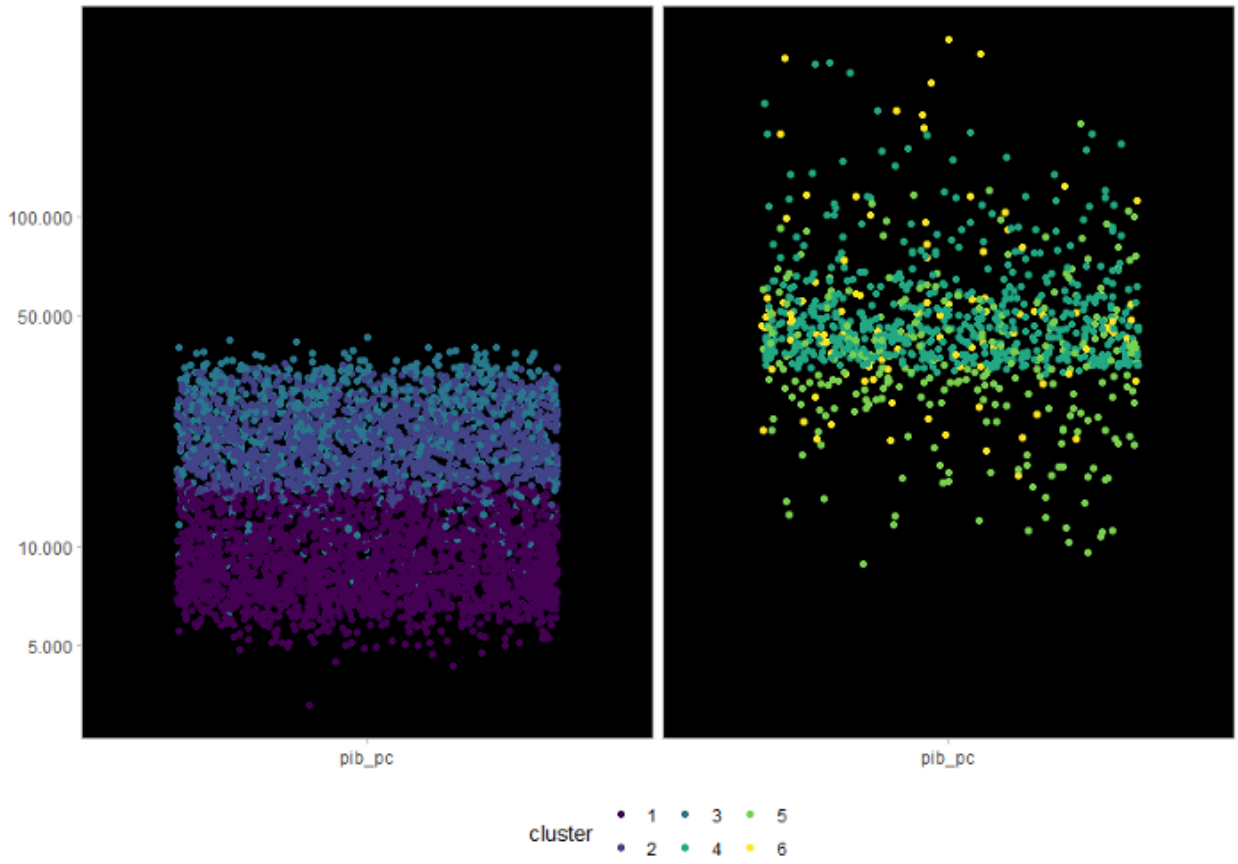


Image by Author

When it comes to GDP per capita, it can be seen in the figure above:

- Cluster 1 is densely situated in the lowest values
- Cluster 5, despite being dispersed in the GDP components, is perceived in the per capita GDP analysis as occupying the range of the highest values

The medoids in the six-dimensional centers

The clustering algorithm used for this work is characterized by the possibility of identifying the so-called medoids. This strange name is associated to the point at the center of the clusters formed in this case by six dimensions: the five variables we analyze here and the total GDP.

The idea of medoids can be compared with the average and median in the descriptive statistics, that is, the identification of an element that best characterizes a set of data. In the case of our study, medoids are the municipalities that from the point of view of the analyzed variables best characterize each of the six clusters.

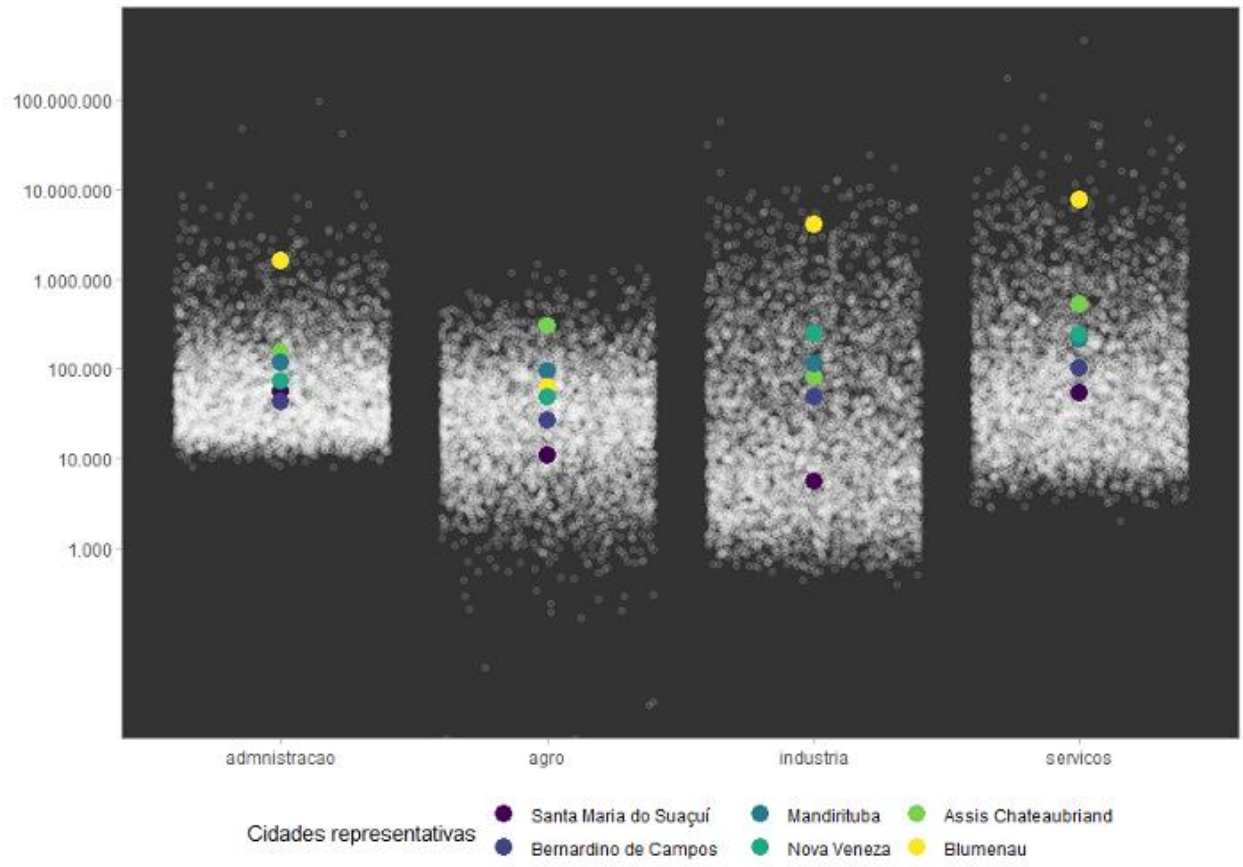


Image by Author

When we analyze the medoids of the economic sectors, we realize that:

- Blumenau, which represents Cluster 6, is far apart of the other medoids with the exception of the agricultural sector. In this case it comes in third place.
- Assis Chateaubriand, which characterizes Cluster 5, leads the agricultural sector with great difference to the others.
- The administration sector is where the medoids of clusters 1 to 5 are most similar to each other.

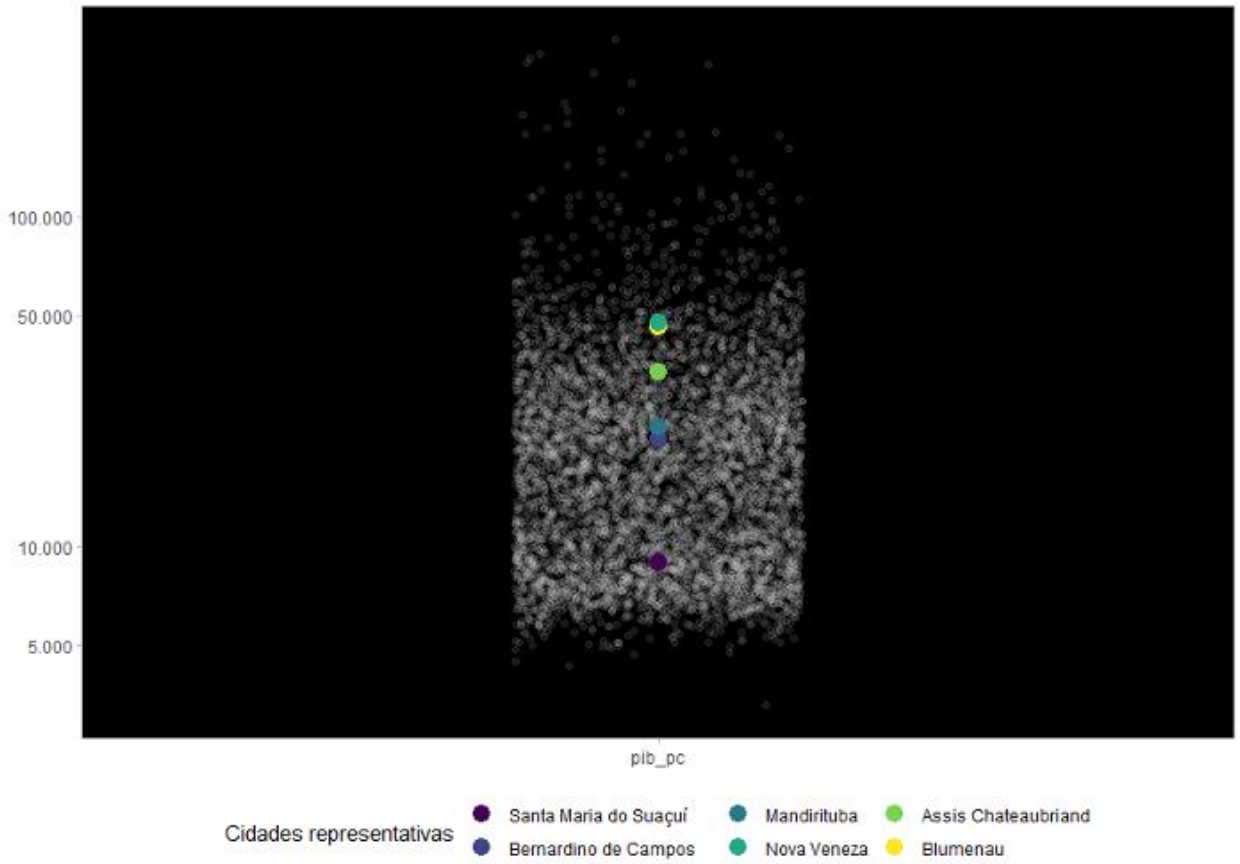


Image by Author

When analyzing the GDP per Capita, one realizes that:

- Blumenau and Nova Veneza (medoids of cluster 4) are quite close
- Santa Maria do Suaçuí, the medoids of cluster 1 is completely displaced from the others, presenting very low value.

Back to Earth and cartographic representations

Now discussing the cartographic representation, what are the impacts of the findings on clusters and their distribution characteristics across municipalities? In the August text we did the first experiments with maps. Now we go to another one that shows the impact of the analysis of a cluster from its medoids.
But first we will paint on the map of Brazil the clusters with their respective municipalities.

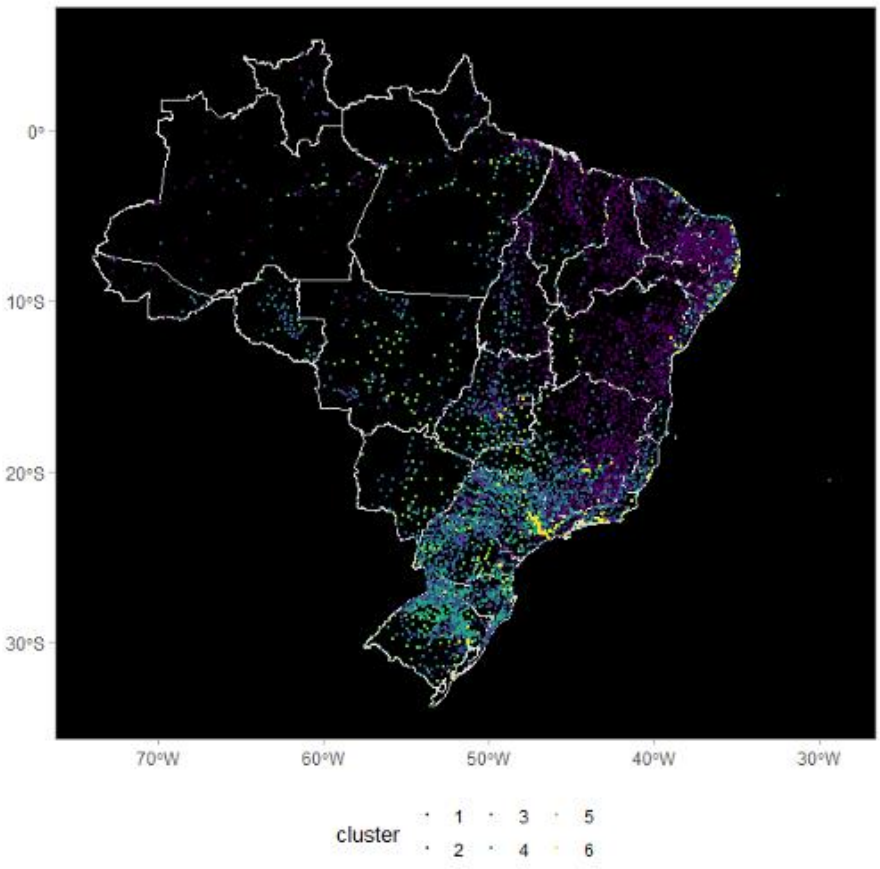


Image by Author

The result of the algorithm I used to identify the clusters informs that group 1 is the most cohesive. It is observed that this cohesion is also geographical. The Northeast states concentrate most of the municipalities in this cluster. The north of Minas Gerais also has a large concentration of municipalities with this characteristic. It is worth remembering here that this cluster has the lowest values of the five variables, especially per capita GDP, which reinforces the condition of poor income distribution and poverty that affects these regions of Brazil.

The map also shows the wealth belt in the state of São Paulo and the dispersion of the other clusters mainly in the states of the South and Southeast. It is also worth noting that the Midwest has its few municipalities quite associated with the high GDP cluster of the agricultural sector.

It is worth remembering that we need to go back to the medoids. I have already indicated which are these municipalities, but I have not yet said where they are. The map below corrects this gap.

Image by Author

The six municipalities are concentrated in the Southeast and South regions. There are two municipalities in Paraná, another two in Santa Catarina, one in São Paulo and another one in Minas Gerais.

In general, given the concentration of the municipalities in the South and Southeast regions, this was expected mainly for the medoids of clusters 2 to 6. Cluster 1 having its medoids in Minas Gerais is also not surprising since Santa Maria do Suaçuí is in the north of that state, therefore in an area that predominates municipalities of that cluster.

There are several analysis with maps that can be made using the medoids as reference. At this point I will highlight the per capita GDP for cluster 1. In the map below the colors in shades approaching blue refer to municipalities with higher per capita GDP than Santa Maria do Suaçuí.

The colors with shades closer to pink are municipalities with lower GDP than our reference. For the two directions stronger colors indicate greater distance from the reference.

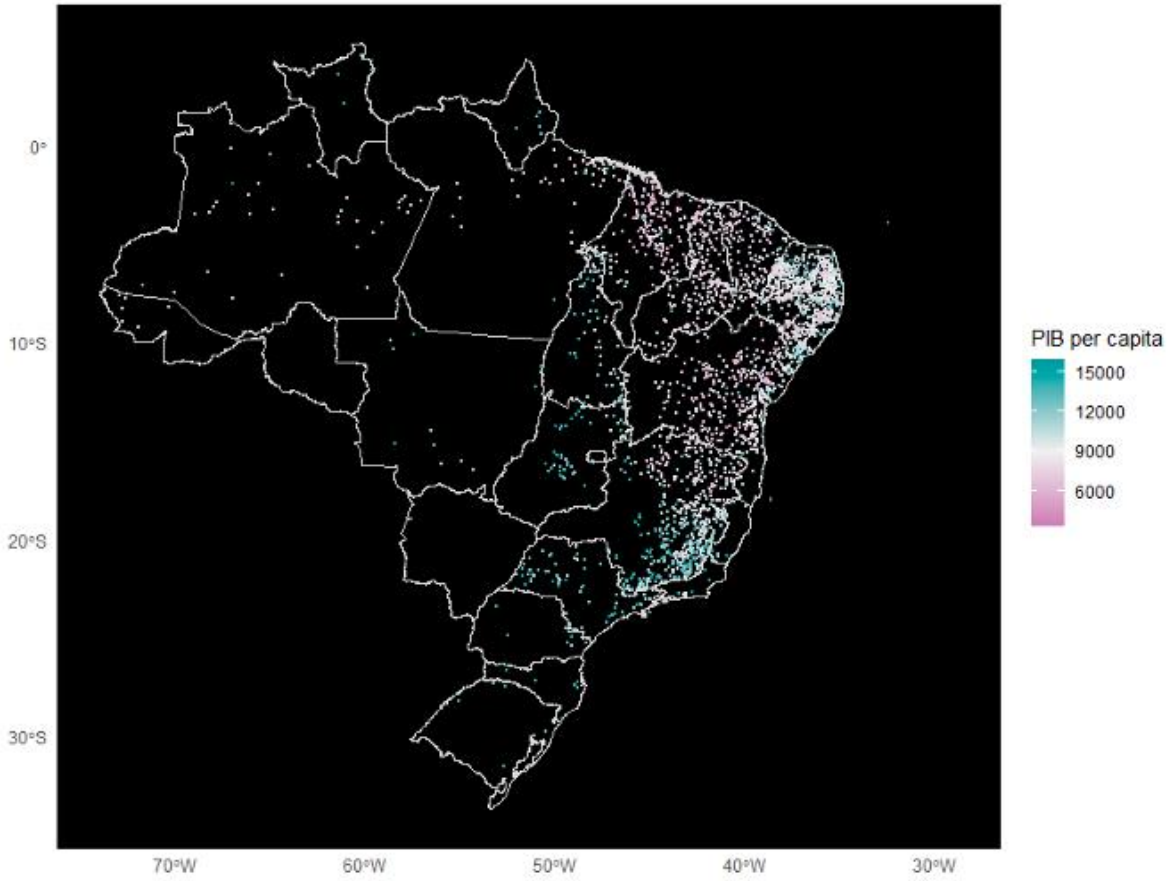


Image by Author

The cluster's medoids and the cities with GDP values close to the medoids are in the transition range between the two shades of color. It can be seen in the legend that this transition band is in the middle of the interval considered. Visually it can be observed that there is a small imbalance in the shades. There are apparently a bit more cities close to the pink tone than the blue one. Actually the distribution for cluster 1 is this:

- 1221 municipalities with a higher per capita GDP than the one of the medoide municipality.
- 1103 municipalities with lower GDP per capita than that of the medoide municipality.

It is important to note on the map above that even within the cluster the geographical pattern repeats itself. The municipalities with lower values are concentrated precisely in the Northeast and North of Minas Gerais. The points with the highest values are scattered in the other regions.

Next interstellar journeys

The findings presented seem to be important for several pieces of research that involve mainly the subjects related to inequalities. Thus, developments may go towards research related to federalism and its objectives of reducing inequalities. It may also be interesting to study public policies that impact the effects of inequality or even its causes.

It is also worth noting that the cluster approach can also be useful for research with qualitative approaches. Medoids and distances from other points in clusters to medoids can help to identify cities for more in-depth *in situ* research where the dimensions of economic output and GDP distribution are relevant to the selection of municipalities. Even the definition of cities for pilot public policy projects can use this approach.

In the end, it is hoped that the elucidation of inequalities can contribute to socio-economic differences between municipalities being increasingly less measured in astronomical units.

Codes and data

The codes are available in my Github. Take a look there. I emphasize the use of the geobr library always a great help when making maps.
The data are the same used in the other text. This is the GDP database for municipalities of 2017.

Acknowledgements

I thank Tiago Maranhão for the most accurate data visualization tips and for introducing me to XKCD. It is recommended to follow him there on twitter.
I also thank Mônica and Dante for choosing the colors.