Project 3: Portugal's 2011 census data analysis

Geospatial Data Mining

Master of Science in Geospatial Technologies
NOVA IMS
2021-22

Lucas Casuccio m20210150@novaims.unl.pt
Andre Kotze m20211199@novaims.unl.pt
Guillem Ulldemolins m20211210@novaims.unl.pt

1. EXPLORATORY DATA & SPATIAL ANALYSIS

1.1. Election of the variables

For working the exploratory data analysis and the spatial analysis we have first chosen the main variables. We decided that 12 would be a good value, into the predefined range of 10 to 15. The distinction of the most important ones was the first task, and we just discarded the ones that were the age distribution of other variables, for example, we have chosen unemployment rate instead of unemployment rate of the population between 15 and 24 years old.

After that, we added some variables either relevant for itself or relevant when comparing to other variables, such as population density (for itself) or proximity to a police station (when working with criminality rates). Our final selection is:

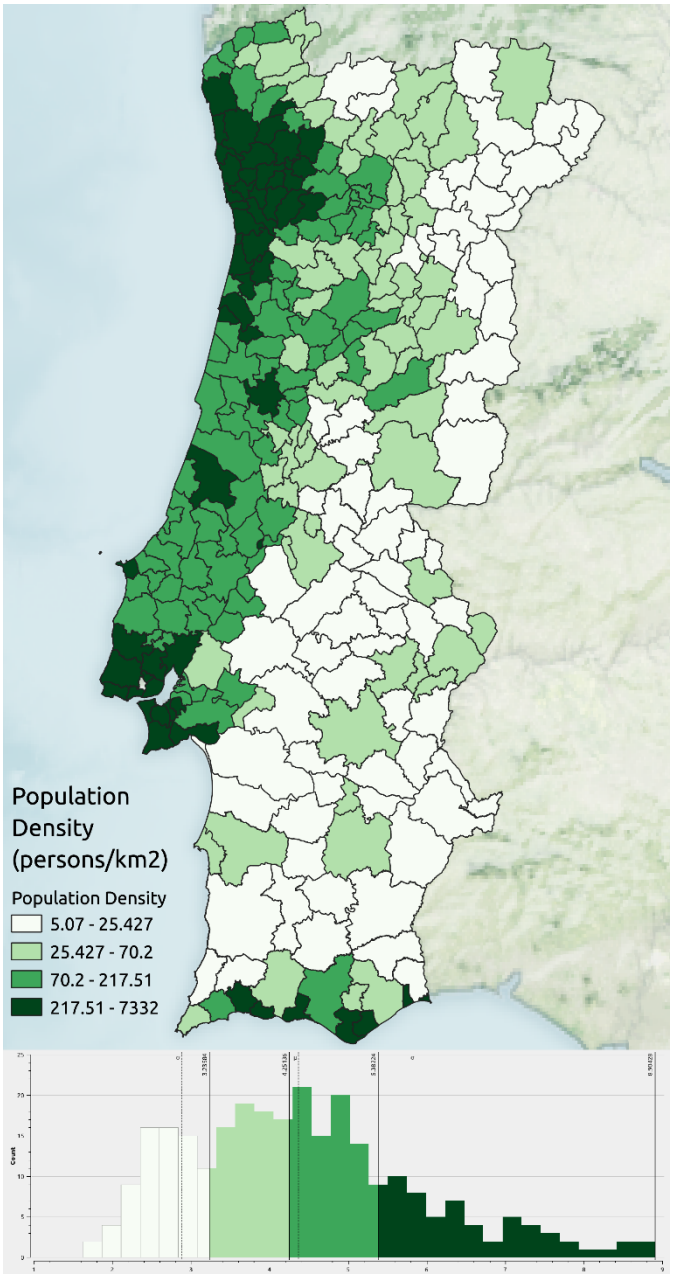
1	(denspop11)	Population density
2	(discapd)	Average distance to the closest district capital
3	(gmm2011)	Average monthly earnings
4	(podc)	Daily purchasing power per capita
5	(tdesmp11)	Unemployment rate
6	(pbsdt11)	Population receiving the unemployment benefit
7	(ne_sup)	Population with a university degree completed
8	(Abanesc)	Percentage of school dropout
9	(pdmcdt11)	Expenses of the municipality on cultural and sports activities
10	(pdiv)	Percentage of divorces
11	(distpolall)	Average distance to the closest police station
12	(tcrmiltot)	Criminality rate

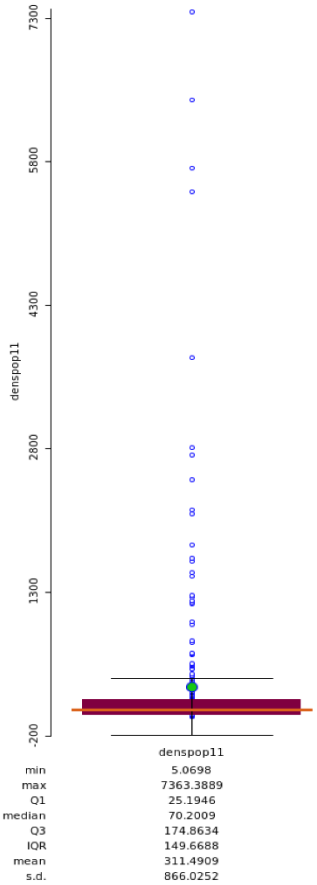
Finally, we started to explore in the data and search for spatial relations in it.

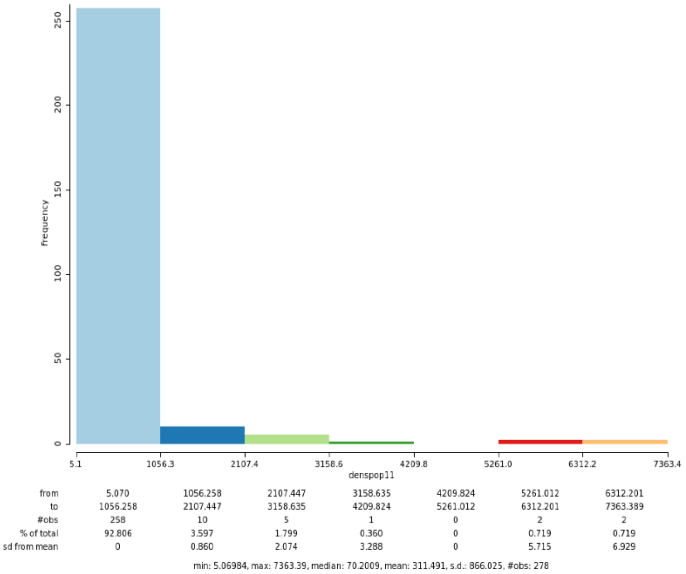
1.2. Exploratory data & spatial analysis

For this purpose, we used mainly the free open code program GeoDa. This tool allows the user to take a first contact with the data analysed. Its main features are to very easily create histograms, box plots or choropleth maps.

1.1.1 Population density (denspop11)



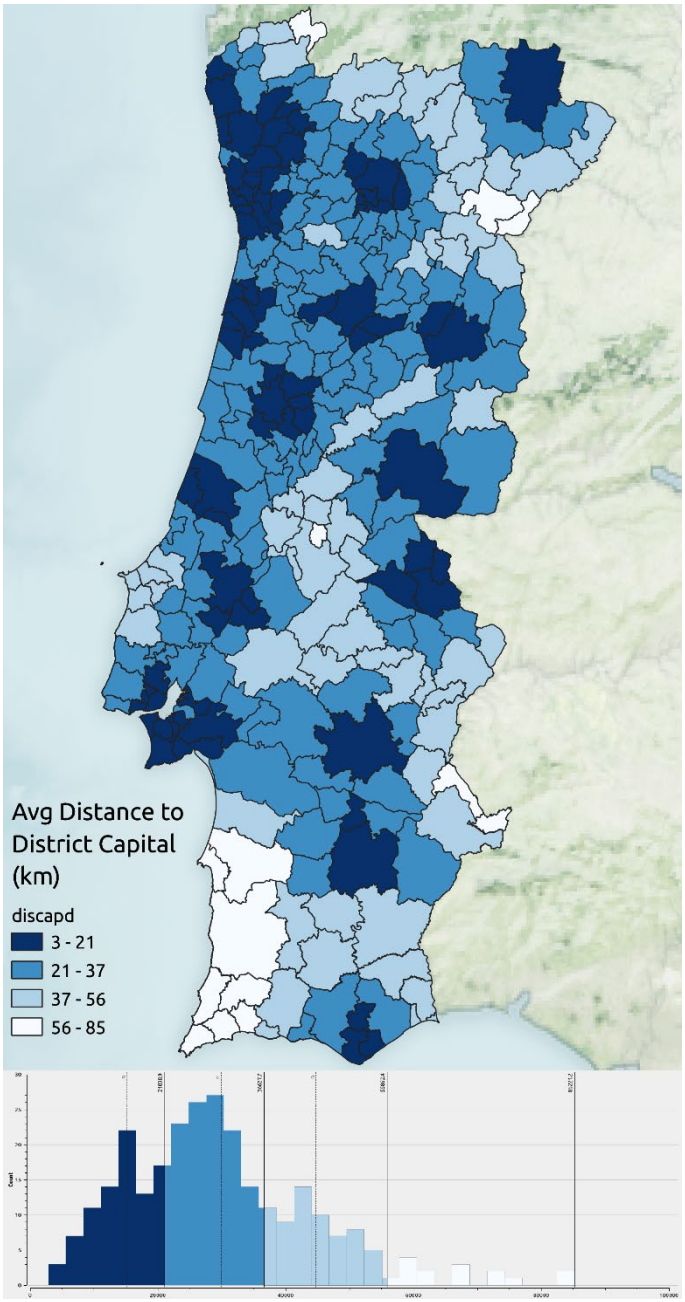




Portugal's population distribution is highly unequal, ranging from 5 to 7363 inhabitants per square kilometre. With an average of 311, but a median of 70, it is clear that the density increases exponentially in highly urbanised areas. A histogram (above) shows how the majority of the country is sparsely populated, with only a few municipalities accounting for the bulk of the total and as a result, it has an exponential structure. Thus, another histogram normalised to the natural log has been included with the choropleth map (below).

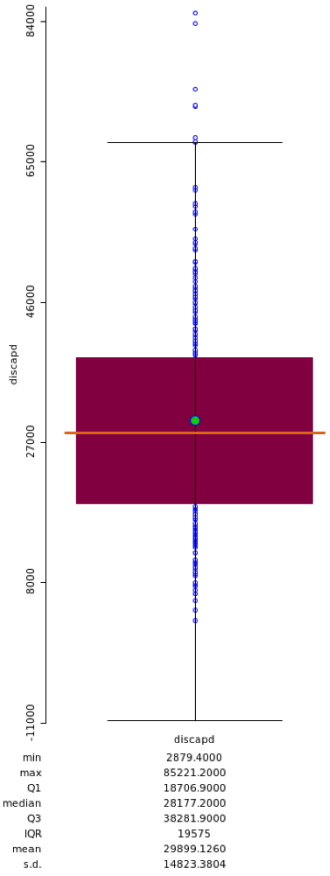
We can appreciate two main clusters around the most important cities, Lisboa and Porto. There are some outliers such as Entroncamento, a municipality located in an area with moderate values of density but with a very small area (13,73km^2) that converts him into an outlier.

1.1.2 Average distance to the closest district capital (discapd)



The values of distance to the closest district capital are expressed in meters. They range from 3km to 85km (average 30km) with only 8 municipalities exceeding 60km. This distribution is positively skewed.

All around the Portuguese geography there are 18 hubs of low values, which correspond to the 18 district capitals. In the South-West, around Ponta de Sagres and in the Nort-East, in the oriental Bragança is where the highest values are located.

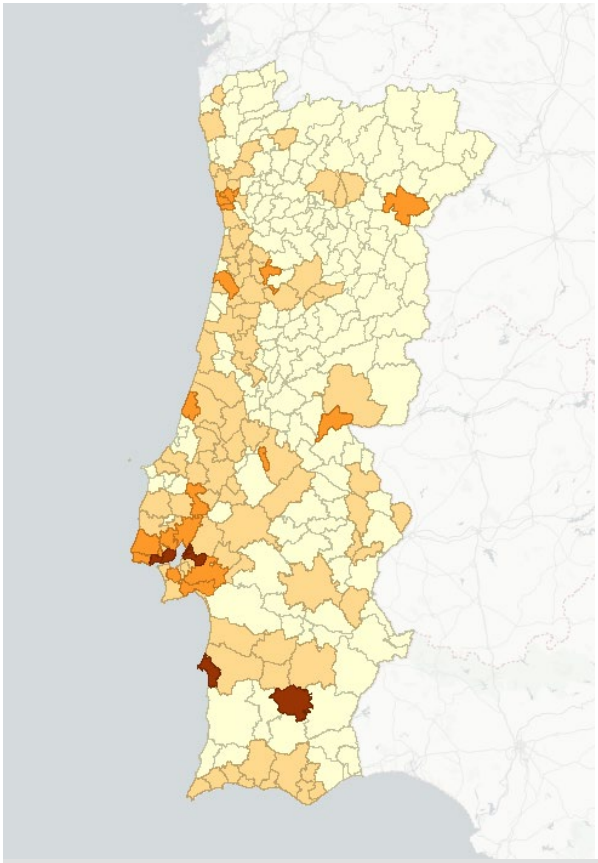


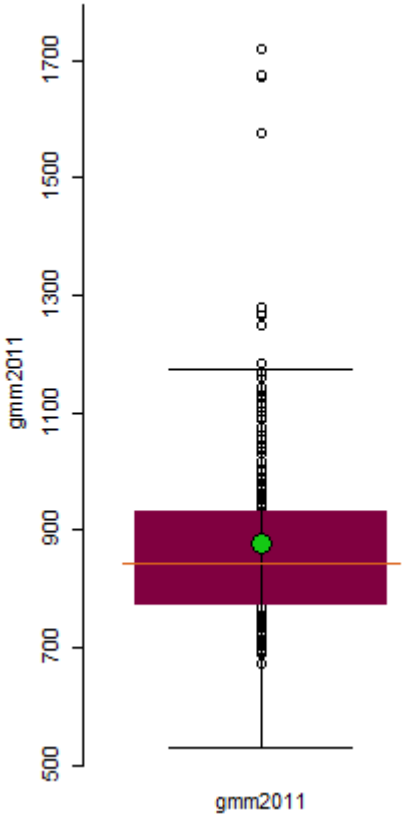
1.1.3 Average monthly earnings (gmm2011)

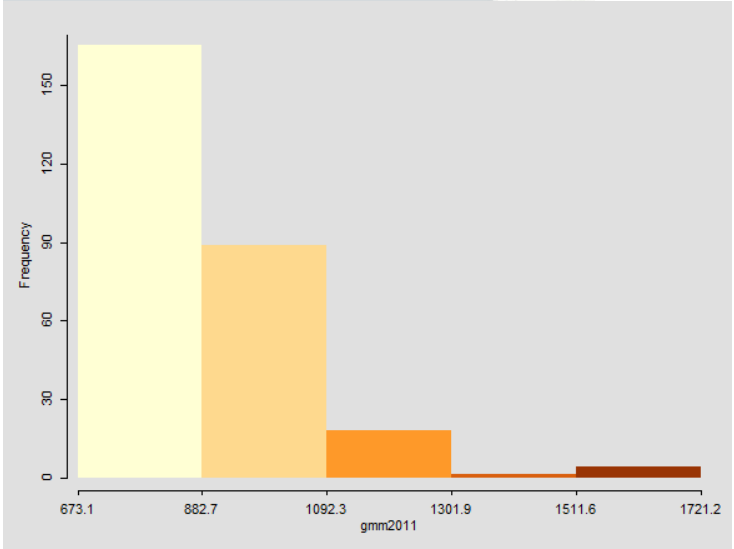
The average monthly earnings are stored in the attribute GMM2011 and are expressed in euros. The mean for this attribute is 876,91€, while the standard deviation has a value of 159,76€. Its values vary from 673,07€ in Freixo de Espada à Cinta (a frontier town in Bragança's district) to the 1721,20€ in Oeiras, attached to Lisbon.

At first handside, the spatial distribution of the average monthly earnings show us higher values on the coastal municipalities than in the interior ones. The biggest high values agglomeration is located around Lisboa, while the main concentration of low values is located around Serra da Estrela and extends up to the North until the Duero River.

We also can appreciate a Low-High cluster located in Bragança, where the municipality of Freixo de Espada à Cinta, the one with the lowest values is attached to Torre de Moncorvo, from the highest decile.



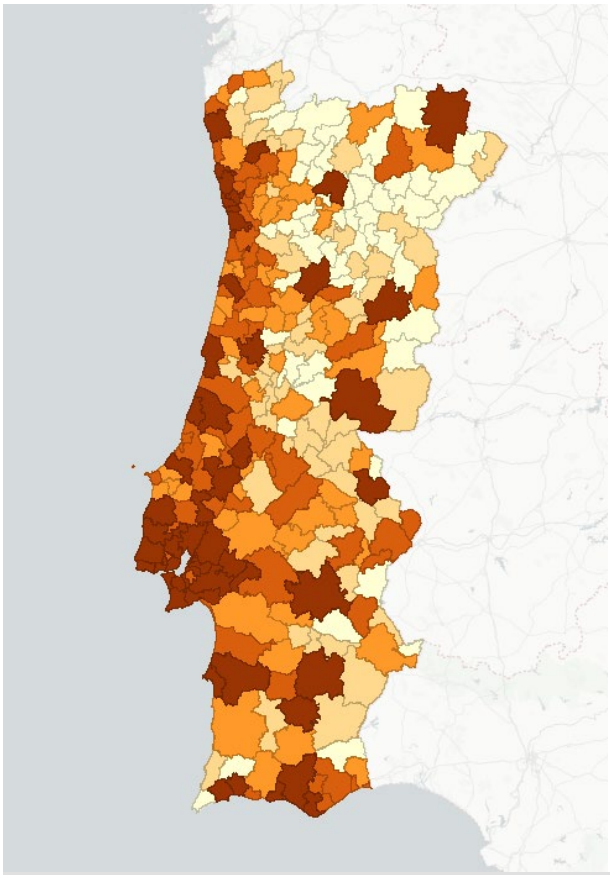


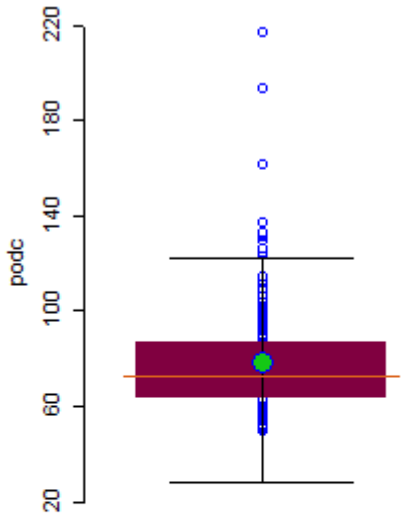


1.1.4 Daily purchasing power per capita (podc)

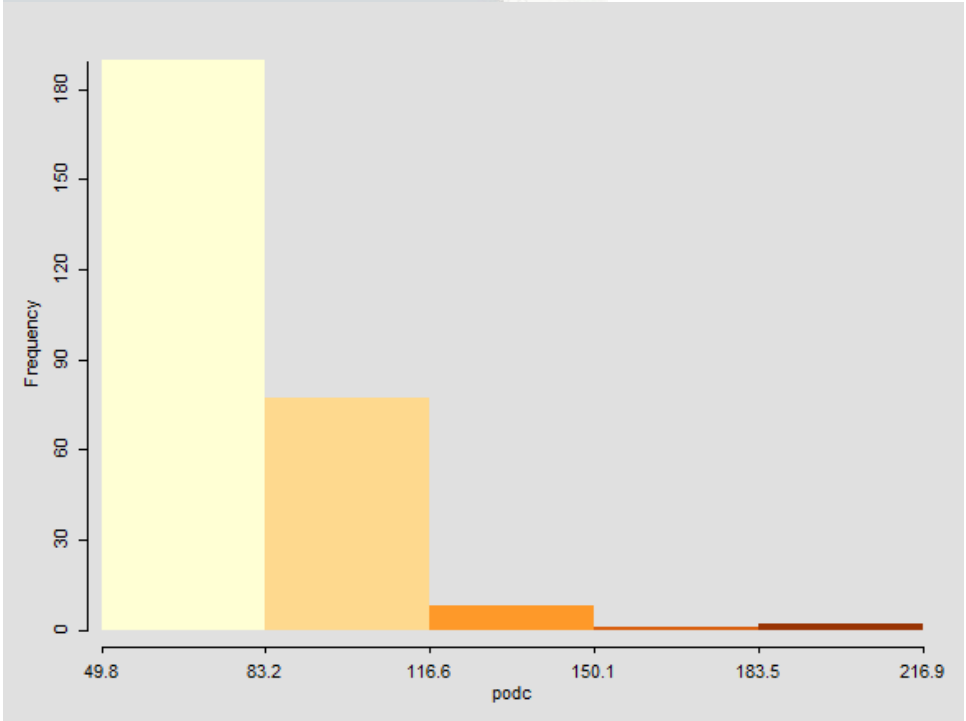
In PODC we have the values for the daily purchasing power per capita. This is expressed in a comparative index, where the country mean is the index 100. The mean for the municipalities values is 78, as the median is 73. The standard deviation has a value of 21.20.

The highest values are in the municipalities located at the coastal regions, being Lisboa the top one. The lowest values are mostly in the interior of the Norte and Centro regions.





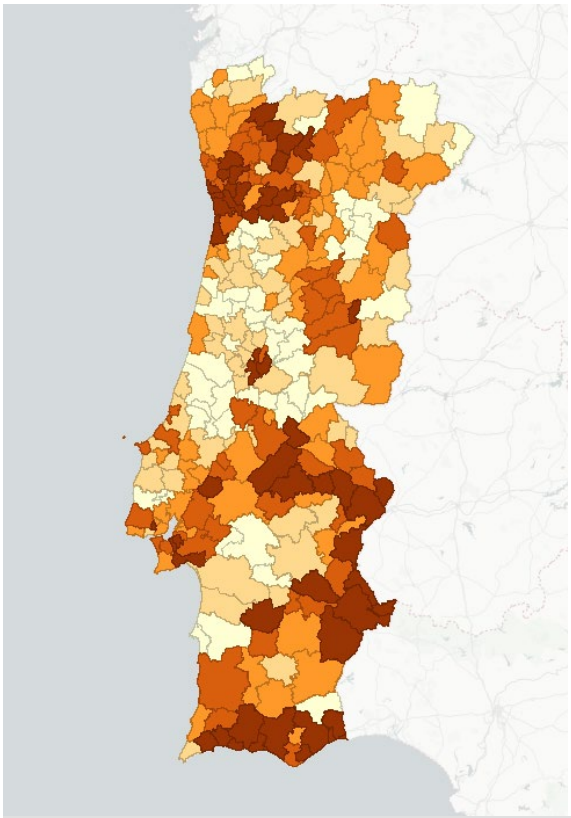
	podc
min	49.8000
max	216.9000
Q1	63.4000
median	72.2500
Q3	87
IQR	23.6000
mean	78.0234
s.d.	21.2047

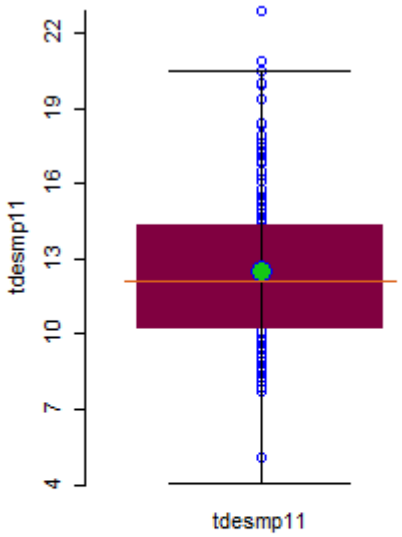


1.1.5 Unemployment rate (tdesmp11)

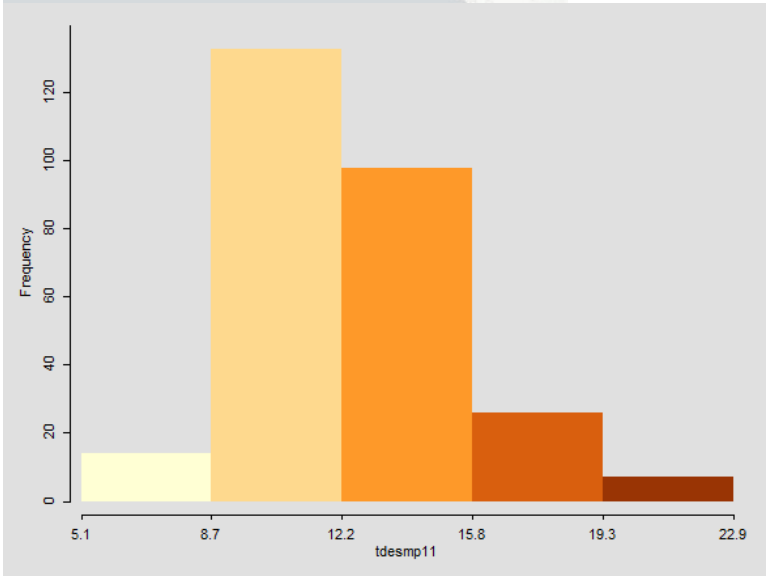
Tdesmp11 is the unemployment rate, it is presented as a percentage and is calculated dividing the number of unemployed people between the potential working population. Its values vary between 5.1% (Oleiros, central Portugal) and 22.9% (Mourão, Alentejo). The distribution of this data tends to a normalization, with the histogram deviated to the lower values. Its standard deviation is 2.8 and the mean is just a bit higher than the median (12.1%, 12.5%).

On the side of the spatial distribution, we can appreciate two major clusters of unemployment, located in the coastal and central Norte region and in the oriental Alentejo. Apart from that, the distribution does not seem to follow any pattern. With some knowledge of the country, we would be able to go further in the qualitative analysis, for example, relating the temporality of the occupation in the Algarve region with the high unemployment rates that are shown in the region.





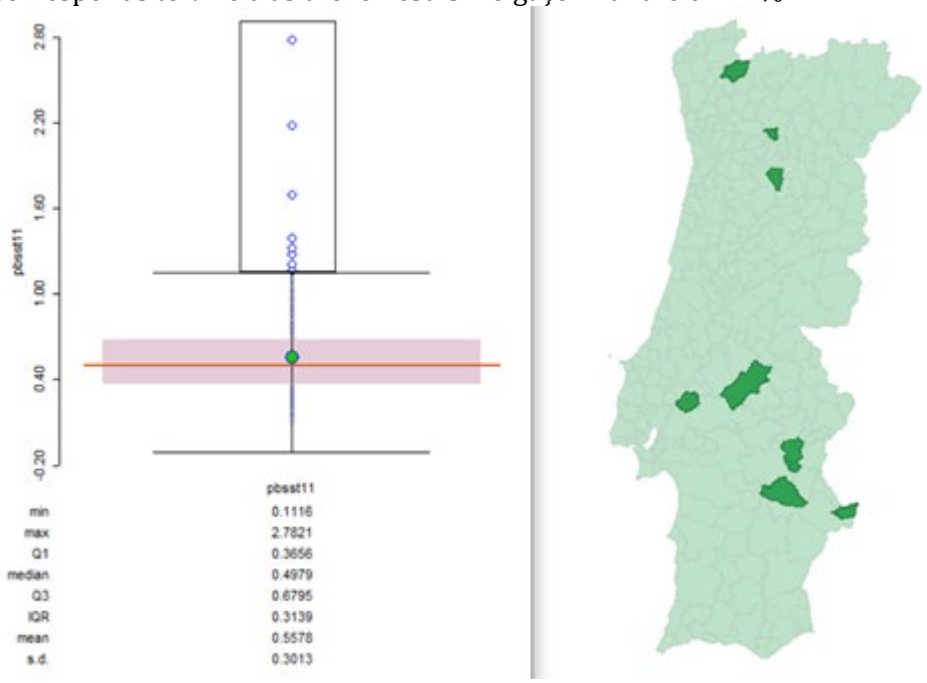
tdesmp11	
min	5.1000
max	22.9000
Q1	10.2000
median	12.1000
Q3	14.3000
IQR	4.1000
mean	12.4978
s.d.	2.8345



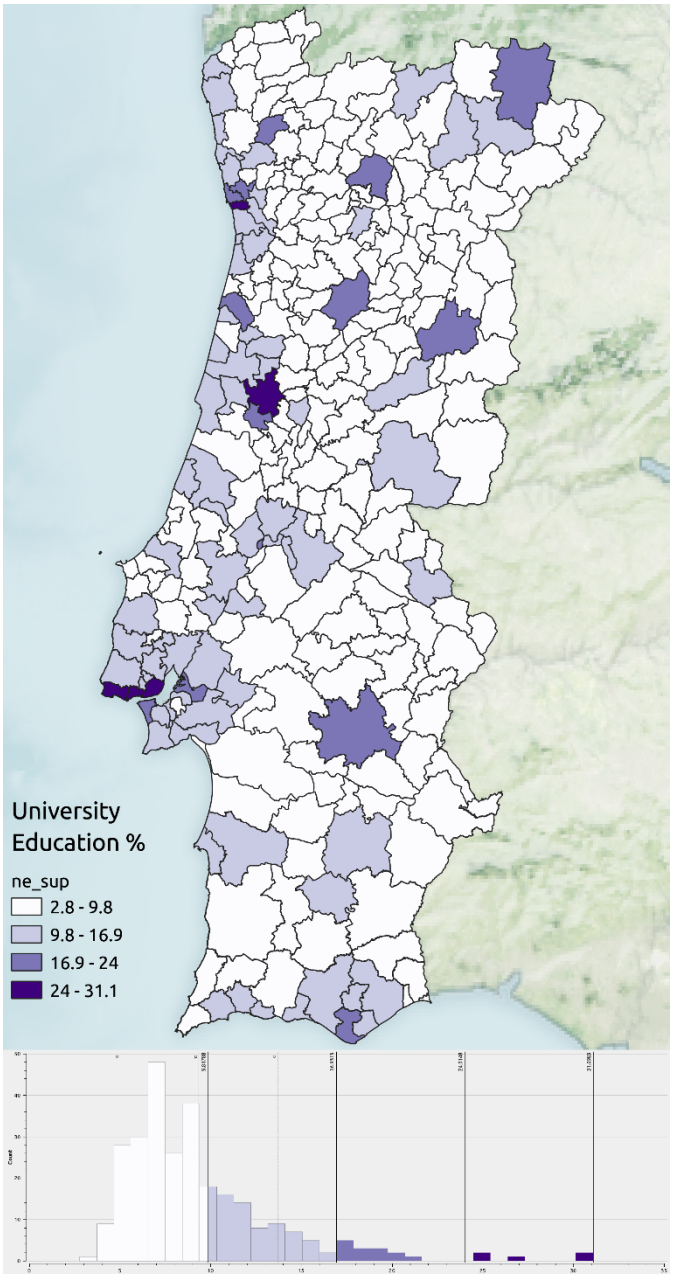
1.1.6 Percentage of the total population receiving the unemployment benefit (pbsdt11)

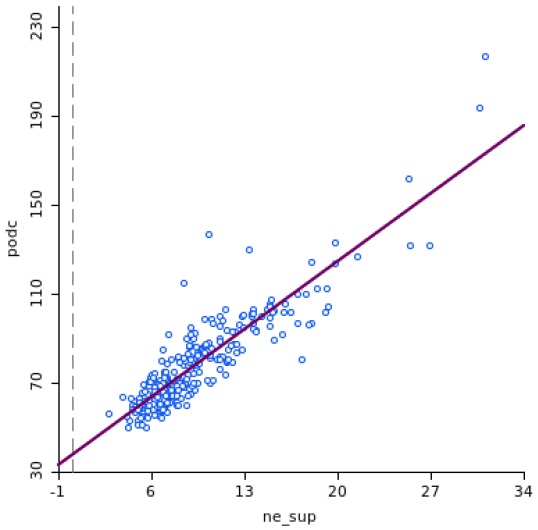
The variable of percentage of the total population receiving social employment benefit shows that most values are concentrated in the Media/Q3 region, the median (0.498) is near the media (0.558), which is higher, and this fact is caused by the outliers that are exposed above the Maximum limit. The standard deviation is 0.3,

The major concentrations of this value are in the coastal municipalities, clustering around Porto, the Tajo River and in Algarve. In opposition, the lowest values are in the interior and in the North. The highest value (2.782%) corresponds to Vizela as the lowest is Melgaço with the 0.112%.



1.1.7 Percentage of population with a university degree completed as education level (ne_sup)





#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
278	0.798	37.298	1.359	27.440	0	4.360	0.132	33.035	0
278	0.798	37.298	1.359	27.440	0	4.360	0.132	33.035	0
0	0	0	0	0	0	0	0	0	0

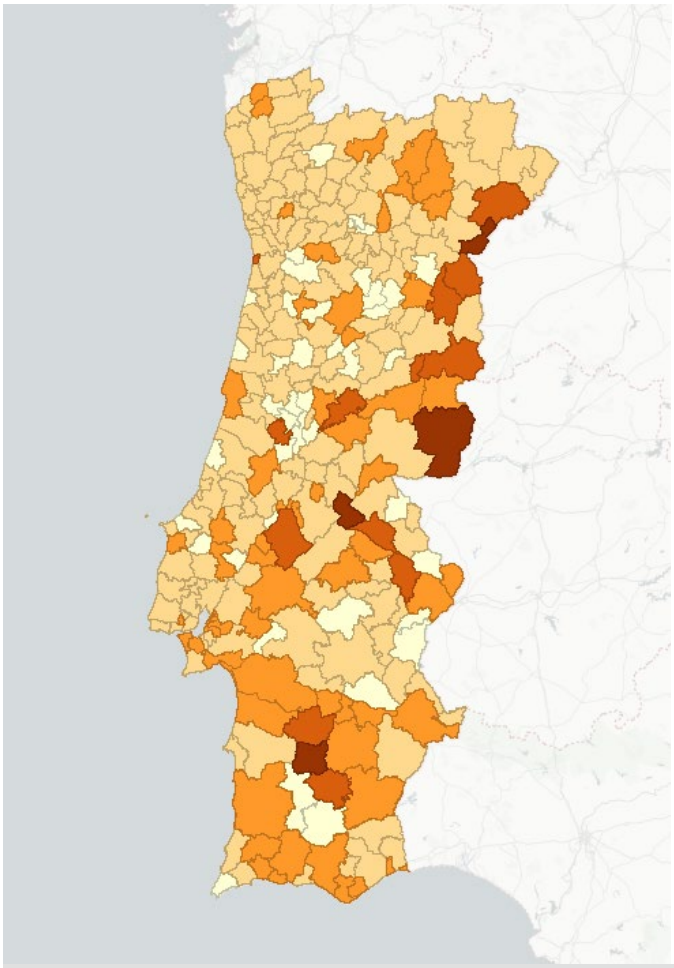
University education is positively skewed, ranging from 3 - 31% but with 60% in the interval 5 - 11%. Lisbon, Oieras and Coimbra lie 3 – 4σ from the mean, and with Porto and Cascais make up the top 5.

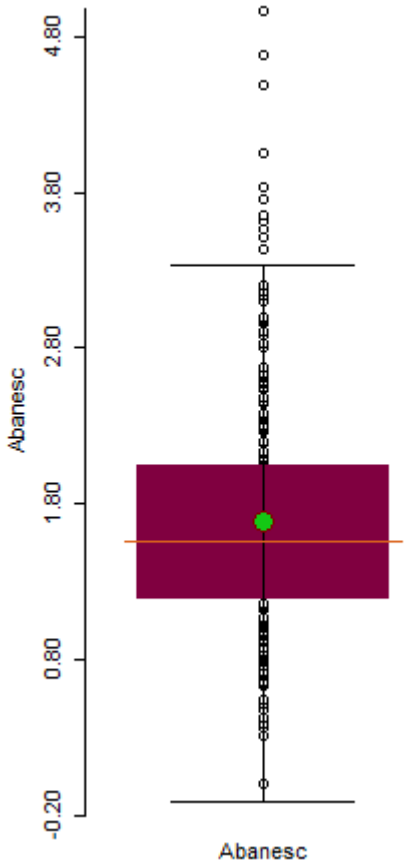
The spatial distribution for this value does not create great clusters or tendences in comparison with some others. We can appreciate some clustering around Lisboa, Porto and Coimbra, the cities with the most important universities. The tendence seems to be, as with so many other variables, to have higher values in the coastal and nearby municipalities.

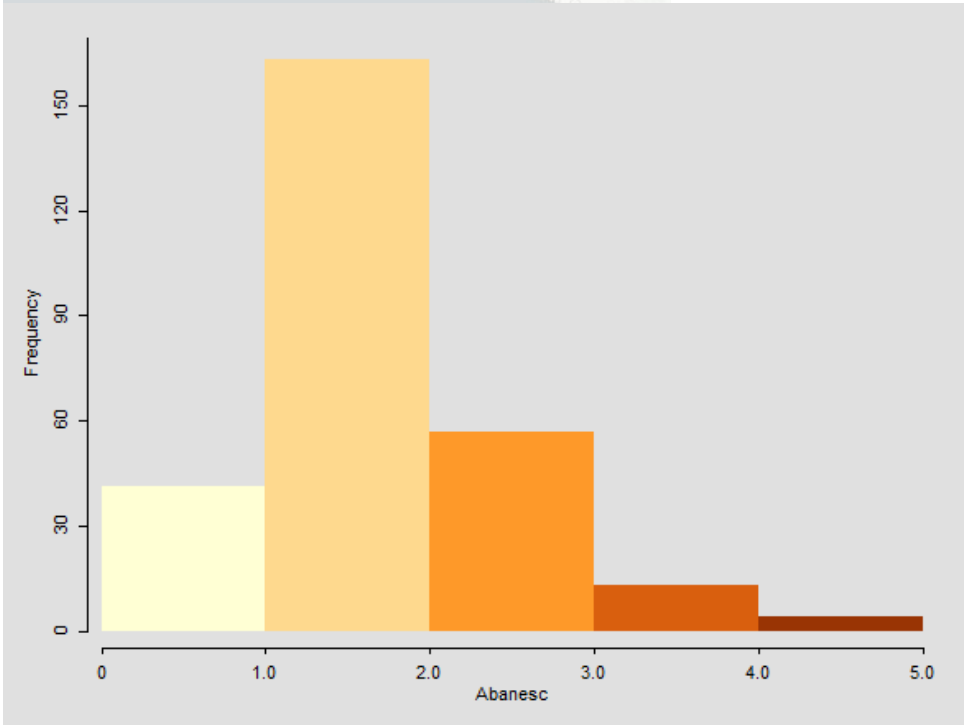
When comparing the percentage of population with a university degree completed to other categories, we can appreciate a strong correlation of this value with purchasing power index (podc, $R^2 = 0.80$).

1.1.8 Percentage of school dropout (Abanesc)

School abandonment is positively skewed as the histogram weights more in his left. The lowest value is 0% and it never exceeds 5%. The mean (1.681%) is slightly higher than the median (1.550%). The higher rates are confined to more rural municipalities in the north-east and south-central region as in most of the country there are values in between 0.9 and 2.5%.



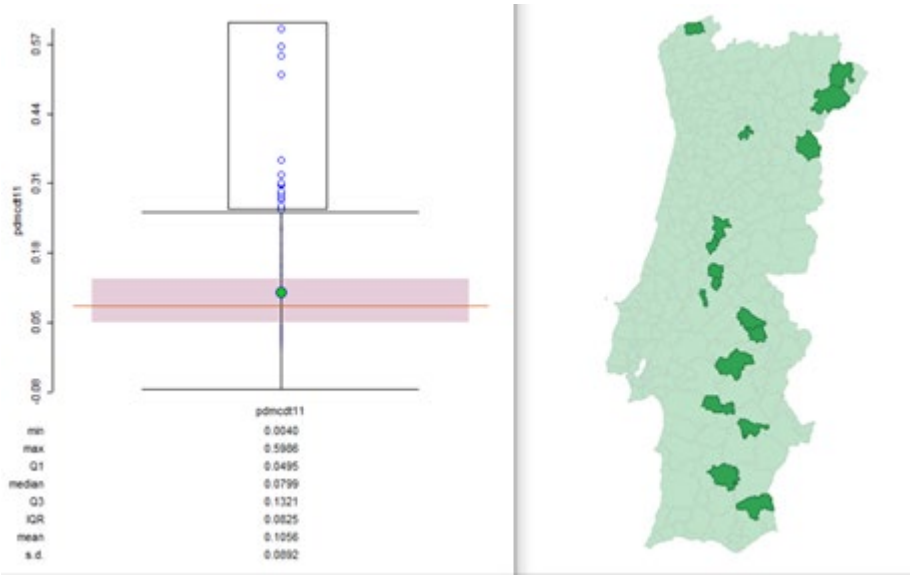




1.1.9 Expenses of the municipality on cultural and sports activities (pdmcdt11)

The variable regarding expenses of the municipality on cultural and sports activities, per inhabitant (1000 EUR) shows that most values are concentrated in the Media/Q3 region, the median is near the media, which is higher, and this fact is caused by the outliers that are exposed above the Maximum limit. The standard deviation is 0.08.

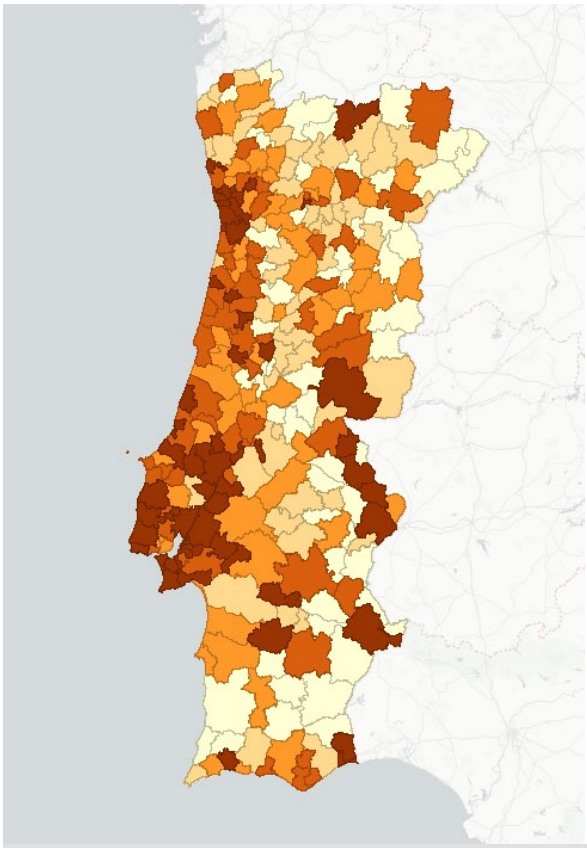
There is a strong gradient of *coastality* in the spatial distribution of the values, as the higher values (over 0.150) are mostly located in the interior municipalities.

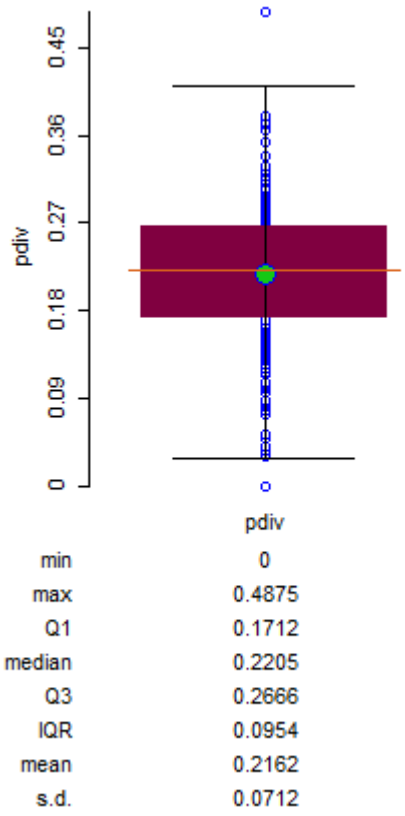


1.1.10 Percentage of divorces (pdiv)

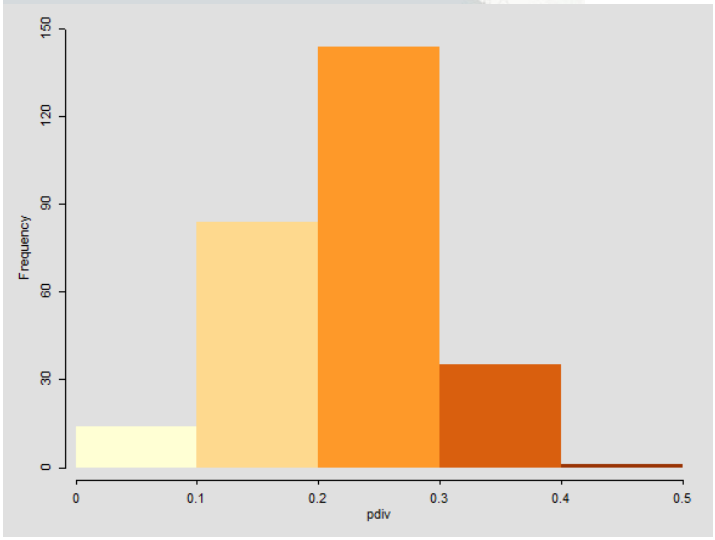
The percentage of divorces is calculated over the married couples, that is why it has so small values. Its smallest value is 0, meaning no divorces in that municipality during 2011, as the highest value is 0.49% (Viana do Alentejo). This data tends to follow a normal distribution. Its mean equals to the median (0.22%) and its standard deviation is 0.07.

The geographical distribution of this data shows us that there tend to be more percentual divorces in municipalities closer to the Atlantic coast, with the major concentration located all around Lisbon. Otherwise, the lower values are more easily located in the interior departments but also in the southern Alentejo.





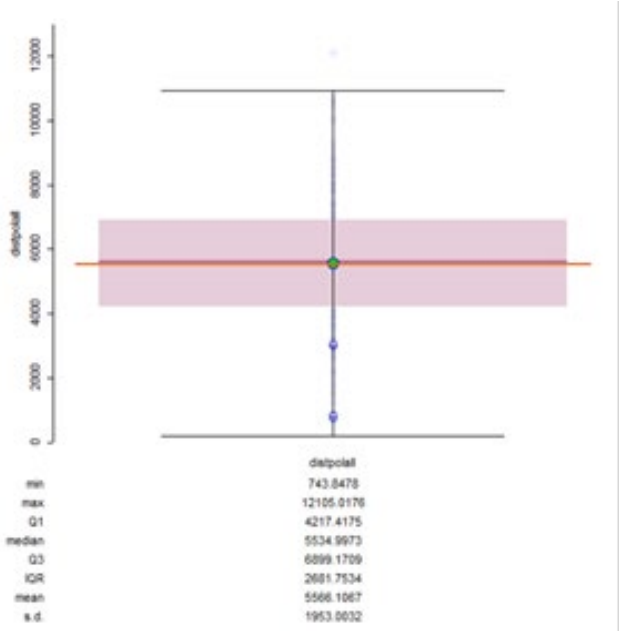
min	0
max	0.4875
Q1	0.1712
median	0.2205
Q3	0.2666
IQR	0.0954
mean	0.2162
s.d.	0.0712

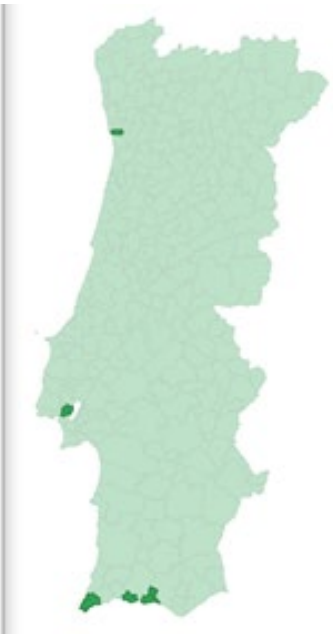


1.1.11 Average distance to the closest police station (distpolall)

The variable of average distance to the closest police station (m) shows low variability between most of the values, the median remarkably similar to the media, caused by the almost null presence outliers, only one value is exposed above the Maximum limit and corresponds to the DICO Chamusca. The standard deviation is 1953 m.

When observed the DICO's distance from police stations with more criminality rates, it is exposed that these outliers on high criminality rates are not related with a lack of density regarding police stations.

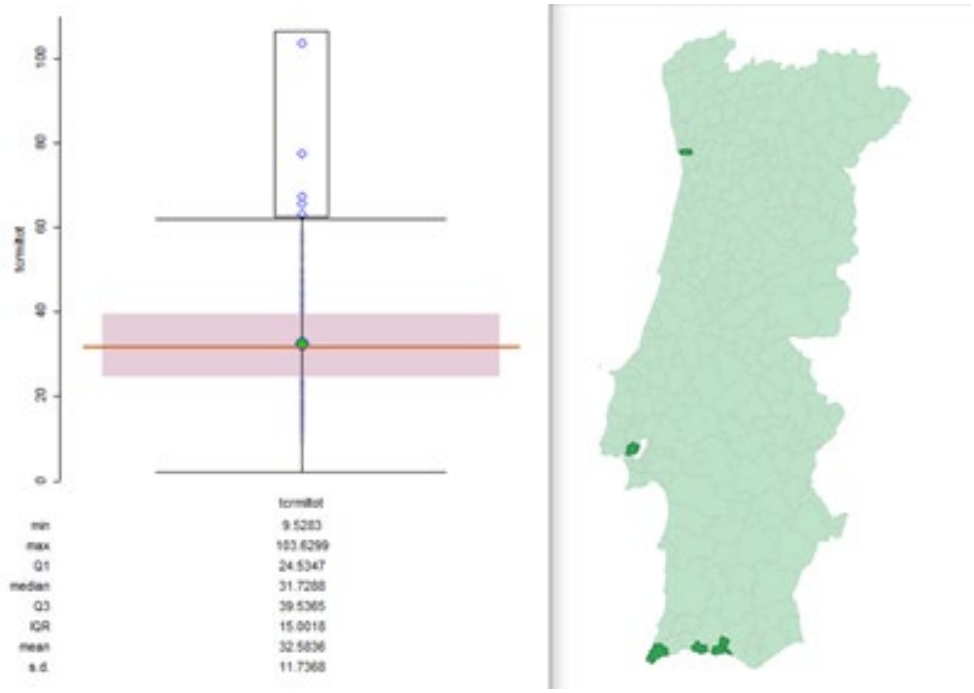




1.1.12 Criminality rate (tcrmiltot)

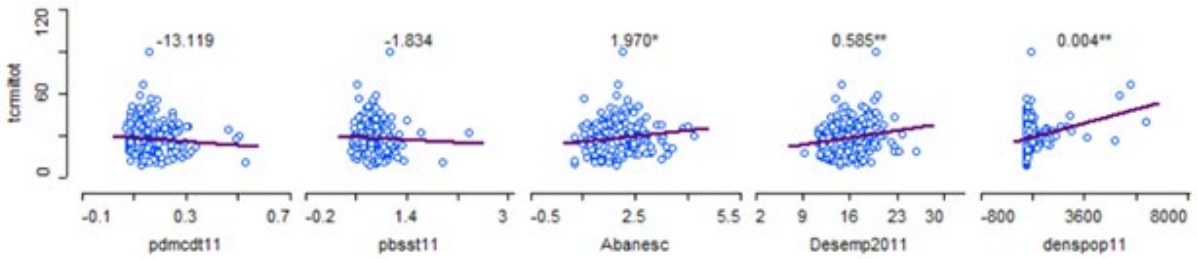
The variable of total criminality rate shows low variability between most of the values, the median is near the media, which is higher, and this fact is caused by the outliers that are exposed above the Maximum limit and are even disperse between themselves. The standard deviation is 11.7.

Albufeira (epicentre of Algarve) is the DICO with the highest criminal rate. Followed by Lisbon, Porto, Lagoa and Vila do Bispo. All cities with high affluence of tourists.

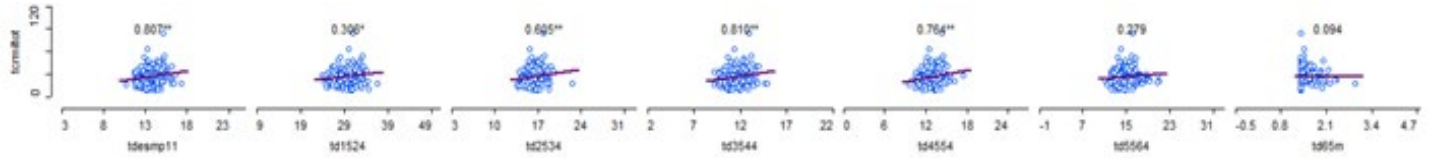


When considering the correlation between this category and some others, first we can notice that the total criminality is directly proportional to the population density.

Also shows negative correlation with the absence of scholar education, as it shows positive correlations with the pct. Of people who have finished secondary studies, median studies, and superior studies; from that, we can affirm that as major is the mean study level in a municipality, les crimes are committed in it.



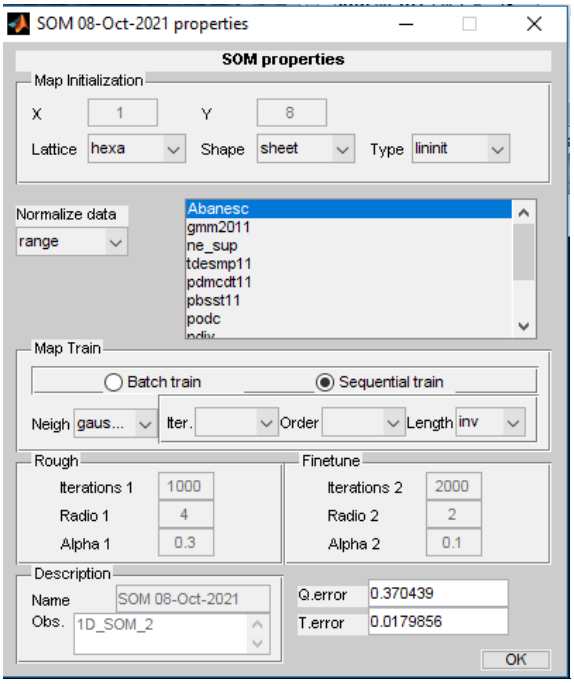
When comparing it to the unemployment rates, it shows positive correlation with high young unemployment, but keeps immutable with people over 55 being unemployed.w

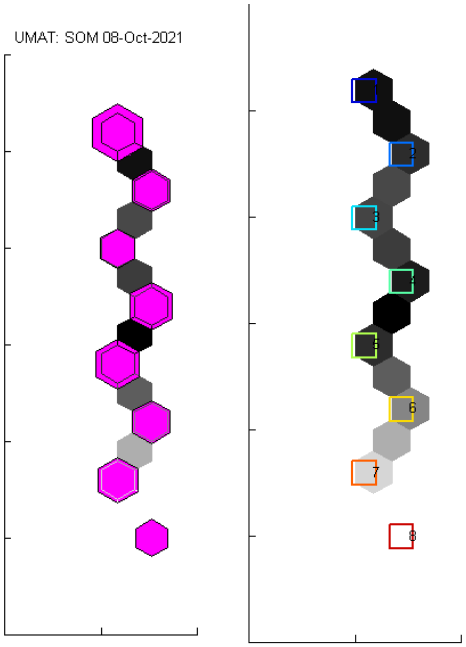


2. MULTIDIMENSIONAL ANALYSIS

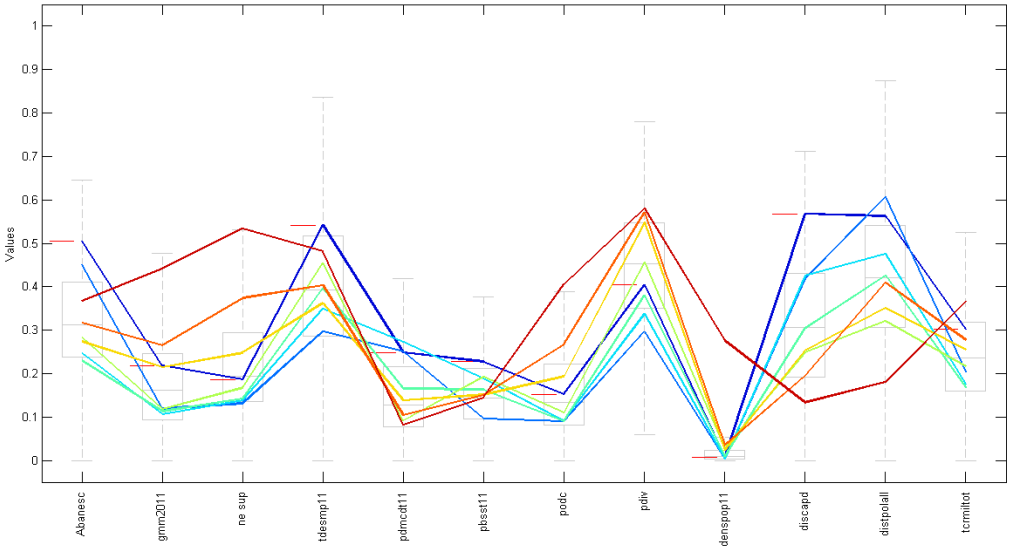
2.1. SOM clustering

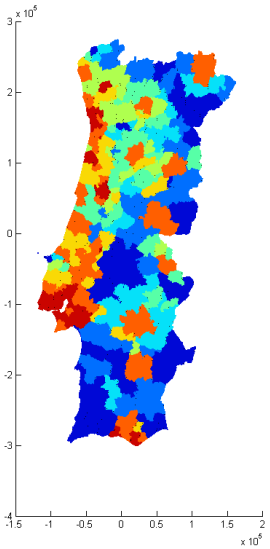
1x8 SOM



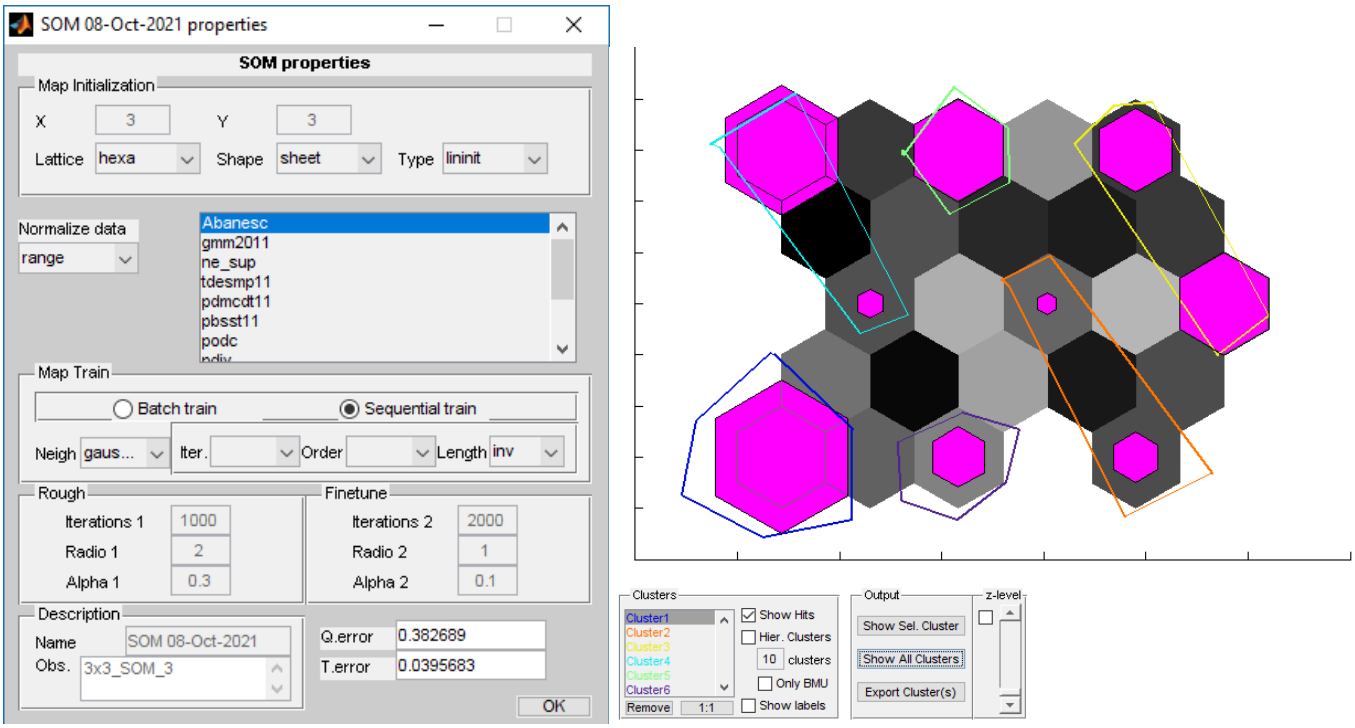


A one-dimensional neural network (1 x 8) was used, in order to maximise variance between clusters. The UMAT shows that hits are comparatively well-distributed, and differences between nodes increase from one end of the neural string to the other. Looking at the component histogram, clusters seem to be defined on the basis of Average Earnings (gmm2011), University Education (ne_sup), School Dropout (Abanesc), Divorce Rate (pdiv) and Distance to Police Station/District Capital (discapd, distpolall). The smallest variance can be seen in Population Density (denspop11) with only one cluster deviating significantly from the mean.

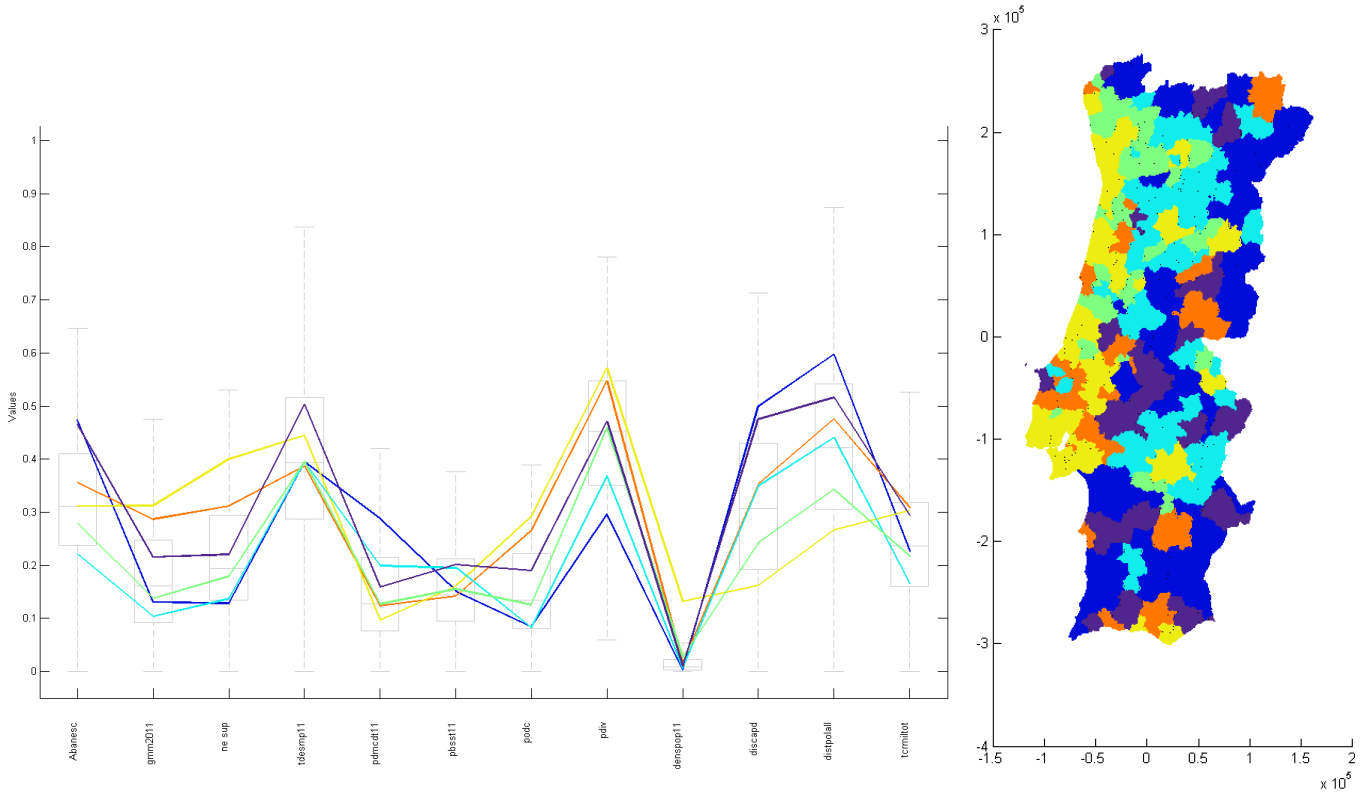




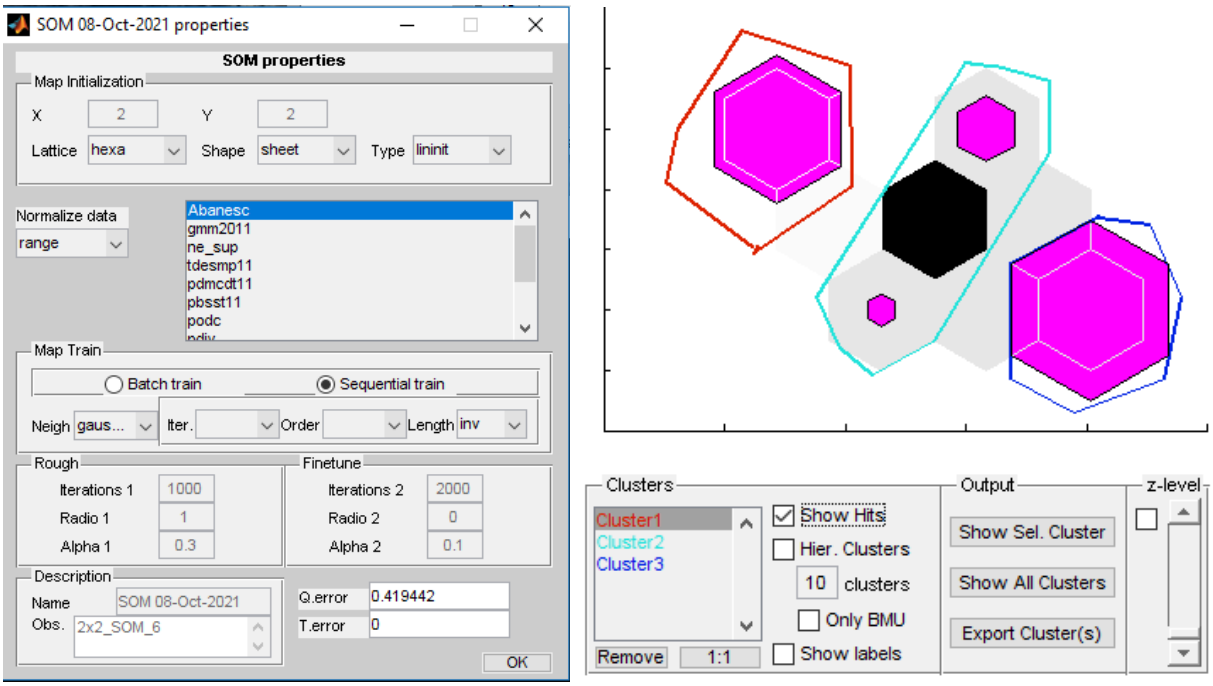
3x3 SOM



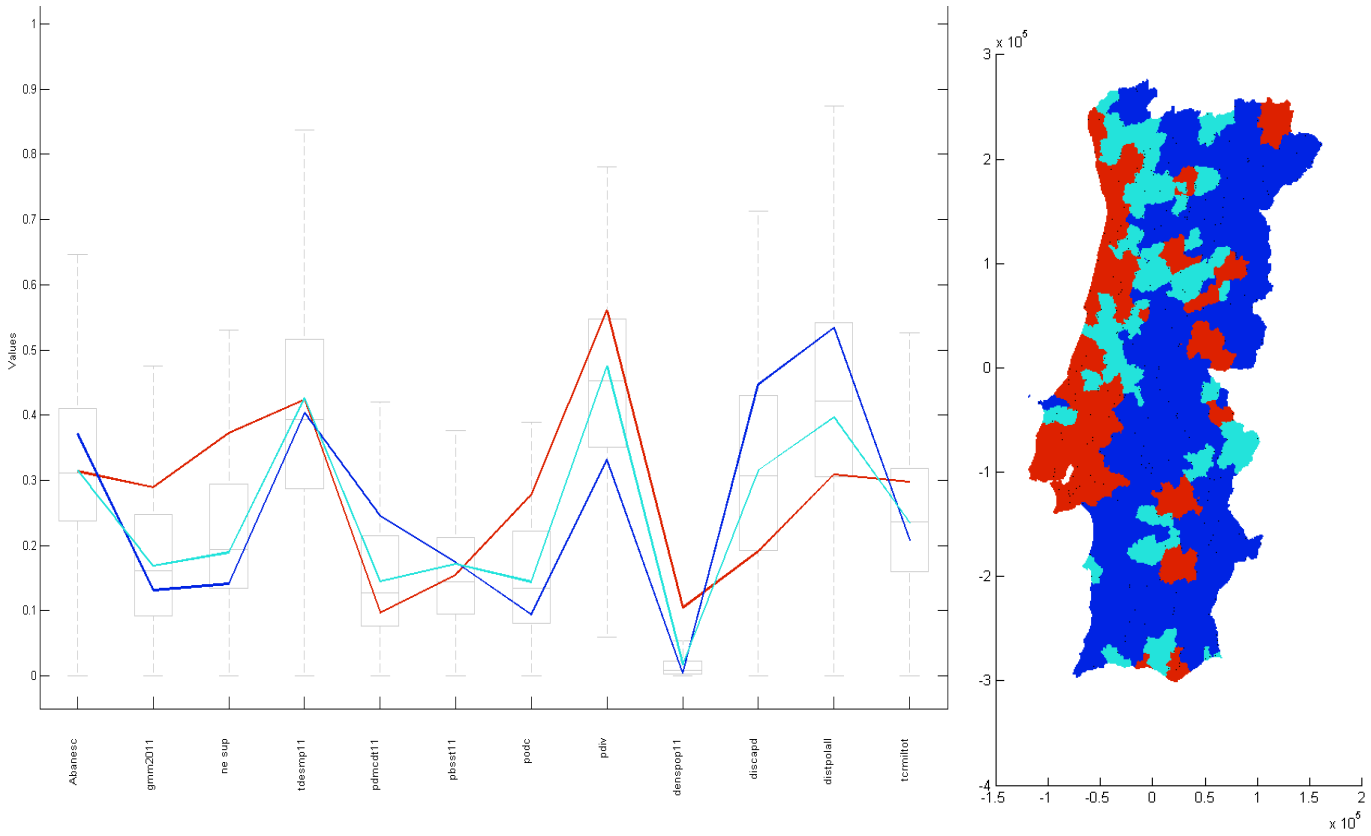
A small (3 x 3) map showed highly unequal hit distribution in the U-matrix, and six broad clusters could be discerned. They separated well based on discapd and distpolall, as well as gmm2011 and ne_sup. On the map, the most highly urbanised municipalities can be clearly distinguished, and more rural municipalities differ most widely in tcrmtot, podc. Because of the combined weight of discapd and distpolall, clusters reflect municipality isolation quite well.



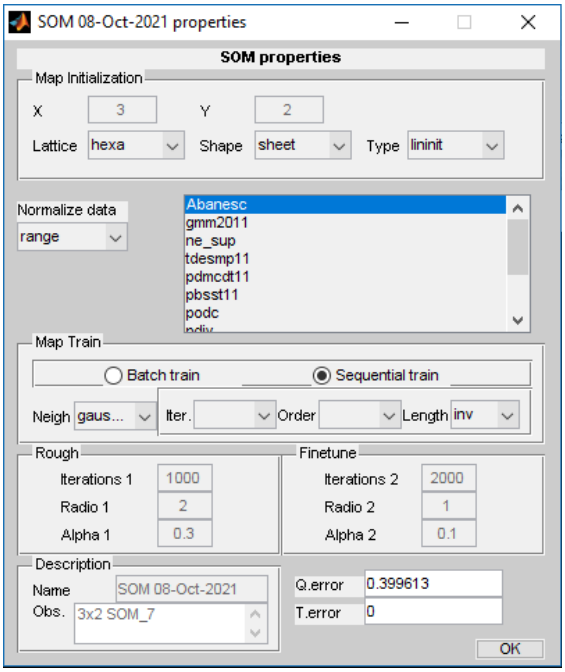
2x2 SOM

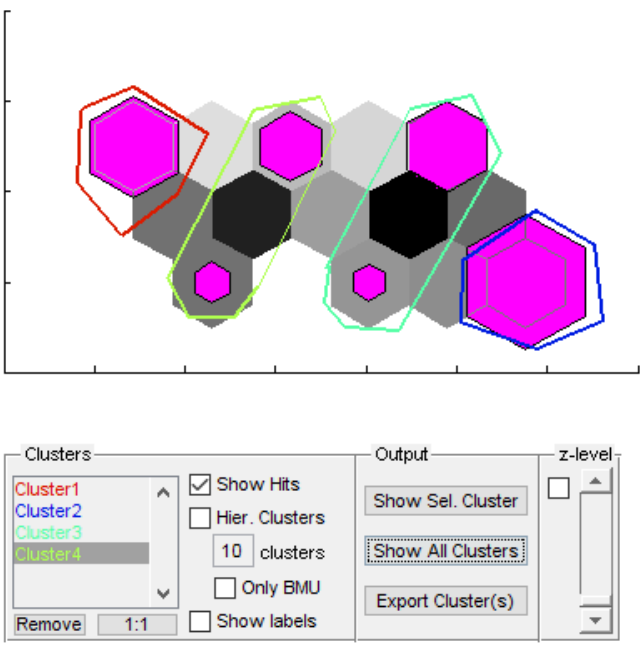


This map is an extreme case, using a 2 x 2 network only. Two neurons with the fewest hits could be joined into one cluster. Using this high level of discrimination, the three generalised clusters showed the greatest variance in gmm2011, ne_sup, pdiv, discapd and distpolall. Very little variance was evident in tdesmp11, pbsst11 and tcrmtot.

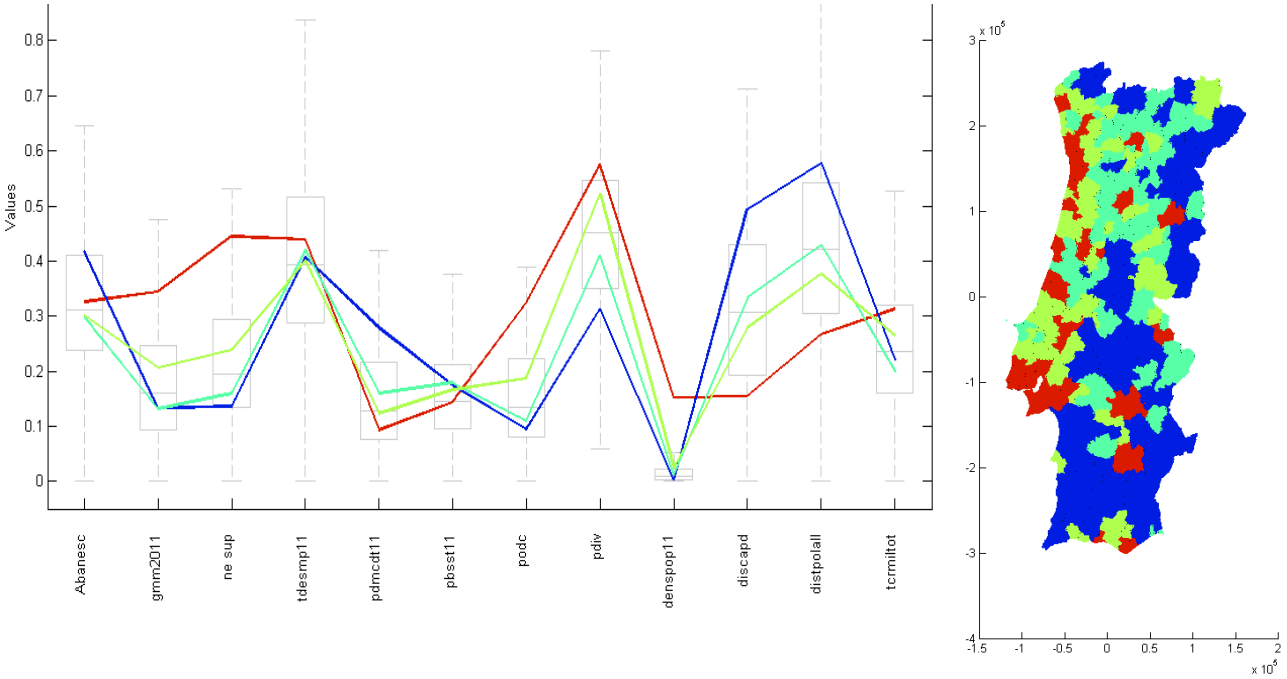


3x2 SOM



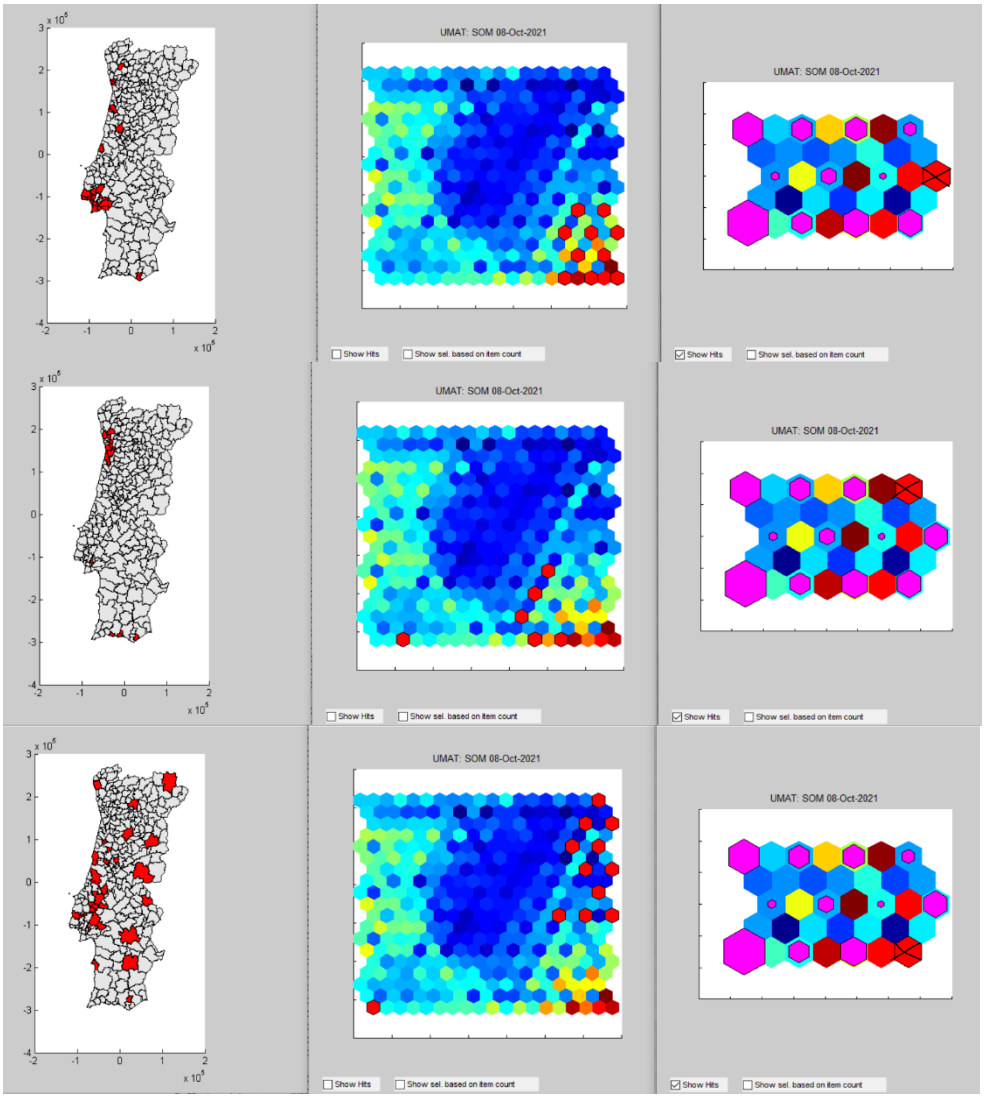


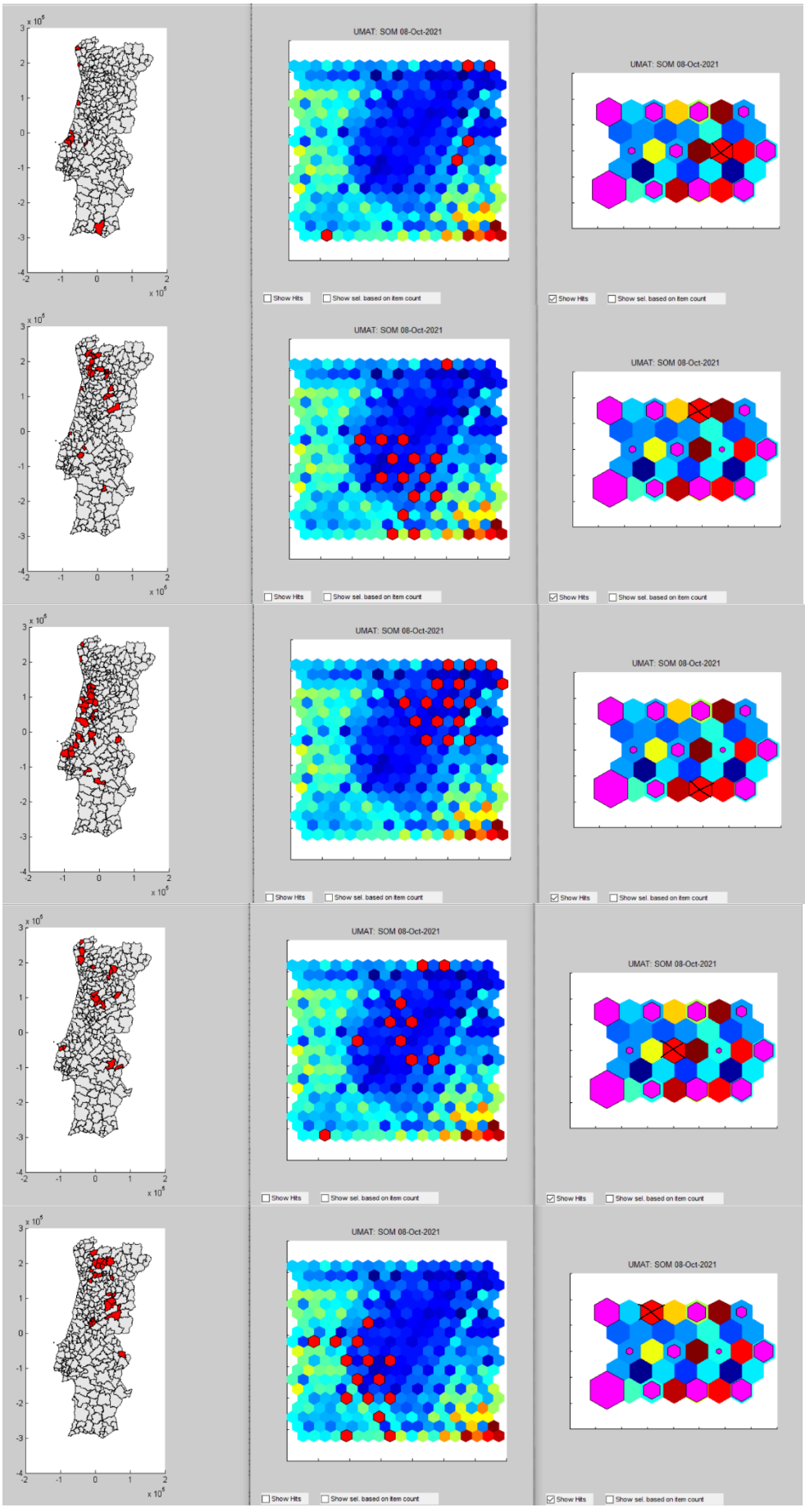
This map (3 x 2) had the highest quantification error, though the number of clusters seemed close to optimal. The 6 neurons could be clustered 1:1, or grouped into 4 (as shown). As with previous clustering results, the overarching differences are traceable to urban development i.e. a graded urban - rural split, evidenced by higher earnings, education and divorce in highly developed municipalities, and higher remoteness (discapd and distpolall) and school dropout in the countryside.

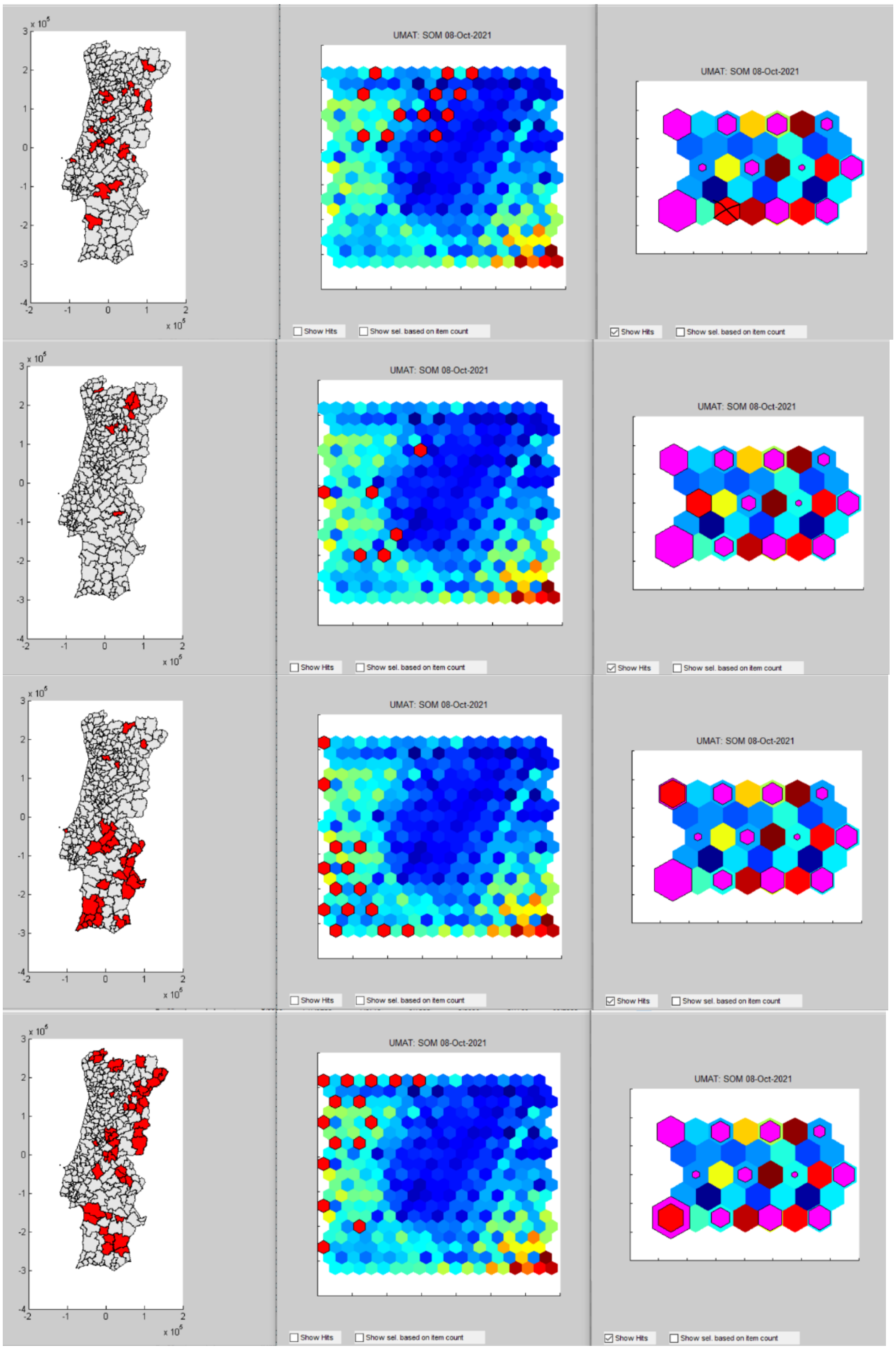


4 x 3 vs 10 x 10 SOM

The following figures show a comparison between 2 SOM performed over the same variables and the way they are distributed on the map regarding the SOM with 12 neurons and also how each neuron groups the neurons of the bigger SOM. The 100 neurons SOM is the one with less error in Q (0.263) and T(0.01) compared to the SOM made with 12 neurons (QError: 0.366, TError: 0.02)

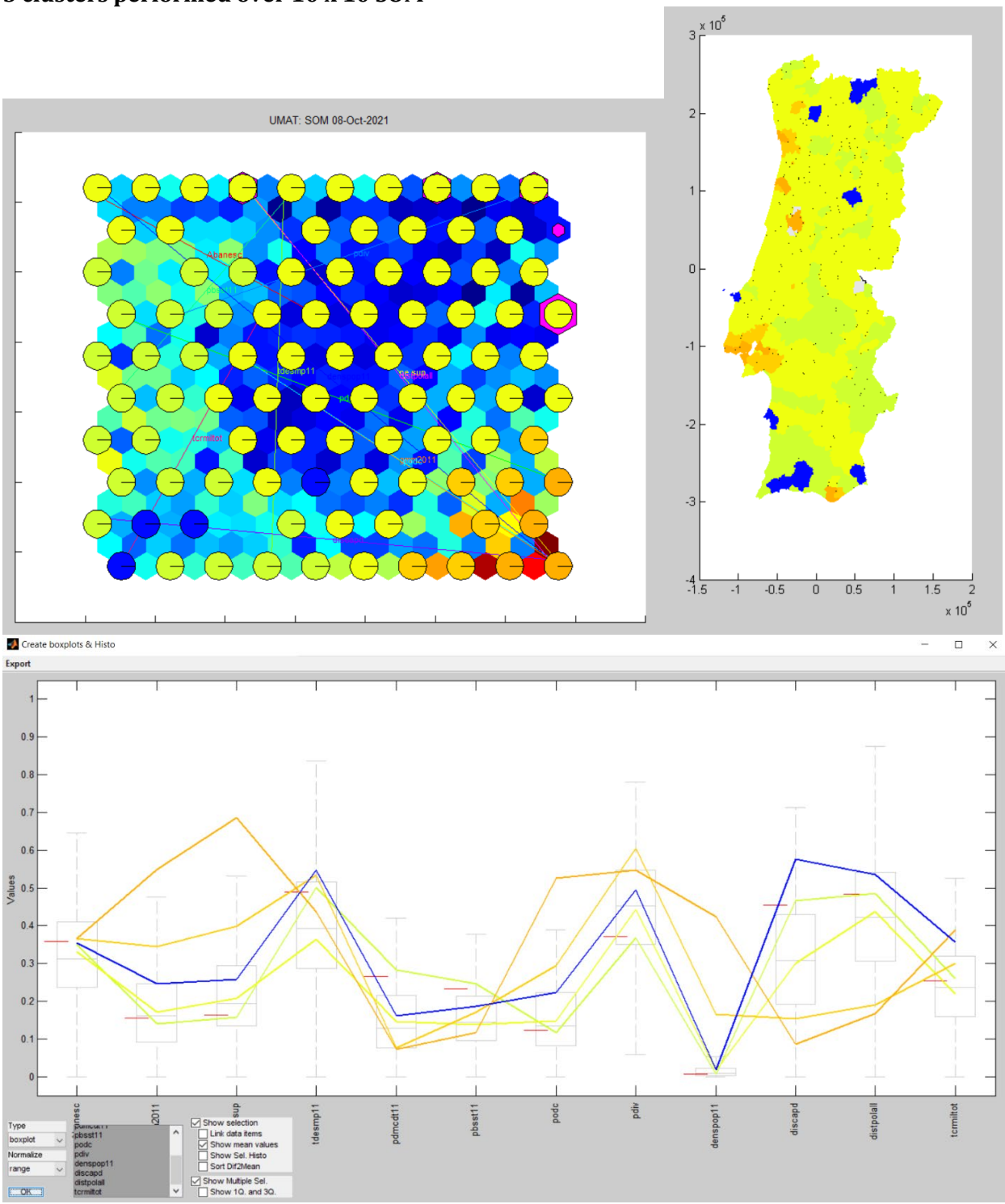



UMAT: SOM 08-Oct-2021
Show Hits
Show sel. based on item count
x 10^6



UMAT: SOM 08-Oct-2021
Show Hits
Show sel. based on item count
x 10^5

5 clusters performed over 10 x 10 SOM



The clustering was made over the 10 x 10 SOM, taking in consideration the previous results of the comparison with the 4 x 3 SOM. The 5 clusters reveal “small” portions of Portugal with a very particular behaviour when compared to the rest. This behaviour is represented in the boxplot performed for every variable and crossed with the clusters. Their biggest variances were reflected in gmm2011, ne_sup, pdoc, discapd, distpolall. The blue and dark orange clusters showed the most different behaviour regarding all clusters among the considered variables.

SOM Parameters and results:

x	y	Rough	Finetune	Initial Learning Rate	Train Length	Qualitative Error	Topological Error
4	3	1000	2000	0.5	linear	0.366	0.02
10	10	1000	2000	0.5	linear	0.263	0.01
5	5	1000	2000	0.3	inverse	0.33	0.02
1	8	1000	2000	0.3	inverse	0.37	0.02
3	3	1000	2000	0.3	inverse	0.38	0.04
5	1	1000	2000	0.5	linear	0.39	0.01
3	4	1000	2000	0.5	linear	0.37	0.004
2	2	1000	2000	0.3	inverse	0.42	0
3	2	1000	2000	0.3	inverse	0.40	0

Data normalised by range. All SOMs were trained sequentially, taking the neighbourhood radius as half of the maximum dimension, and reducing it by 50% for the fine tune phase. The learning rate decay curve was varied, along with the initial learning rate itself.

Conclusion

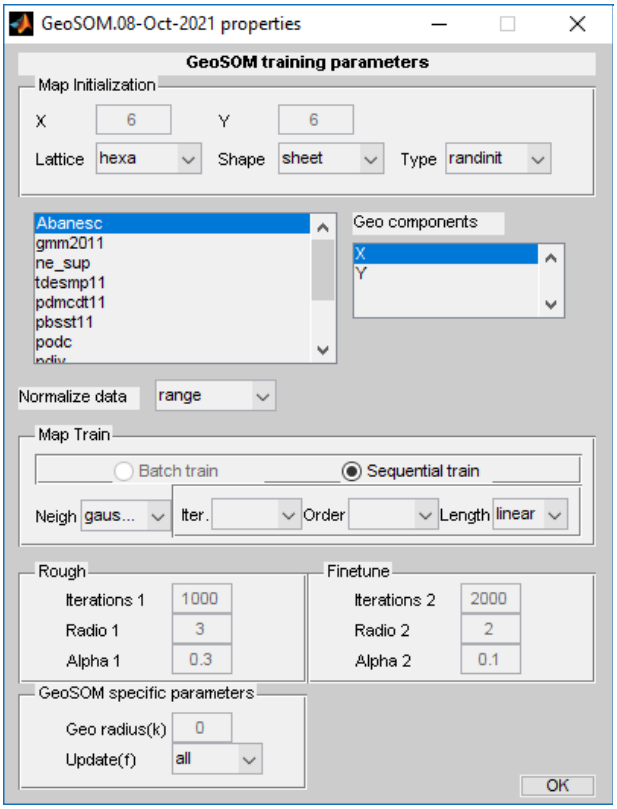
Linear learning rate decay yielded generally better results, with greater errors in the inverse experiments. This is due to the high reduction in neuron mobility within the first 200 epochs. From 9 SOM analyses, the most satisfactory classification was achieved with 100 neurons (10 x 10 map), with additional interpretation using smaller maps. The five clusters demonstrated clearly the rural-urban divide in Portugal's municipalities, as well as the concentration of development closer to the coast.

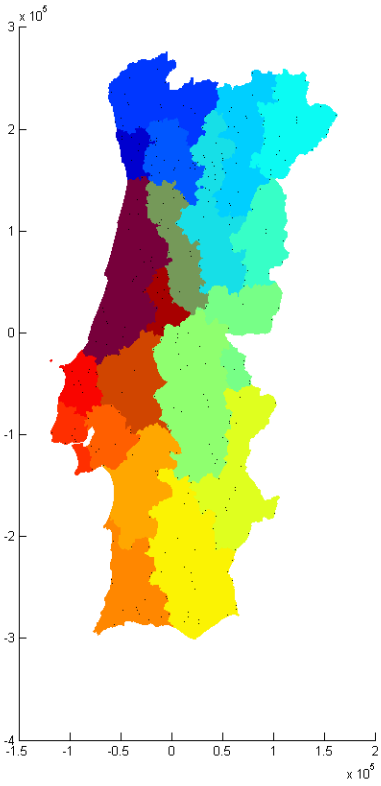
The 5 clusters can be generally described as:

	1	2	3	4	5
Abanesc	medium	medium	medium	medium	medium
gmm2011	very high	high	medium	very low	medium
ne_sup	very high	high	medium	very low	medium
tdesmp11	medium	very high	very high	high	low
pdmcdt11	low	very low	medium	very high	medium
pbsst11	low	medium	medium	very high	medium
podc	very high	high	medium	very low	low
pdiv	very high	very high	high	low	medium
denspop11	very high	high	medium	very low	very low
discapd	very low	low	very high	very high	medium
distpolall	very low	low	very high	very high	medium
tcrmiltot	high	medium	high	low	low

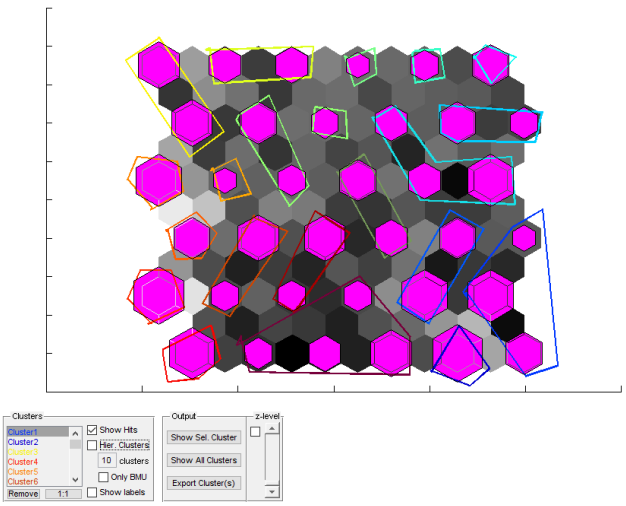
2.2. GeoSOM clustering

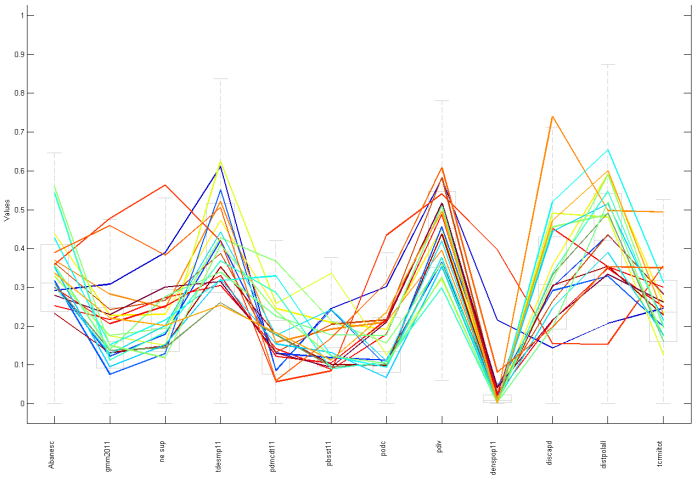
6x6 GeoSOM



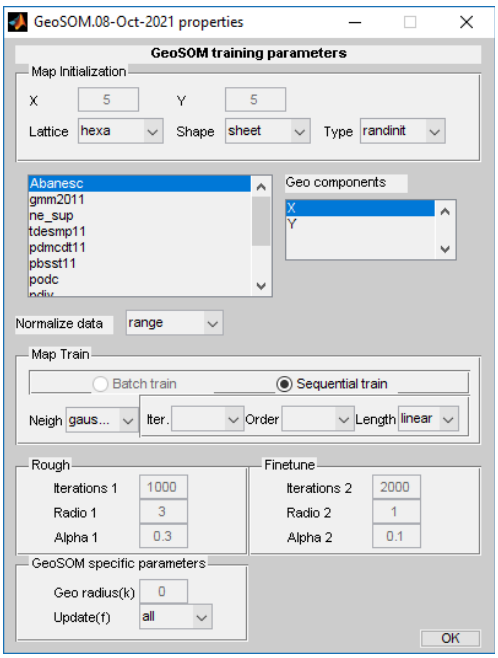


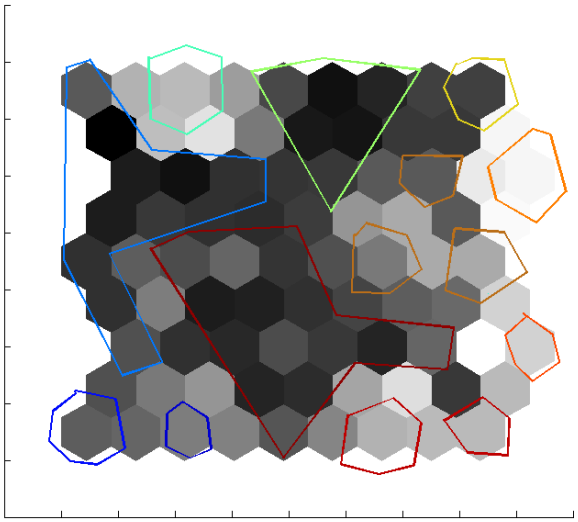
This analysis used 6 x 6 dimensions in order to broadly ascertain the number of neurons needed to map the clusters when geography is considered. The 36 nodes showed significant separation except for a few cases, and the number of clusters drawn could be reduced to 21. The map shows well-defined regional trends and the box histogram shows that component variation is optimal. The number of clusters is high enough to capture the scale and variance of Portugal, and low enough to be a useful simplification.

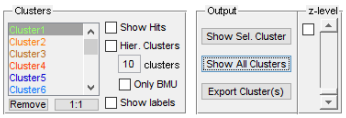




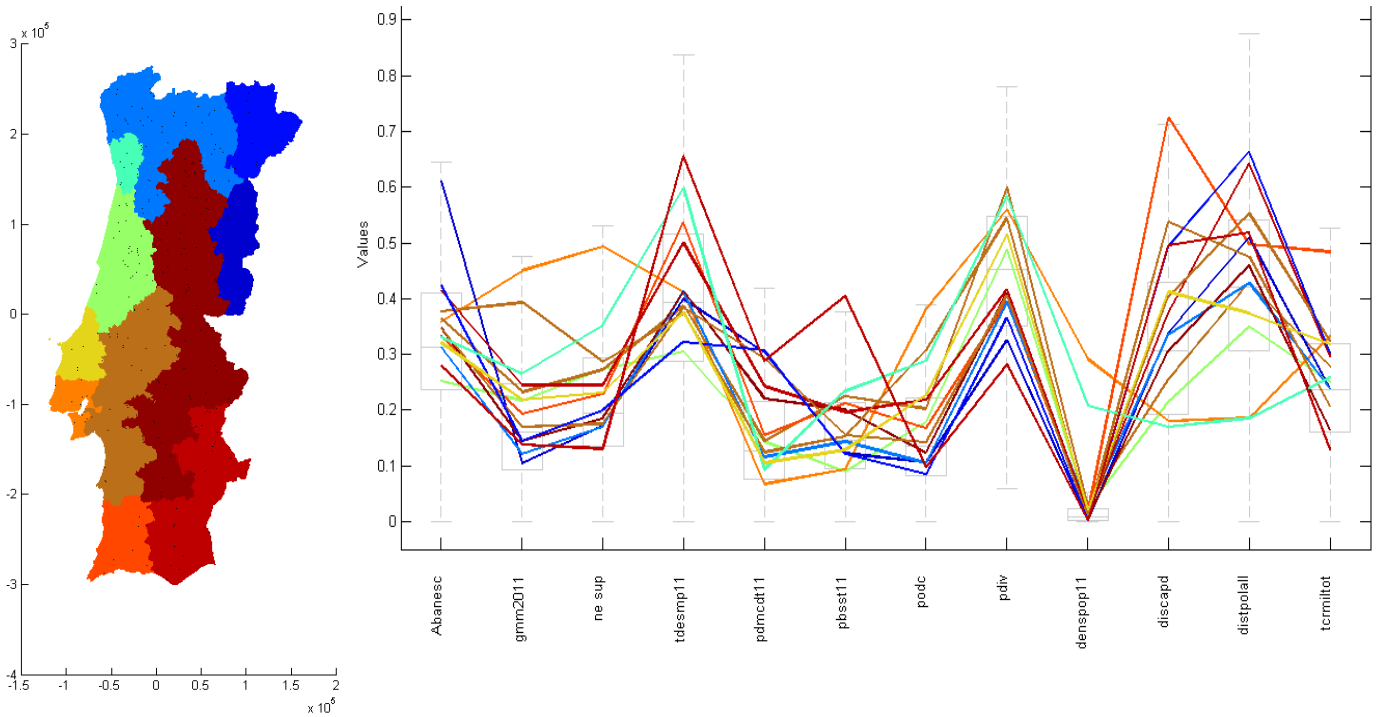
5x5 GeoSOM



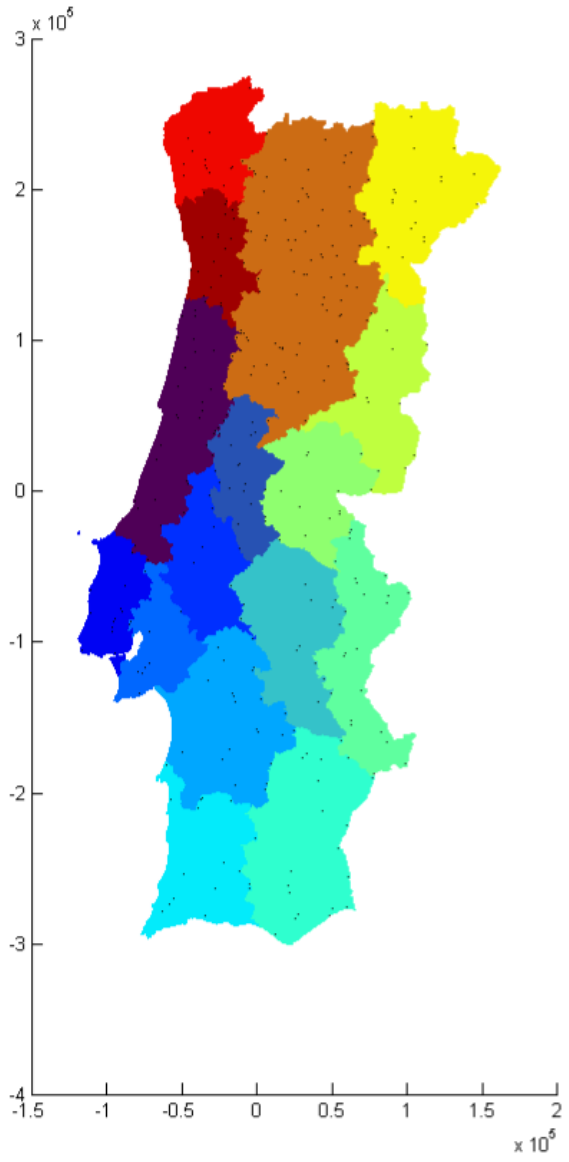


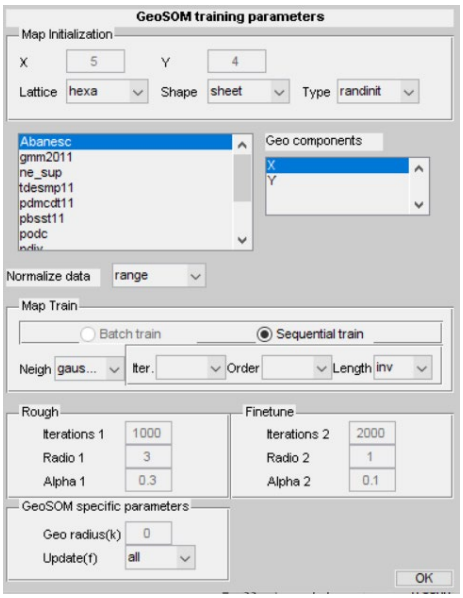


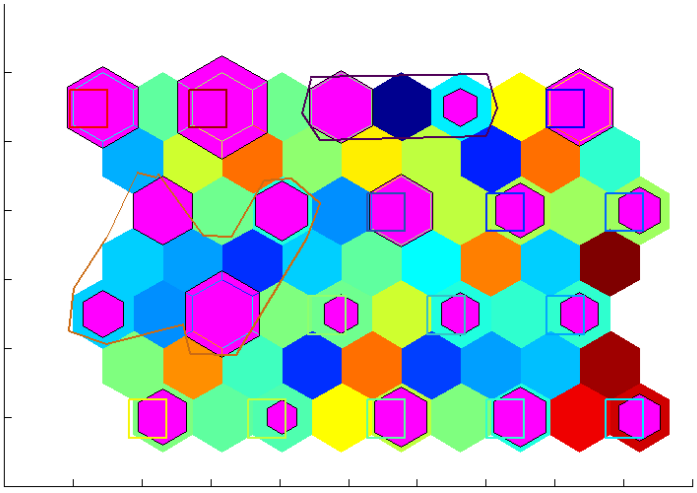
In an attempt to see whether the same information could be captured in fewer clusters, this GeoSOM used a square map of 5 x 5. The final output shows much broader trends in 14 clusters, and compared to 20+ spatial clusters it is too generalised.

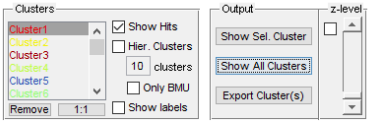


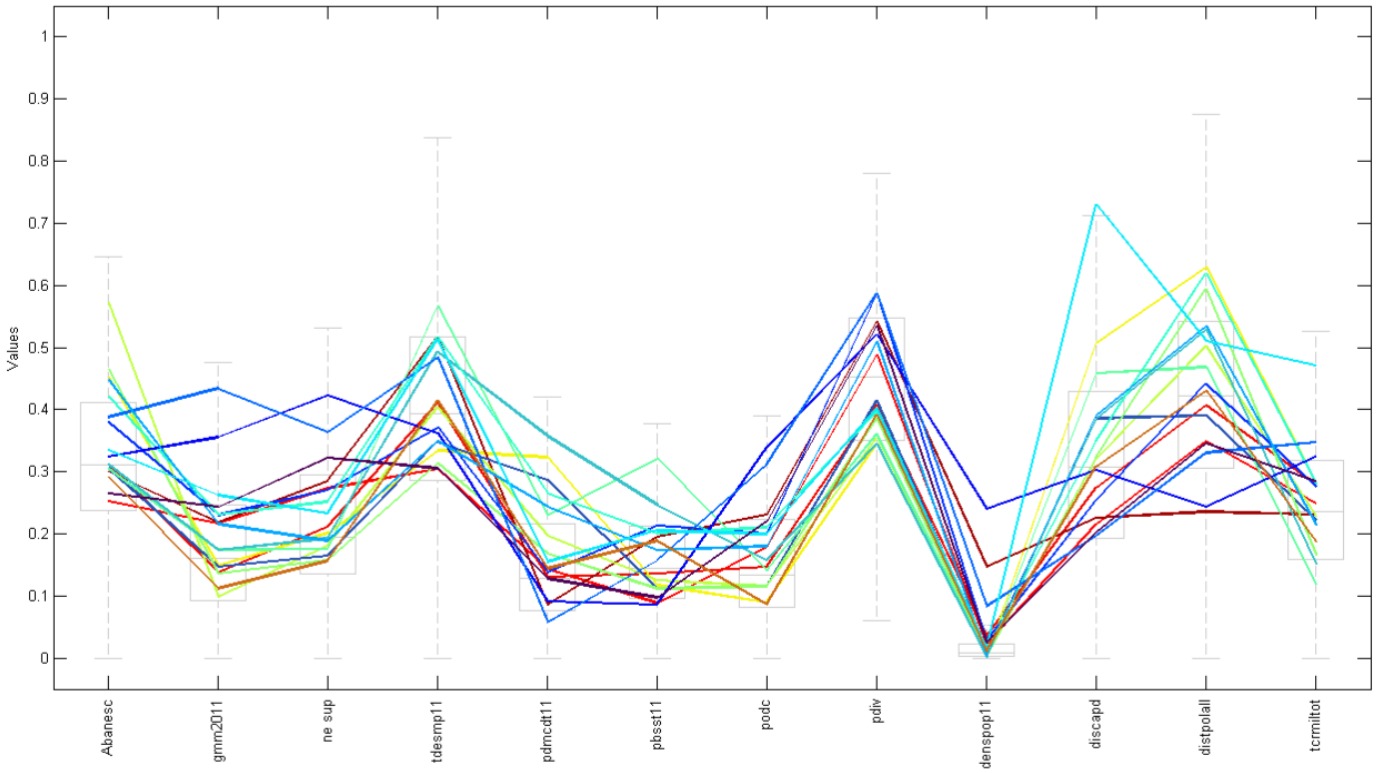
5 x 4 GeoSOM


x 10^5
3
2
1
0
-1
-2
-3
-4
-1.5
-1
-0.5
0
0.5
1
1.5
2
x 10^5


GeoSOM training parameters
Map Initialization
X
5
Y
4
Lattice
hexa
Shape
sheet
Type
randinit
Abanesc
gmm2011
ne_sup
tdesmp11
pdmcdt11
pbsst11
podc
Geo components
X
Y
Normalize data
range
Map Train
Batch train
Sequential train
Neigh
gaus...
Iter.
Order
Length
inv
Rough
Iterations 1
1000
Radio 1
3
Alpha 1
0.3
Finetune
Iterations 2
2000
Radio 2
1
Alpha 2
0.1
GeoSOM specific parameters
Geo radius(k)
0
Update(f)
all
OK




Clusters
Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Cluster6
Remove
1:1
Show Hits
Hier. Clusters
10
clusters
Only BMU
Show labels
Output
Show Sel. Cluster
Show All Clusters
Export Cluster(s)
z-level


Values
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Abanesc
gmm2011
ne sup
tdesmp11
pdmcdt11
pbsst11
podc
pdiv
denspop11
discapd
distpolall
tcrmiltot

This analysis resulted in 16 clusters, most of them are per neuron except for 2 that are groups of 2 and 4 neurons, put together due proximity evidenced by the colour. This analysis shows a similar approach compared with the last two except for a few clusters and the way they are grouped.

Conclusion

After defining 5 clusters into which the 278 municipalities could be classified into, and ascertaining optimal SOM parameters to obtain them, we ran several GeoSOM analyses. The best results are shown here, with modest map sizes of 36, 25 and 20 neurons. The final map consists of 16 geographically distinct clusters which adequately represent spatial variations in development and demography. As previously mentioned, 14 clusters were too few and lacked detail, but might be suitable for use cases that require a higher degree of simplification. Conversely, the 21 cluster division looks very promising. However, from the U-matrix it appeared that neurons were not quite as well distributed as in the final 5x4 GeoSOM. Portugal officially has 18 districts (established in 1835) and although they do not correspond directly with our geographically weighted clustering results, it is a good indication of the ideal number of administrative zones. In addition, the GeoSOM results provide divisions based on modern demographic realities. Soon the traditional districts will be 200 years old, and it may be time to reassess their usefulness.

3. Annexes

	DENS POP11	DISCAPD	GMM2011	PODC	TDES MP11	PBSST11	NE_SUP	ABAN ESC	PDMCDT 11	PDIV	DISTPOL ALL	TCRMIL TOT
Minimum	5.070	2879.400	673.07	49.8	5.1	0.112	2.764	0	0.004	0	743.85	9.53
Q1	25.195	18706.90	770.25	63.4	10.2	0.366	6.587	1.180	0.050	0.17	4217.42	24.53
Median	70.201	28177.20	841.87	72.25	12.1	0.498	8.273	1.550	0.080	0.22	5535.00	31.73
Q3	174.863	38281.90	932	87	14.3	0.680	11.101	2.040	0.132	0.27	6899.17	39.54
Maximum	7363.389	85221.20	1721.20	216.9	22.9	2.782	31.098	4.970	0.599	0.49	12105.02	103.63
Mean	311.491	29899.12	876.91	78.02	12.5	0.558	9.341	1.681	0.106	0.22	5566.11	32.58
St. Dev	866.025	14823.38	158.76	21.20	2.8	0.301	4.345	0.784	0.089	0.07	1953.00	11.74

Figure 1: Statistical summary of the evaluated variables.

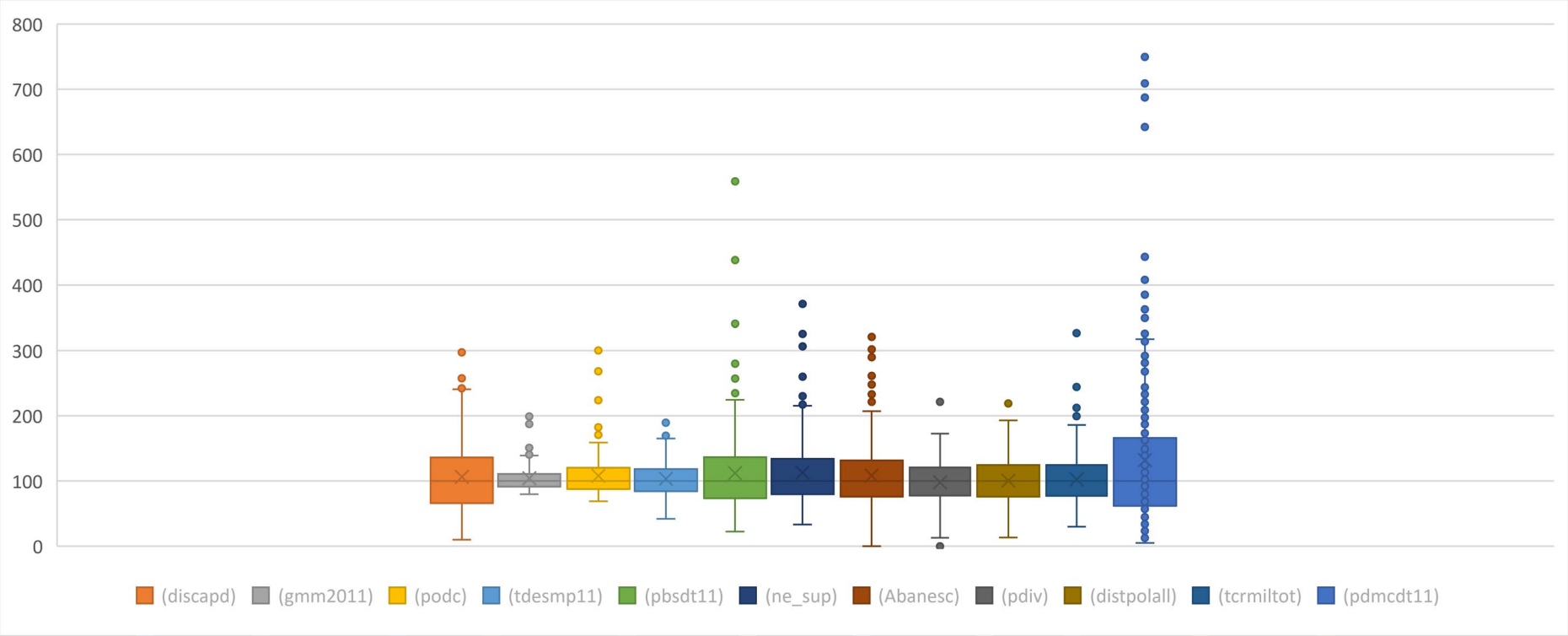


Figure 2: Box plots of the values indexed at median = 100. Density of population has not been included as even normalized it had so despair values that just made illegible the other box plots.