Clustering EU-SPI

ALBERT BUSCA LOPEZ[1]

ALBERT CARRERAS DE ODRIOZOLA[2]

Abstract

We dig into the EU-Social Progress Index published in 2020 to explore its territorial/regional clustering patterns. Are EU internal frontiers –clusters- the same as those of the member States? Do they have other influences? What determines belonging to different clusters? Is there any resemblance between the GDP per capita and the SPI regional distribution? We divide the EU NUTS-2 map into clusters using hierarchical clustering. We look for the optimal number of clusters and we compare the outcome with State borders, paying attention to discrepancies or to State combinations. Our main finding is that the optimal clustering is two, and that they are very robustly defined. Another major finding is that the EU-SPI clustering reveals major discrepancies with the per capita GDP clustering. Some NUTS-2 regions perform in SPI terms much better than expected by per capita GDP. On the contrary, some NUTS-2 regions perform much worse than expected by per capita GDP. The discrepancies suggest major public policy successes and failures.

[1] Graduate in Political Science, Universitat Pompeu Fabra
[2] Department of Economics and Business, Universitat Pompeu Fabra.

1. Introduction

The European Social Progress Index (EU-SPI) is an initiative from the European Commission aimed at creating alternative indices that do not include economic metrics to assess social progress in Europe and to help policy decision-makers. The EU-SPI is created in the context of the GDP and Beyond[1] initiative. It also stems from the Social Progress Imperative (SPI) the index in which the EU-SPI is based, issued annually at the country level and globally.

The EU-SPI is not the only one of its kind. In recent decades, various initiatives have appeared, trying to depart from exclusively economic measures such as the GDP. Particularly, the Human Development Index (HDI) created by the United Nations Development Programme (UNDP) has made great progress in creating an alternative to GDP that embraces other noneconomic components and is widespread in its use. Others, such as the Better Life Index made by the OECD or the World Happiness Report, also follow this line of work.

Among all these, the EU-SPI, first published in 2016, has some particularities that make it stand out. Firstly, the EU-SPI makes a great effort to explicitly and totally depart from economic metrics, and particularly those measured in monetary values. Composed by 55 indicators in the 2020 (second) version, not one of these is economic in nature. It focuses solely on other aspects of societal progress. Moreover, the EU-SPI is the only one to present its results at the subnational level, specifically at the second level of the Nomenclature of Territorial Units for Statistics (NUTS-2).

The EU-SPI's philosophy and expected applicability can be seen in the choice of its indicators. They are chosen to have a set of common characteristics. One of these is that

[1] More information here: https://ec.europa.eu/environment/beyond_gdp/index_en.html

they must “cover matters that can be addressed by policy intervention” (European Commission et al., 2020). This shows that one of the objectives of this index is to aid policymakers and the evaluation of public policies. Furthermore, a second characteristic to be shared by the indicators is that they must “measure outcomes, not inputs” (ibid), which in turn converts the EU-SPI in a tool for those same policymakers in measuring the outcomes of the policies that are implemented in each region regarding its impact to social progress. In summary, this demonstrates that the EU-SPI is designed to be an important and innovative tool that can help any actor interested in better designing policies. At the same time, it promotes more efficiency at allocating public resources to maximize the EU-SPI or, in other words, the social progress of the region.

The EU-SPI is not only distinct in its nature and application but also has been shown to be robust and better suited than other alternatives. The latter argument has been proven in relation to the GDP per capita and the HDI, demonstrating that the EU-SPI performs better at predicting social issues and outcomes than the other indices (Siddique et al., 2017). More specifically against the GDP per capita, the methodological paper presented by the European Commission delves into this relationship and states that, while some correlation can be found (0.62), the GDP per capita alone is unable to explain all variability in the EU-SPI (European Commission et al., 2020).

Furthermore, the internal consistency and robustness of the index has also be assessed. Recent literature has found that the EU-SPI is robust in its results across various methods of unbalance penalization (Annoni and Scioni, 2022). Beltran-Esteve et al., 2023, also conclude that the EU-SPI is also robust against changes in the normalization or aggregation criterion. Worried about this issue we tested for robustness on opinion variables by extracting them and recomputing the EU-SPI but found no significant change. The correlation between the original EU-SPI and the alternative was of 0.95.

In conclusion, the EU-SPI does not only present methodological innovations but also constitute a robust and consistent index that performs better to estimate social progress than traditional alternatives such as the GDP per capita or the HDI.

Taking all of this into account, the goal of this research is to study regional similarities and disparities regarding social progress in Europe. This will be achieved through the extensive analysis of the results of the EU-SPI second edition (2020). As such, the research question emerges as: How do EU regions differ from each other on social progress?

The initial hypotheses on this objective are the following:

H1. There are geographical patterns that affect social progress, specially at the country level. Exploring the relationship between the components of the EU-SPI, we suspect that geographical patterns will appear, highlighting regional differences of multicausal nature.

H2. Difference in social progress also reflect differences in the economic and political context of the regions. Despite the EU-SPI not including any of those variables, we suspect that, inevitably, the differences between regions will also respond to those contexts to some degree.

2. Methodology

EU-SPI was first published in 2016. The second edition was presented in 2020 and the third edition -labelled as EU-SPI 2.0- is the last edition available at the time of this writing, published in 2024.

When choosing to analyse the second edition as opposed to either the first or the third, several factors were taken into account. Firstly, from the entirety of the indicators of the EU-SPI 2.0, three quarters use data from 2021 to 2023. The second edition mainly extracts

its data from the 2016 to 2018 period while the first edition is mostly from the time period of 2011 to 2013.

As such, a lot of data that forms the third EU-SPI is from when the COVID-19 pandemic took place in Europe or when the extraordinary lockdown and health protecting measures were still in action. The first edition also represents abnormality extracting its data from the years in the aftermath of the 2008 financial crisis and during the sovereign debt crisis, with its effects probably still impacting the EU-SPI results. The second edition is the only one that responds to a situation of normality, making the findings based on it expectedly more loyal to the true situation of social progress in Europe.

However, choosing the second edition also has its drawbacks. Compared to the first, the United Kingdom is not present in the database and consequently 40 NUTS-2 regions are lost. On the other hand, the EU-SPI 2.0 has obviously more recent data than the second one. Ultimately, a decision was made to use the second edition to eliminate the effect of both the pandemic and the financial and sovereign debt crisis at the expense of losing some observations (compared to first edition) and using slightly older data (compared to the third edition).

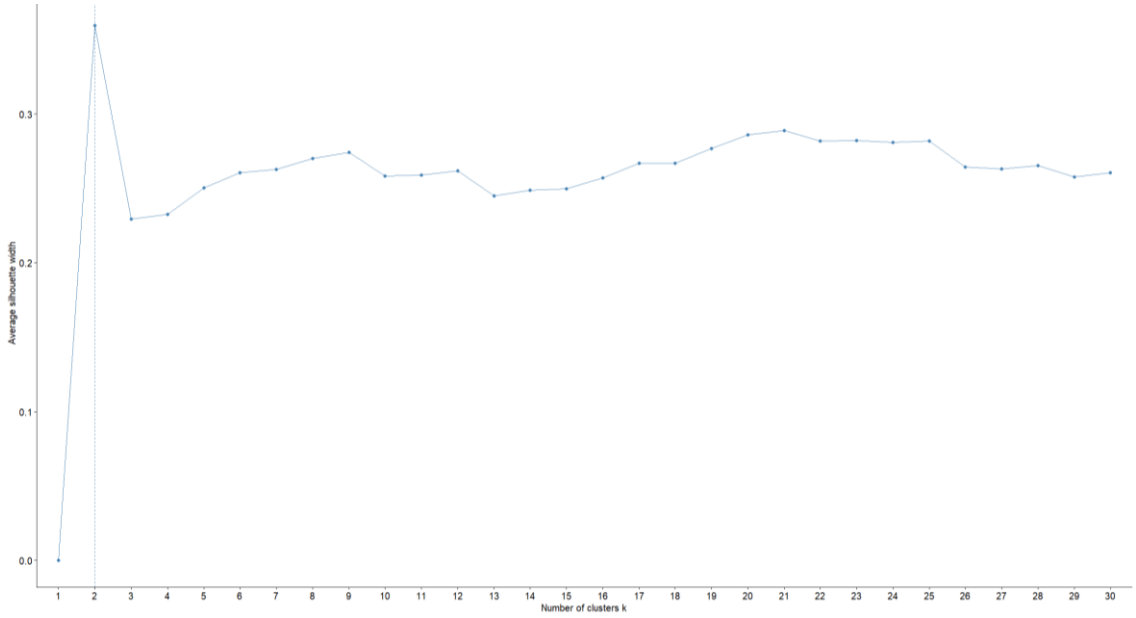
The impact of this decision is arguably limited. The EU-SPI 2.0 is made up of 53 indicators. Of those, only 13 are new, while the other 40 (75%) are the same. The components only suffer changes in labelling but not in nature, and the dimensions are the same. In turn, the second edition introduces 14 new indicators among its 55. Moreover, the methodology in the normalization and aggregation is mostly the same across the editions with minor changes. Although full comparability has to be discarded, the influence of these changes is limited. For that, the benefit of using the second edition exceeds the drawbacks discussed previously. See Annex 1 for details on the composition of the second edition of the EU-SPI by indicators, components, and dimensions.

To test for the existence of geographical patterns stemming from the EU-SPI, in line with our first hypothesis, a cluster analysis will be performed. The twelve components of the EU-SPI will be used to compute the clusters. More specifically, a hierarchical cluster method will be used. Hierarchical clustering has the benefit of not needing to specify the number of clusters when applied and allows to choose dynamically the granularity of the clusters by cutting the results at different levels. Among the methods available, the Ward method (Ward, 1963) was chosen.

For cluster validation, several steps were taken. First, the validity of the dendrogram was evaluated using the cophenetic correlation coefficient (Sokal and Rohlf, 1962) -or CPCC-, which has been proven to be a good tool to test for the global fit of the dendrogram on the data when validating clusters (Dubes and Jain, 1979). The dendrogram resulting of the Ward method not only had a satisfactory CPCC value of 0.64, but the value was higher than those of other methods tested. The Ward method was also chosen because, given the nature of this method, it maximizes internal variance. In other words, the observations that make up each cluster share similar characteristics across the twelve components of the EU-SPI, which aligns with our research objective.

The next step was to validate the best number of clusters to perform. Two indices were used: for its wide-spread use and good performance (Chouikhi et al., 2015; Arbelaitz et al., 2013), the Silhouette Index (Rousseeuw, 1987); and for being proven as one of the best performing indices (Chouikhi et al., 2015; Arbelaitz et al., 2013; Milligan and Cooper, 1985), the Calinski-Harabasz (CH) Index (Calinski and Harabasz, 1974). The mean Silhouette for every number of clusters ranging from 2 to 30 was computed. The resulting graph can be seen in Figure 1. The Calinski-Harabasz index was also computed for each number of clusters from 2 to 30.

Figure 1. Mean Silhouette for each number of clusters



As can be seen in Figure 1, the best performing number of clusters is two, with an average Silhouette of (0.36). The second-best value is found at twenty one. Also relevant, the first relative maximum after two is found at nine clusters. The CH index has similar results. The highest index results for two clusters (164.38). Then, for each cluster added the index diminishes. Following these results, the main analysis will be performed partitioning the dendrogram in two clusters.

After the first cluster analysis, we expect the existence of the geographical patterns will be either proven or disproven. After this, we aim to test the nature of this geographical patterns. For that, further analysis techniques will be used. First, a Linear Discriminant Analysis (Huberty, 1975) or LDA, secondly a Random Forest Analysis (Breiman, 2001; Genuer et al., 2010) or RFA, and lastly a Factor Analysis (Alhija, 2010) or FA.

The LDA finds a Linear Discriminant (LD1) as a combination of the twelve components so that it better separates the two clusters. With this, we can assess how the two clusters may be separated using the LD1, and which components contribute the most to that separation.

While the LDA focuses on linear relationships, the RFA will be useful in understanding the impact on cluster formation when the relationships are non-linear. At the same time, RFA can also discern the impact of each component separately for each cluster.

Finally, the FA will result in the calculation of several factors, each of which will represent a latent variable in the dataset. The grouping of the components in these factors will aid in understanding if there are core aspects of the components that cannot be seen in the original variables.

Even though this main analysis is deemed to be the most relevant, further partitions are also believed to be important. Firstly, a partitioning at nine clusters will be performed. Both Silhouette index and CH index are relatively good compared to most alternatives and it allows seeing how those two initial clusters are further subdivided without atomizing the groups.

The final partition studied will be at 21 clusters. Despite the value being lower in the CH index, it has the second highest Silhouette index. It also represents the optimal subdivision close to the same number of countries included in the EU-SPI (27). This partition will, therefore, be of the utmost relevance to understand the impact of country boundaries to social progress.

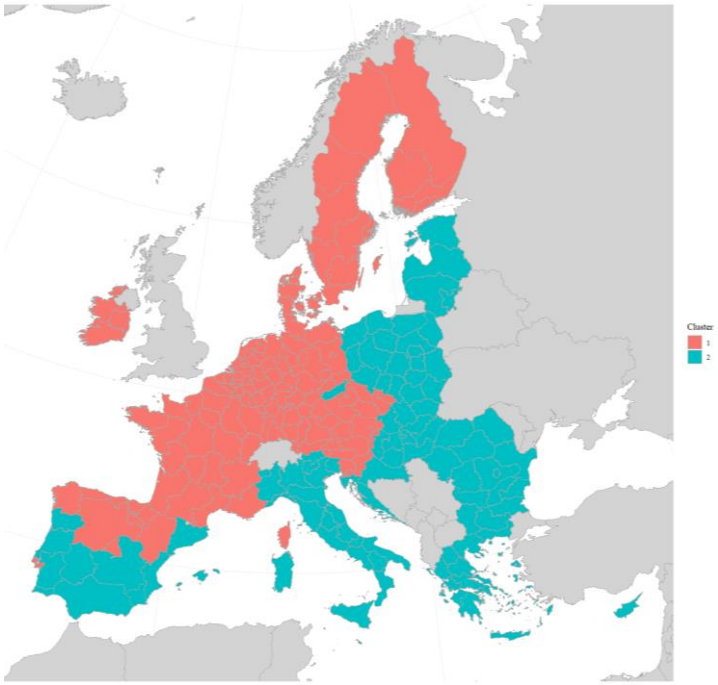
The cluster analysis, aligned with the two hypotheses, will be conducted using the components that make up the EU-SPI (H1), but also contrasted with other important contextual variables: the per capita GDP and the national borders (H2). By doing this, we will be able to test whether the clusters reflect economic, political or social characteristics.

3. Results

3.1. Two clusters

The first step was to compute a dendrogram stemming from the values of the 12 components in all NUTS-2 regions. The resulting dendrogram and the one used throughout the study, can be seen in Appendix 2 The map in Figure 2 represents cluster membership of each region when cutting the dendrogram in two clusters. At this level, most national borders are preserved. The most notable exception is Spain. Europe is divided into two regions, the first comprised by central and northwestern countries and the second comprised by eastern and southern countries.

Figure 2. EU Map at the NUTS-2 level of the SPI clustering results



The mean of each component by cluster was computed and the results can be seen in Table 1. The first cluster has higher average values compared to the second cluster in every component except for *Personal Security*. Therefore, the first cluster performs better than the second one across components. The largest absolute and relative difference between clusters can be found in *Access to Advanced Education*.

Table 1. Mean for each EU-SPI component by cluster

EU-SPI Components	Cluster 1	Cluster 2
Nutrition and Basic Medical Care	85.73	75.33
Water and Sanitation	93.83	83.80
Shelter	88.55	70.63
Personal Security	70.23	73.45
Access to Basic Knowledge	78.77	68.82
Access to ICT	82.64	65.60
Health and Wellness	71.19	58.70
Environmental Quality	52.84	37.57
Personal Rights	56.32	39.59
Personal Freedom and Choice	72.30	54.24
Tolerance and Inclusion	70.45	49.02
Access to Advanced Education	66.55	42.36

With all this, it can be stated that there are two different geographical areas (northwest and southeast) with distinct social progress levels in Europe. The two clusters are well separated and there is homogeneity within the clusters. However, to know the nature of these patterns requires a deeper analysis. As stated before, we perform a Linear Discriminant Analysis (LDA), a Random Forest Analysis (RFA) and a Factor Analysis (FA).

The LDA aids in discovering the linear relationships between the various components and cluster membership. The values shown in Table 2 are the scaling factor for each component. A larger -in absolute terms- scaling factor is more impactful to the first linear discriminant (LD1) and, therefore, more impactful to the separation of the clusters. Negative values are associated with membership to the first cluster while positive values are associated with membership to the second cluster.

Table 2. First Linear Discriminant (LD1) loading for each EU-SPI component

EU-SPI Components	LD1
Nutrition and Basic Medical Care	0.055
Water and Sanitation	0.032
Shelter	-0.163
Personal Security	0.034
Access to Basic Knowledge	-0.031
Access to ICT	0.005
Health and Wellness	-0.062
Environmental Quality	0.005
Personal Rights	-0.038
Personal Freedom and Choice	-0.016
Tolerance and Inclusion	-0.011
Access to Advanced Education	-0.007

Shelter stands out as the most important component, with a scaling factor twice that of the second highest -*Health and Wellness*-. The third most important component is *Nutrition and Basic Medical Care*. These are the components that have a stronger linear relationship with cluster membership. Because *Shelter* has a negative scaling factor, higher values of *Shelter* will be associated with membership with cluster 1. The same happens with *Health and Wellness*. On the contrary, high values of *Nutrition and Basic Medical Care* are associated with membership to the second cluster.

Next, we will compute an RFA. Random Forest Analysis is a technique mainly used for classification. In this case, we will use it to account for non-linear relationships between the components of the EU-SPI -as predictors- and cluster membership -as the predicted value-, which will aid us in nuancing the findings in LDA.

The results can be seen in Table 3. The first column presents global performance for each component. As such, *Shelter* stands out again as the most important variable. It has the largest Mean Decrease Accuracy, meaning that the model heavily relies on *Shelter* to accurately predict cluster membership. *Access to ICT* has the second largest value, and therefore it is seen as a great variable to classify observations to both clusters. The second and third column show the same metric but disaggregated to each cluster. This gives us valuable information. For example, *Health and Wellness* and *Nutrition and Basic Medical Care* are particularly useful when classifying to the first cluster compared to the second. On the contrary, *Personal Freedom and Choice* is more relevant for classifying to the second cluster compared to cluster 1.

Table 3. Random Forest Analysis results

EU-SPI Components	**Mean Decrease Accuracy**	**Cluster 1[a]**	**Cluster 2[a]**
Nutrition and Basic Medical Care	12.408	11.646	6.007
Water and Sanitation	3.873	-1.038	4.220
Shelter	24.657	20.384	16.576
Personal Security	4.059	3.303	2.569
Access to Basic Knowledge	13.080	10.782	10.998
Access to ICT	18.763	17.734	10.549
Health and Wellness	11.910	11.232	6.806
Environmental Quality	6.514	5.359	3.855
Personal Rights	14.086	9.628	10.638
Personal Freedom and Choice	13.477	8.051	11.148
Tolerance and Inclusion	15.599	12.931	10.608
Access to Advanced Education	9.013	6.369	6.752

[a]Decrease in accuracy for the selected cluster

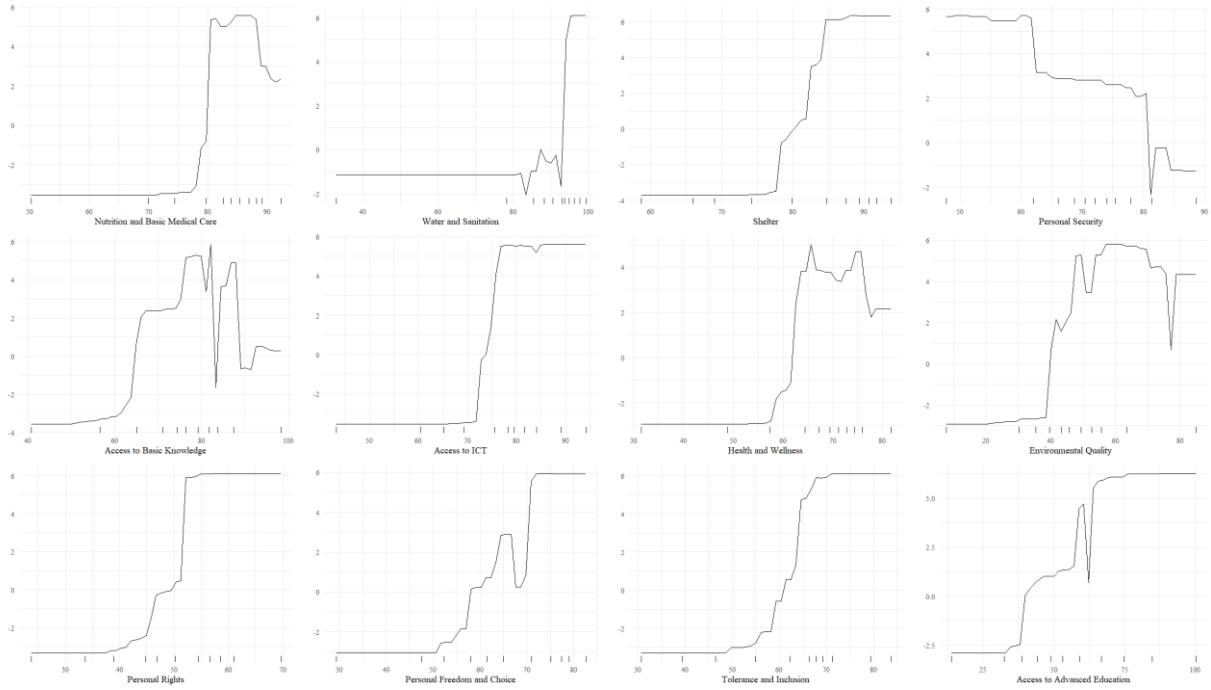
In summary, *Shelter* is the most important component in both models. *Health and Wellness* and *Nutrition and Basic Medical Care* diminish their impact when non-linear relationships are considered in favour of *Tolerance and Inclusion* or *Personal Rights*. *Access to ICT* has a strong non-linear relationship with cluster membership. Overall, these variables seem to be the most important at determining the classification of the observations.

According to both LDA (Table 2) and RFA (Table 3), *Shelter* is very important for classifying observations to cluster 1. The two components related to health or physical well-being, are relatively more important for this cluster than to cluster 2 classification. *Access to ICT* is very important in RFA, signalling a non-linear relationship. With the information in table 1, cluster 1 is then characterised by having high levels of those components.

On the other hand, for classifying to cluster 2, *Shelter* is again the most important variable. *Personal Freedom and Choice*, which also includes employment related indicators, is also relevant specially when compared to cluster 1. Once more, with table 1 we can see that cluster 2 is defined by having lower levels of these components.

To understand exactly the relationship between each component and cluster membership, whether linear or not, partial dependence plots (PDPs) were computed. These show, according to the Random Forest model, the log odds of being in a selected cluster (in this case membership to the first cluster; the PDPs for the second cluster would be the reverse image), for any value of the component. The results of the PDPs can be seen in Figure 3.

Figure 3. Partial dependence plots for cluster 1 by components



As can be expected, the probability of being in cluster 1 generally increases as the values for any component increase, except for *Personal Security*. The short lines seen along the x-axis represent the deciles of the data, to understand the observations distribution relative to the PDPs.

Most of these graphs show very low probabilities of being in cluster 1 until a certain threshold, where the probabilities of being in cluster 1 rise sharply. This happens in most components and most notably in the most relevant components according to previous models such as *Shelter* or *Access to ICT*.

This has a great impact on the analysis of the clusters. It implies that for most components there is a certain critical value that separates cluster 1 from cluster 2, with very little gradation. This validates even further both cluster distinctiveness.

It also shows tipping points were the impact of public policies either improving or diminishing the component's value may have a very critical impact on cluster membership, and at the same time other range of values for where larger changes may not result in a change on cluster membership. This is the case for most components, but

particularly so for those components that where assigned more importance by LDA and RFA. That is, among others, *Shelter* or *Access to ICT*.

After assessing the importance of each component, we performed a Factor Analysis (FA). The objective is to discover if there are any latent variables that affect the components and can explain cluster membership. This model was made using an oblique rotation, which means that the factors can be correlated to each other. The model has four factors.

Prior to computing the model, the validity of the data was tested for an FA. Firstly, we computed the Kaiser-Mayer-Olkin measure (Kaiser, 1970) or KMO. KMO indicates the common variance that each component share with each other in relation to the unique variance. Above 0.8 is considered the best threshold (Kaiser, 1970) although all values above 0.5 are considered acceptable. In our case, the global KMO value was 0.86. All components except three passed the 0.8 threshold and only one was below 0.5: *Personal Security* with 0.19. Bartlett's test (Barlett, 1937) assesses whether the variables presented (in our case the EU-SPI components) are correlated or not with each other, which would make them suitable for FA. With a p-value close to 0, it is tested against the null hypothesis that there is no correlation.

Taking this into account, the model was calculated as stated above, although *Personal Security* was omitted from the model. Furthermore, to validate for the number of factors, several models were performed, and the model fit was calculated for each. We used both the Root Mean Square Error of Approximation (Steiger & Lind, 1980; Browne & Cudeck, 1992) or RMSEA and the Bayesian Information Criterion (Schwartz, 1978) or BIC. RMSEA mainly considers the residuals, and BIC takes into account the global fit, while both are penalised by complexity. The four-factor model was found to have optimal values for both indices, so the model was computed using this number. The loadings for every

component in each factor, alongside the communality and complexity, can be found in Table 4.

Table 4. Factor Analysis results

EU-SPI Components	F1	F2	F3	F4	Communality	Complexity
Nutrition and Basic Medical Care	1.03	0.01	-0.08	-0.01	0.94	1.01
Water and Sanitation	0.39	0.03	0.11	0.44	0.63	2.10
Shelter	0.22	0.58	0.31	-0.11	0.86	1.95
Access to Basic Knowledge	-0.15	0.16	-0.02	0.82	0.70	1.15
Access to ICT	0.29	0.13	0.57	0.08	0.89	1.64
Health and Wellness	0.99	-0.11	0.06	-0.11	0.89	1.06
Environmental Quality	-0.26	0.06	0.94	-0.17	0.56	1.23
Personal Rights	0.10	0.27	0.58	0.13	0.88	1.61
Personal Freedom and Choice	-0.14	0.88	-0.02	0.30	0.92	1.29
Tolerance and Inclusion	0.38	0.19	0.56	-0.11	0.93	2.15
Access to Advanced Education	0.03	-0.29	0.89	0.21	0.74	1.34

A higher loading represents higher importance of that component in the given factor. Communality and complexity are metrics that are useful in understanding how well the model explains each component. Communality is the proportion of the variance in a given component explained by the model. As such, we can see how *Nutrition and Basic Medical Care* is almost entirely explained by the model while only just above half *Environmental*

Quality is within the model. Complexity is a measure that informs us of in how many factors is the component present. Again, *Nutrition and Basic Medical Care* has low complexity, meaning that it is both almost entirely explained by the model and almost entirely explained by only one factor (F1).

Regarding the components that were assessed to be most important by RFA and LDA, *Shelter* is particularly important in the second factor. It is almost entirely within the model, with a communality of 0.86. The model also explains around 90% of other important components such as *Access to ICT*, *Personal Freedom and Choice* or *Tolerance and Inclusion*.

Focusing on the factor loadings, we see that *Nutrition and Basic Medical Care* (with 1.03) and *Health and Wellness* (with 0.99) are the components that primarily make up the first factor. It is therefore named as Health. Factor 2 is mostly *Personal Freedom and Choice* and *Shelter*, and as such named Freedom, Employment and Housing (FEH). It should be noted that more than half of *Personal Freedom and Choice* indicators are related to the job market although it is not implied in its name. The fourth factor is composed primarily by *Access to Basic Knowledge*. It is named as "Basic Education".

The third factor is most importantly composed by *Environmental Quality* and *Access to Advanced Education*, but it has minor loadings on several other components: *Access to ICT*, *Personal Rights* and *Tolerance and Inclusion*. It is of relevance that *Personal Rights* although not implied by its name, is mostly composed by indicators relating to institutional trust. This factor groups most components that were not relevant in previous factors. It has a broad variation of indicators inside as a result. It will be named "Eco-social and Institutional Quality (ESI Quality)".

Table 5 shows the variance explained by each factor. The first row is directly correlated to table 4, as it is the sum of the squared loadings for each factor. The second

and third rows represent the proportion of variance in the data explained both individually and cumulatively. The fourth and fifth rows represent the proportion of variance explained within the model. As can be seen, the most variance in the original data is explained by ESI Quality, followed by Freedom, Employment and Housing. The four factors combined explain 81% of the variance in the data.

Table 5. Variance explained by each factor

	F3	F2	F1	F4
SS loadings	3.236	2.783	1.697	1.231
Proportion Var	0.294	0.253	0.154	0.112
Cumulative Var	0.294	0.547	0.701	0.813
Proportion Explained	0.362	0.311	0.190	0.138
Cumulative Proportion	0.362	0.673	0.862	1.000

The correlation matrix between factors can be found in Table 6. The highest correlation between components is between Eco-Social and Institutional Quality and Health and the least is between Health and Basic Education. As such, higher levels of Health tend to generally be correlated with higher levels of ESI Quality.

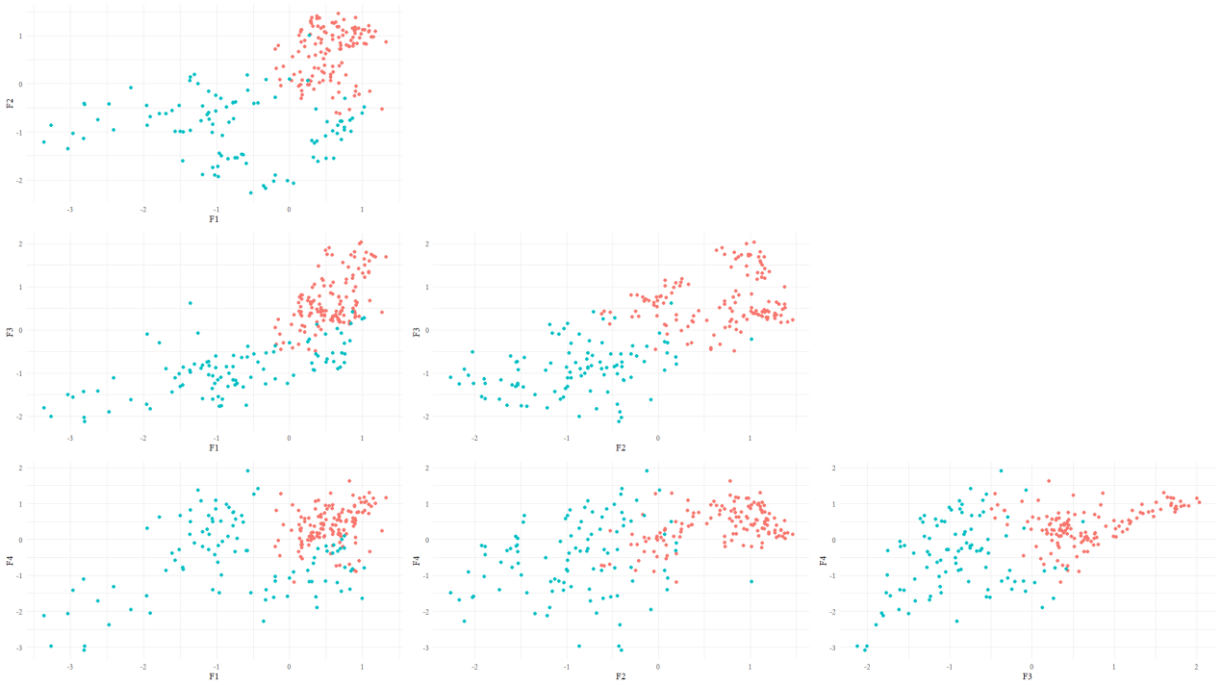
Table 6. Correlation matrix between factors

	F1	F2	F3	F4
F1	1.000	0.491	0.733	0.408
F2	0.491	1.000	0.672	0.432
F3	0.733	0.672	1.000	0.487
F4	0.408	0.432	0.487	1.000

The scatter plots of these correlations can be seen in Figure 4. The colours represent cluster membership, with red being cluster 1 and blue being cluster 2. For most of the correlations we see a linearity where higher values of both factors are related to cluster 1

membership. A detailed look has to be made, however, in the relationship between Health and Basic Education. Here, lower values of both are related to cluster 2 membership. However, the graph shows a fork at a certain point at which cluster 2 regions are either improving only Health or only Basic Education and only regions in cluster 1 seem to be excelling in both areas at the same time.

Figure 4. Graphical correlation matrix between factors



To know whether these factors can accurately predict cluster membership, we compute the clusters again. The clusters are computed the same way as before, using the Ward method in hierarchical clustering and then cutting the dendrogram at two clusters. Now, however, the predictor variables are the four factors instead of the EU-SPI components. A confusion matrix was performed between the two cluster models. Only 12 regions were misclassified -different between the original clusters and the factor clusters- which represented 5.19% of the total.

The misclassified regions can be seen in Table 7. All the regions in Spain that were in cluster 1 -with the exception of Madrid, the Basque Country and Asturias- were predicted to be in cluster 2 by the factors. All the Portuguese regions except Lisbon changed to

cluster 1 which, considering Lisbon was already in cluster 1, unified the country. Malta changed from cluster 2 to cluster 1 and Bratislava changed from cluster 1 to cluster 2.

Table 7: Regions that change cluster membership

NUTS Code	Region Name	Original Cluster	Factor Cluster
ES11	Galicia	1	2
ES12	Principado de Asturias	1	2
ES13	Cantabria	1	2
ES23	La Rioja	1	2
ES24	Aragón	1	2
ES41	Castilla y León	1	2
MT00	Malta	2	1
PT11	Norte	2	1
PT15	Algarve	2	1
PT16	Centro	2	1
PT18	Alentejo	2	1
SK01	Bratislavský kraj	1	2

The original clusters showed to be robust, and the factors were validated as good predicting variables for cluster membership. The region with the most confusion was the Iberian Peninsula. If we were to consider these changes, the country borders would be almost intact except for three regions in Spain and one region in the Czech Republic (Severozápad).

A Random Forest model with the factors as independent variables was computed to test for the importance of each factor at classifying the observations. The results can be seen in Table 8. Eco-social and Institutional Quality (F3) which is composed mainly by *Environmental Quality* and *Access to Advanced Education* and secondly by *Access to ICT*, *Personal Rights* and *Tolerance and Inclusion*, was the best factor at classifying the

observations on average as can be seen by its highest value in Mean Decrease in Accuracy. Freedom, Employment and Housing (F2) followed. Health (F1) was the best classifier for cluster 1 and the worst for cluster 2, pointing out to the discrepancies between clusters. Mainly, it highlights that cluster 1 has a very distinct range of values (mostly high) for this factor, while cluster 2 is more disperse.

Table 8. Random Forest results with Factors as predictor variables

Factors	Mean Decrease Accuracy	Cluster 1	Cluster 2
F1	30.47503	30.48155	14.57364
F2	32.91022	26.95012	20.16020
F3	34.70631	25.45501	25.69503
F4	20.70812	16.13208	16.37120

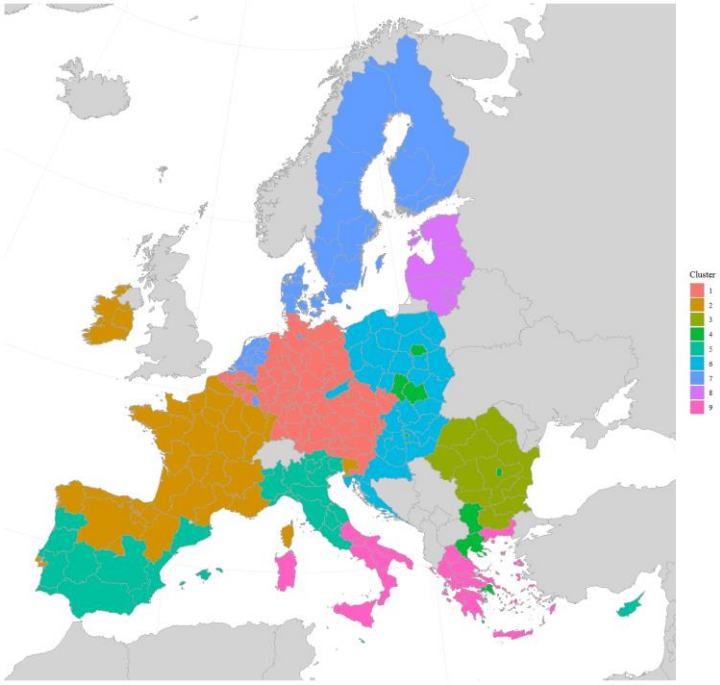
3.2. More clusters

To understand how the two clusters operate internally, it is important to see how they further divide. Namely, we will use the same dendrogram computed prior but cut it at the 9-cluster level (Appendix 5) and at the 21-cluster level (Appendix 6). Twenty-one clusters has the second highest Silhouette Index behind the two clusters, so it is important to see how they distribute themselves. On the other hand, 9 is the first local maximum after two, so it can be useful to see how the two clusters are first divided.

The results of subdividing the clusters into 9 can be seen in Figure 5. The former cluster 1 is divided into three clusters (1, 2 and 7) while the former cluster 2 is divided into six clusters (3, 4, 5, 6, 8, and 9). The national borders that had been mostly preserved for 2 clusters are now broken, mostly in southern Europe. The most northern regions in Spain (but Catalonia) are joined with France while the Spanish eastern and southern regions are combined with northern Italy. At the same time, the south of Italy goes with most parts of Greece. On the other hand, France, Germany, and most of central Europe

remain with their national borders. Most notably, the Nordic countries are joined into only one cluster.

Figure 5. NUTS-2 Map at 9 Clusters



Additionally, to better understand the clusters, in Table 9 it can be seen, for each cluster, the mean per capita GDP and the mean EU-SPI level. The highest SPI cluster (7) corresponds to the Northern countries, that also include Hamburg, the Netherlands and Luxemburg, which represent the top performers of the former cluster 1. Their SPI advance over the second SPI ranked (2) -France, Ireland, Northern Spain, Lisbon, Wein and Western Slovenia- is of twelve percent, while their per capita GDP advance is of seventeen per cent. Along with the third SPI ranked (1) (most of Germany, Austria, Czechia, Belgium and Eastern Slovenia), these three clusters represent the former cluster 1.

Out of the nine clusters, the difference between the cluster 2 and cluster 1 is of one per cent in SPI in favour of cluster 2 but of five per cent in per capita GDP in favour of cluster 1. Cluster 2 is thus getting more social progress out of their per capita GDP than cluster 1. The difference between these two clusters is that cluster 1 has higher values in

Personal Freedom and Choice but lower values in *Access to Advanced Education* or *Environmental Quality*. This points out that cluster 1 represents a more industrial and labour oriented social progress while cluster 2 bases its social progress in other areas such as environmental policies and higher education levels.

Table 9. Mean GDPpc and EU-SPI values by clusters at 9 clusters

Cluster	Mean GDPpc	Mean EU-SPI
1	33,496.72	71.01177
2	31,917.07	71.71985
3	15,500.00	47.42561
4	32,262.50	59.55308
5	27,128.57	63.22684
6	18,533.33	59.25026
7	37,493.55	80.43914
8	26,700.00	66.22437
9	17,015.79	55.83056

Regarding the former cluster 2, the fourth SPI cluster (8) consists of the three Baltic countries, that are well above their per capita GDP ranking (sixth). The fifth SPI cluster (5) -that is also the fifth in per capita GDP- consists of Northern and Central Italy, Eastern and Southern Spain, almost all continental Portugal (but Lisbon), Cyprus and Malta. The main difference between cluster 8 and 5, is that while 8 excels in education-related components, 5 excels in health-related components. The aggregated SPI level difference is of just 3 points. The sixth SPI cluster (4) comprises a number of regions that are the location of Eastern European capitals or major towns. Their SPI ranking is worse than their per capita GDP standing (fourth). They include the Attica, Central Macedonia (where Thessaloniki is), the Sofia region in Bulgaria (Yugozapaden) the Bucarest region,

Budapest, Malopolskie (lower Poland, capital Cracow), Slaskie (capital Katowice) and the Warszaw region. Although this cluster (4) does good in *Personal Freedom and Choice* and education-related components, it has the lowest value for *Environmental Quality*, low values in components inside the Health factor, and in *Tolerance and Inclusion*. As such, the difference with other clusters inside the former cluster 2, is that cluster 4 is more oriented towards its industry and labour market potentially costing the region lower values for Health and *Environmental Quality*. The seventh SPI cluster (also the seventh in per capita GDP) corresponds to most of Poland, Slovakia and Hungary without their capital regions, all Croatia and a Czech region – Severozapad (Karlovy Vary region). The eight (in both SPI and pc GDP) cluster (9) includes most of Greece (but Athens and Thessaloniki regions) and Southern Italy, Sardinia and Sicily – the classical Mezzogiorno. Similar to cluster 5, this cluster differs from others in that it has better values for Health related components but low values for education-related components. The worst cluster in SPI and pc GDP (3) is made of Bulgaria and Romania, except their capital regions.

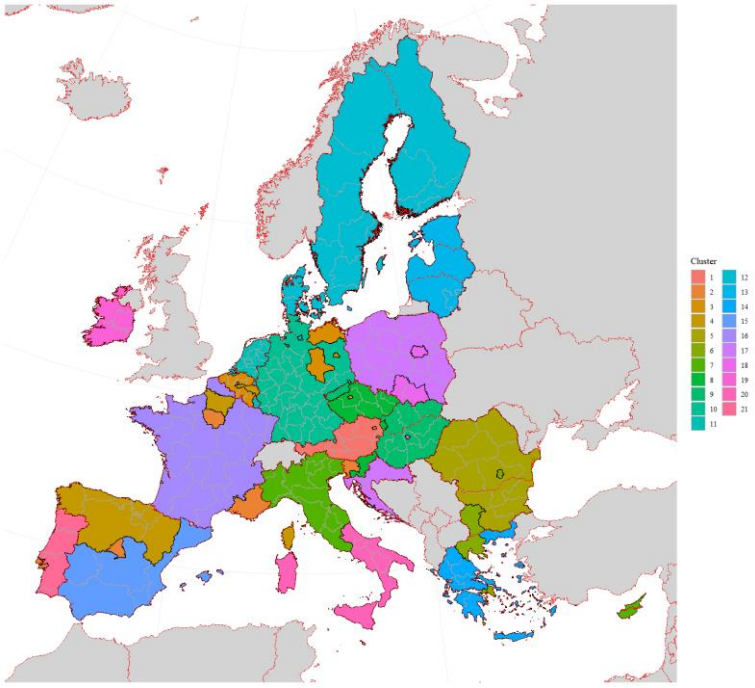
In summary, there are parts of Europe that have been excellent at promoting their SPI with the economic resources at hand, while others not so much. It is also clear that being a capital region matters very positively for the SPI the more economically backward the country is. It is also impossible to ignore that some regions that have been for a very long period the EU goal of supporting regional policies have failed to perform economically and socially – the Mezzogiorno is a case in point. Just on the other side of the coin, the regions of the Eastern European Countries have performed, generally speaking, better than expected by their economic performance. This is clearly the case of the Baltic countries and, even more, the case of the former DDR.

From this, we can also see different models of social progress in Europe. Former cluster 1 is divided into a more industrial oriented approach or a more green and

educational approach. Former cluster 2 is divided into more models, where each one prioritises different aspects of social progress such as health for clusters 5 and 9, education for cluster 8, or industry and economic activity for clusters 4 and 6.

Moreover, the results of further subdividing into 21 clusters can be seen in Figure 6. For ease of visualising the map, the cluster borders are marked with a black outline while national borders are marked with a red outline. In this map, clusters almost always respect national borders, with few exceptions between Romania and Bulgaria, the Baltic countries and the Nordic countries.

Figure 6. NUTS-2 Map at 21 Clusters



Additionally, to better understand the clusters, in Table 10 it can be seen, for each cluster, the mean per capita GDP and the mean EU-SPI level. The richest cluster is cluster 19, comprised exclusively by Ireland. On the other hand, the poorest cluster is number 14, comprised entirely by Greek regions. By EU-SPI, the highest mean value is in cluster 12, in the Nordic regions; while the lowest mean value can be found in cluster 5, in Romania and Bulgaria.

Table 10. Mean GDPpc and EU-SPI values by cluster at 21 clusters

Cluster	Mean GDPpc	Mean EU-SPI
1	36,025.00	73.88641
2	42,900.00	70.68821
3	31,650.00	69.34763
4	27,425.00	69.81758
5	15,500.00	47.42561
6	29,475.00	55.43730
7	32,428.57	62.01472
8	26,725.00	67.87469
9	17,481.82	57.59260
10	35,196.97	71.68054
11	40,500.00	78.07988
12	35,017.65	82.38205
13	26,700.00	66.22437
14	15,036.36	55.38756
15	21,355.56	64.39801
16	26,155.56	72.85288
17	19,256.25	60.38989
18	35,050.00	63.66885
19	55,166.67	75.28185
20	19,737.50	56.43970
21	22,680.00	64.51266

Of the 21 clusters, seven are one-country clusters. Most notably, cluster ten comprises most parts of Germany, cluster 15 all Austria except for its capital, and cluster 16 almost all France. On the other end, cluster 2 is present in 7 countries. It contains the regions for Brussels, Lisbon, Ljubljana, Madrid, Marseilles, Paris, Prague and Vienna. Berlin, one of

the most important capital regions that is omitted in cluster 2, is in cluster 3 with several Belgian regions and a few other German regions. Overall, we can see the effect of being the capital regions. Almost all capital regions in Europe are separated from the main cluster in the country. This suggests that the capital regions have a distinct social progress level than the rest of its country. Even more, for the capitals in cluster 2, it suggests that they are more similar between each other than with any other region in their respective country. For the southern region of Europe, it is worth noting that both Spain and Italy have two clusters dividing most part of their countries in two halves. This kind of division within the country is unique to those two countries. Noticeably, Catalonia that used to be in the wealthiest advanced part of Spain is now on a second tier in social progress. On the contrary, some countries are quite internally homogeneous according to their SPI clustering, even at 21 clusters. This is the case of each of the Nordic countries, the Netherlands, Ireland, Croatia and Lithuania

3.3. Clustering per capita GDP and EU-SPI

Lastly, we will check the impact of per capita GDP levels on the SPI clusters. The EU-SPI has a clear objective of presenting an alternative to GDP, so seeing if the SPI clusters performed differ from the GDP levels would be validating that this goal has been achieved. If, on the other hand, the GDP is proved to be an explanatory factor of SPI cluster membership it would mean that the clusters represent not only social but also economic factors.

The GDP will be measured as GDP per capita at purchasing power parity (PPP). Although the per capita GDP levels are estimated at NUTS-2, the PPP index is only calculated at the national level (NUTS-1) by Eurostat.

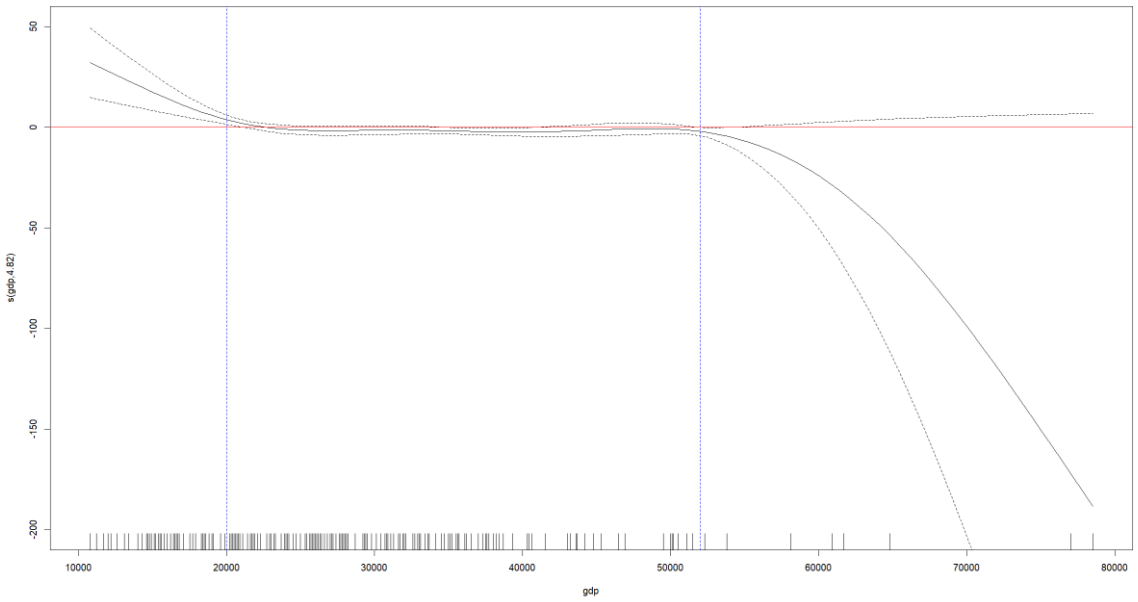
To understand the relationship between per capita GDP and cluster membership, several regression models were computed. They all take the per capita GDP as a predictor and cluster membership as the dependent variable. Cluster membership is treated as a dichotomous variable with values 1 or 2, and all the models try to estimate the probability of being in cluster 2 opposite to being in cluster 1 (the reference level).

The models computed were, firstly a linear model, secondly a logarithmic model, thirdly a quadratic model and lastly several General Additive Models (GAM). The GAMs model are non-linear and non-parametric. The benefits of using GAM, therefore, are that we can see more complex relationships between the GDP and cluster membership. At the same time, being non-parametric means that they don't assume any formula making the model more flexible to the data structure. To test the model fit to the original data the BIC index was considered in an attempt to minimise it. As explained before, the BIC index takes into account the global fit while penalising for complexity in the model, which is crucial when using GAM models.

Of all the linear models, the logarithmic model was the one with the best BIC (216.45). The GAM model differed in the amount of basis functions (which allows for non-linearity) for the smooth term of GDP. A higher amount of basis functions allows for more non-linearity at the expense of potential overadjustment to the data. The numbers tried were 3, 4, 6, 9 and 10. Of these, 6 had a lower BIC index with a value of 194.5, lower than the logarithmic function. For that, the GAM model with a maximum of six basis functions was used.

The results of the model can be seen graphically in Figure 7. The y-axis represents the probability of being in cluster 2 with the reference category being cluster 1. The x-axis represents the per capita GDP. The solid black line is the expected probabilities from the GAM model. The dashed black lines represent the confidence interval at 95%.

Figure 7. GAM expected probabilities of being in Cluster 2 by GDP level

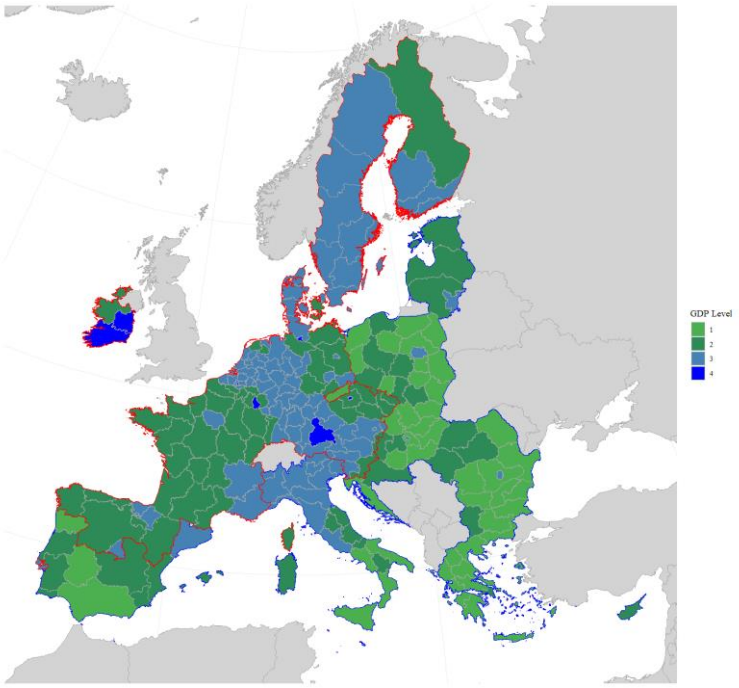


As it can be seen, there are three distinct regions in the graph. A first region, for the poorer regions below 20000€ in GDP per capita, are associated with pertaining to cluster 2. On the contrary, richer regions above 52000€ per capita are associated with cluster 1. There are 49 regions on the poorest end of the graph and 7 regions on the richest end. The other 175 are in the middle region, where per capita GDP does not seem to have any effect on cluster membership.

The effects of this can be seen in Figure 8. This map has been divided into four per capita GDP levels. The first level, in light green, represent the poorer regions below 20000€. According to the GAM model, these are the regions associated to cluster 2 based only on per capita GDP. The fourth level represents the richest regions above 52000€. These are the regions associated to cluster 1 from their per capita GDP. The other two levels (2 and 3) are not associated with any cluster based on per capita GDP. They differ in that regions in level 2 have a GDP per capita below the European mean while regions in level 3 have a per capita GDP above the European mean. To compare with cluster

membership, the borders of each cluster can be seen, the red outline being cluster 1 and the blue outline being cluster 2.

Figure 8. NUTS-2 Map with per capita GDP levels and cluster borders



Dividing the map into four levels instead of three (as the GAM did) provides us with another insight. By separating the regions below the European mean (in green) and above the European mean (in blue), one can see what the two clusters would look like if it only accounted for per capita GDP. As such, it is clear that the clusters computed using the EU-SPI differ greatly in some regions. For example, most of the regions in France are below the per capita GDP EU average, although all those regions are assigned to SPI cluster 1. This suggests a good territorial cohesion and high SPI performance. Public policy perhaps deserves some merit. At the other end we find Italy. All northern regions are above the per capita GDP European average. However, they are assigned to cluster 2. This suggests just the contrary: public policy has failed to transform high per capita GDP into good SPI performance. The permanence of internal economic and social cleavages stresses that the Northern failures do not profit the Southern regions: public policy might reasonably be under scrutiny. Other success stories are the former DDR NUTS-2, that

enjoy SPI levels close enough to those of the former FDR, although the per capita GDP levels are smaller. Also, the Czech Republic NUTS-2 regions are performing in SPI terms clearly better than in per capita GDP. Looking at Spain, it is worth noticing the remarkable performance of several Northern Peninsula NUTS-2 regions, that manage to be in the top SPI cluster although having a below the average per capita GDP. Just the contrary happens with the Catalonia case. This case seems quite different to Northern Italy as its failure to translate economic prosperity into SPI performance seems to be in benefit of other regions and not a pure inefficiency loss.

After all, it seems that GDP has mixed effects on social progress throughout Europe. While on some regions (those in level 1 and 4) it does have an effect, for most regions the economic level can not be said to either benefit or impede the social level of that region.

4. Conclusions

With the clustering analysis, the first hypothesis can be at least partially confirmed. It is true that there are distinct geographical regions with distinct social progress levels. It has been shown that these geographical regions are the southeast and northwest of the European Union. However, it cannot be said that these patterns originate at the country level. On the contrary, it is the macro regions represented by the two clusters the ones which better separate Europe in social progress. Nonetheless, within these two macro regions, country borders were mostly preserved.

As for the second hypothesis, it has also been partially proved. Social factors cannot account for everything regarding cluster membership and social progress levels. Geographically close regions tend to be more similar to each other, especially if they are within the same country. The exception to this norm is for capital regions. It has

consistently been found that capital regions are more similar to each other than to other regions in their respective country. As such, geographical factors can be said to play a role in cluster formation. It is hard to say if this stems from political factors or other variables not studied here.

Additionally, economic factors do not seem to have any effect for most of the regions. Per capita GDP levels only have an impact to the top and bottom performers, but not to the majority of regions. As such, hypothesis 2 has to be partially but not fully discarded, although it is true that social aspects do play a major role in cluster formation, other aspects like economic, geographical or political variables can have a great impact on social progress level depending on the specific context of each region.

Furthermore, some key variables were found to be exceptionally relevant to separating the two clusters and shed light on the nature of this macro regions. Among the EU-SPI components, *Shelter* was found to be the most prominent overall. Among the different latent variables detected, Eco-social and Institutional Quality was found to be the most relevant. It has also been shown, however, that the key components that determine cluster membership and social progress can change over specific clusters and specific regions. Also, with the insights from the partial dependence plots, it can also be stated that for each component there are critical values in which minimal changes will have the maximum impact on cluster membership and therefore on social progress.

Additionally, the subdivision of two clusters into 9 was useful in finding different models of driving social progress in each region. The nine resulting clusters try to prioritise different variables of social progress, such as health, education, or the labour market, with various degree of success.

In conclusion, EU-SPI has been shown to provide for an alternative and non-economic perspective on societal progress. The clusters also show this complementary view and

don’t respond to the economic levels. A more in-depth analysis of the results presented in this paper is needed if the aim is to use this as a guide for public policy making or expenditure management for a concrete region. Through an extensive review of the data published here, looking at the specific values of the components and factors for any region and analysing where they fall along the partial dependence plots can be a starting step to detect any potential tipping points from cluster 1 to cluster 2 or vice-versa, and to drive public policy data based decisions.

5. Bibliography

Alhija, F. A. N. (2010). Factor Analysis: An Overview and Some Contemporary Advances. *International Encyclopedia of Education, 162–170.* doi:10.1016/b978-0-08-044894-7.01328-2

Annoni, P., & Bolsi, P. (2020). The regional dimension of social progress in Europe – presenting the new EU Social Progress Index. European Commission, Directorate-General for Regional and Urban Policy, *Publications Office of the European Union*. https://doi.org/10. 2776/288544

Annoni, P., & Scioni, M. (2022). The unbalance penalisation method for metrics of social progress. *Social Indicators Research*, *162*, 1093–1115. https://doi.org/10.1007/s11205-021-02876-4

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256. https: //doi.org/10.1016/j.patcog.2012.07.021

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *160*(901), 268–282. Retrieved October 15, 2024, from http://www.jstor.org/stable/96803

Beltran-Esteve, M., Peiró-Palomino, J., Picazo-Tadeo, A. J., & Ríos, V. (2023). Is the European Social Progress Index robust? Implications for the design of European Union regional Cohesion policy. *Regional Studies*, *57*(11), 2285–2306. https://doi.org/10.1080/00343404. 2022.2159022

Breiman, L. Random Forests. *Machine Learning, 45*, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research, 21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Chouikhi, H., Charrad, M., & Ghazzali, N. (2015). A comparison study of clustering validity indices. *2015 Global Summit on Computer & Information Technology (GSCIT)*, 1–4. https://doi.org/10.1109/GSCIT.2015.7353330

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, *11*(4), 235–254. https://doi.org/https://doi.org/10.1016/0031-3203(79)90034-7

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31* (14), 2225-2236. https://doi.org/10.1016/J.PATREC.2010.03.014.

Huberty, C. (1975). Discriminant Analysis. *Review of Educational Research, 45* (4), 543 - 598. https://doi.org/10.3102/00346543045004543.

Kaiser, H. F. (1970). A second generation little iffy. *Psychometrika*, *35*, 401–415. https://doi. org/10.1007/BF02291817

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*, 159–179. https://doi.org/10.1007/ BF02294245

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https: //doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Ann. Statist. 6* (2) 461 - 464. https://doi.org/10.1214/aos/1176344136

Siddique, A., Waseem, A., & Mamoon, D. (2017). Did we find alternate to GDP to measure national progress? Analysis of Harvard University's social progress index. *Turkish Economic Review*, *4*, 352–368. https://api.semanticscholar.org/CorpusID:56253185

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *TAXON*, *11*(2), 33–40. https://doi.org/10.2307/1217208

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 777–781. https://doi.org/10.1080/10705511.2016.1217487

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

6. Appendices

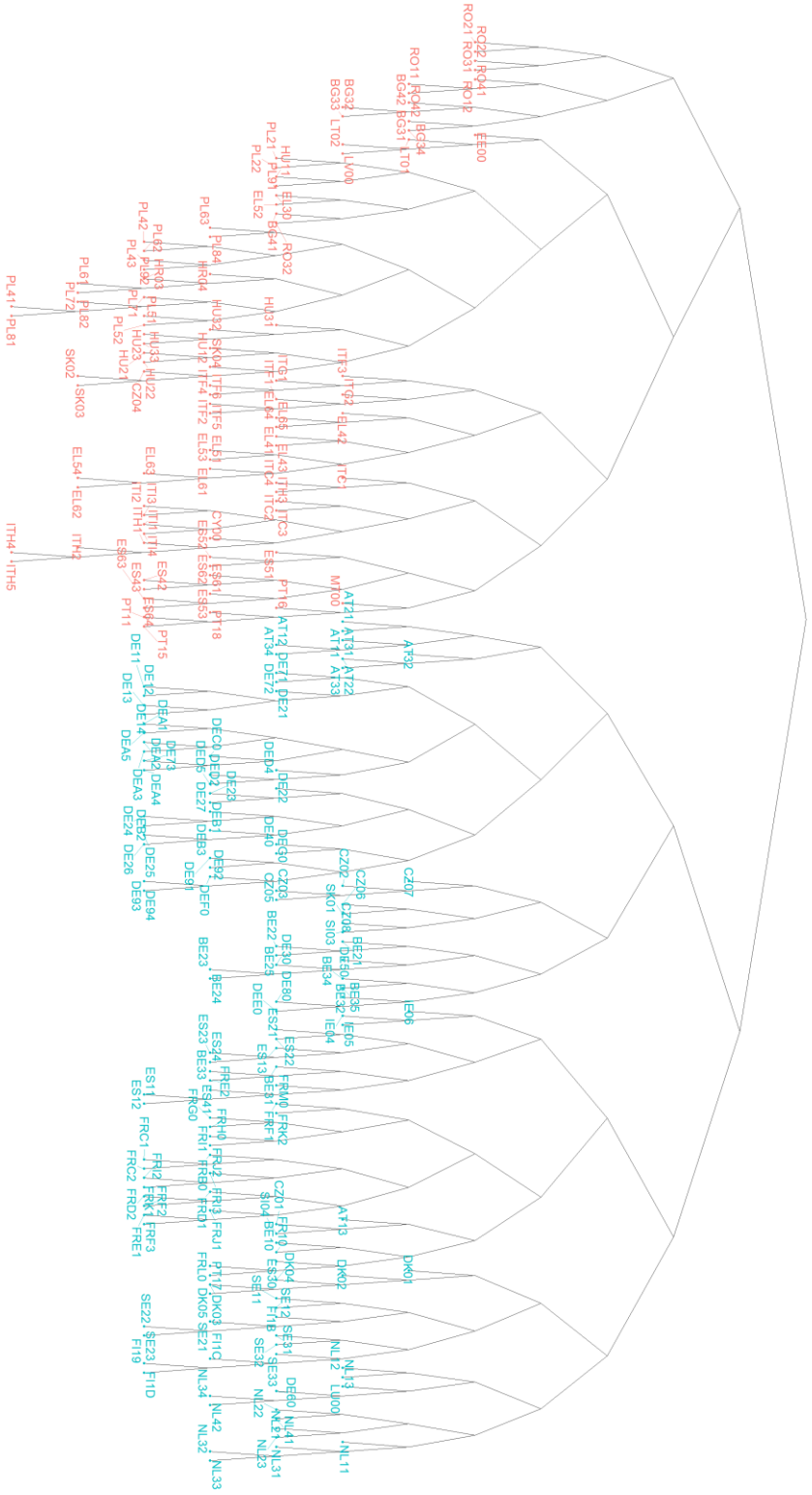
Annex 1. Dimensions, components and indicators within the EU-SPI

Dimension	Component	Indicator
Basic Human Needs	Nutrition and Basic Medical Care	Premature mortality (<65)
	Nutrition and Basic Medical Care	Infant mortality
	Nutrition and Basic Medical Care	Unmet medical needs
	Nutrition and Basic Medical Care	Insufficient food
	Water and Sanitation	Satisfaction with water quality
	Water and Sanitation	Lack of toilet in dwelling
	Water and Sanitation	Uncollected sewage
	Water and Sanitation	Sewage treatment
	Shelter	Burdensome cost of housing
	Shelter	Housing quality-dampness
	Shelter	Overcrowding
	Shelter	Lack of adequate heating
	Personal Security	Crime
	Personal Security	Safety at night
	Personal Security	Money stolen
	Personal Security	Assaulted/Mugged
Foundations of Wellbeing	Access to basic knowledge	Upper-secondary enrolment rate (age 14-18)
	Access to basic knowledge	Lower-secondary completion only
	Access to basic knowledge	Early school leavers
	Access to ICT	Internet at home
	Access to ICT	Broadband at home

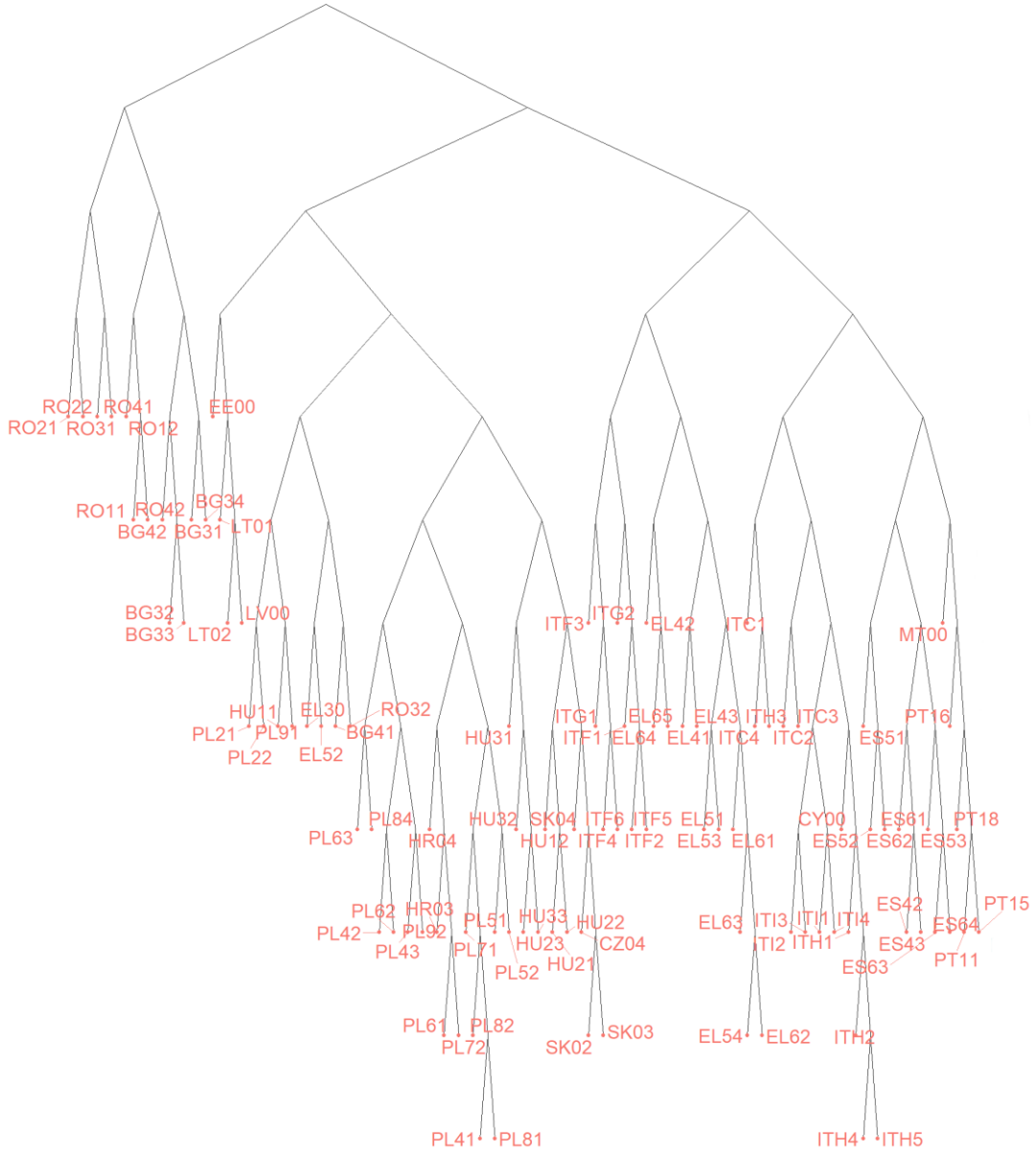
	Access to ICT	Online interaction with public authorities
	Access to ICT	Internet access
	Health and Wellness	Life expectancy
	Health and Wellness	Subjective health status
	Health and Wellness	Standardised cancer death rate
	Health and Wellness	Standardised heart diseas death rate
	Health and Wellness	Leisure activities
	Health and Wellness	Traffic deaths
	Environmental Quality	Air pollution NO2
	Environmental Quality	Air pollution Ozone
	Environmental Quality	Air pollution pm2.5
	Environmental Quality	Air pollution pm10
Opportunity	Personal Rights	Trust in the national government
	Personal Rights	Trust in the legal system
	Personal Rights	Trust in the police
	Personal Rights	Active citizenship
	Personal Rights	Female participation in regional assemblies
	Personal Rights	Institution quality index
	Personal Freedom and Choice	Freedom over life choices
	Personal Freedom and Choice	Job opportunities
	Personal Freedom and Choice	Involuntary part-time/temporary employment
	Personal Freedom and Choice	Young people, not in education, employment or training (NEET)
	Personal Freedom and Choice	Institutions corruption index
	Tolerance and Inclusion	Institution impartiality Index
	Tolerance and Inclusion	Tolerance towards immigrants
	Tolerance and Inclusion	Tolerance towards minorities
	Tolerance and Inclusion	Tolerance towards homosexuals

Tolerance and Inclusion	Making friends
Tolerance and Inclusion	Volunteering
Tolerance and Inclusion	Gender employment gap
Access to Advanced Education	Tertiary education attainment
Access to Advanced Education	Tertiary enrolment
Access to Advanced Education	Lifelong learning
Access to Advanced Education	Lifelong learning - female

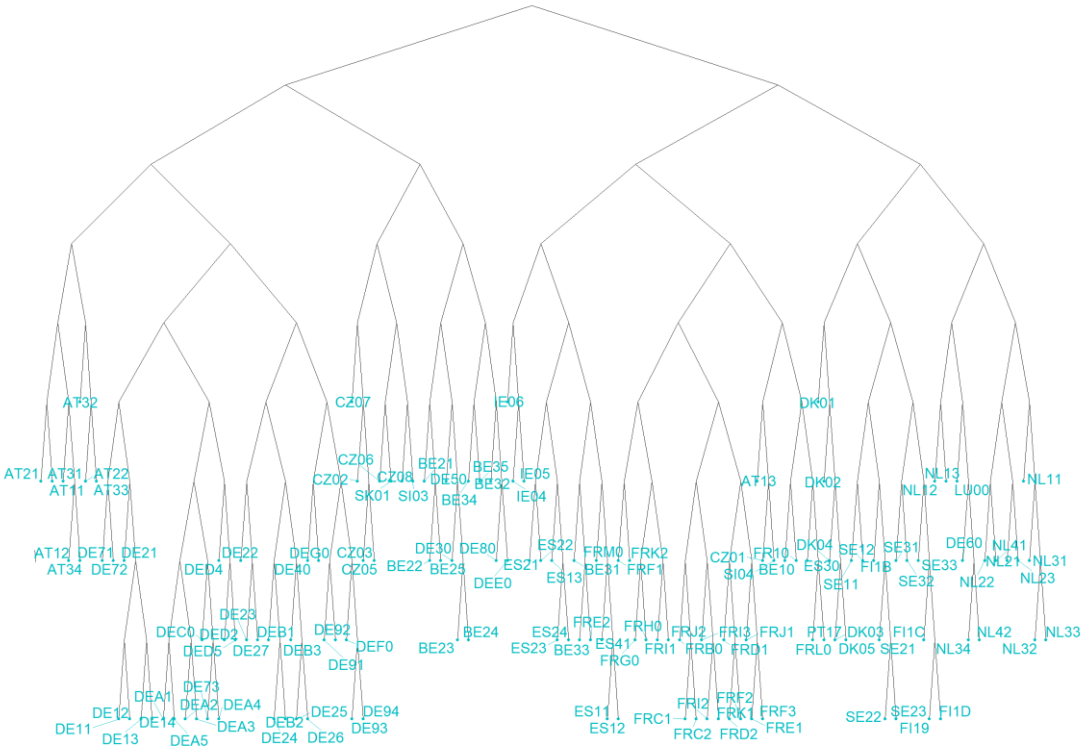
Appendix 2. Dendrogram of the clustering model coloured by cluster membership at 2 clusters



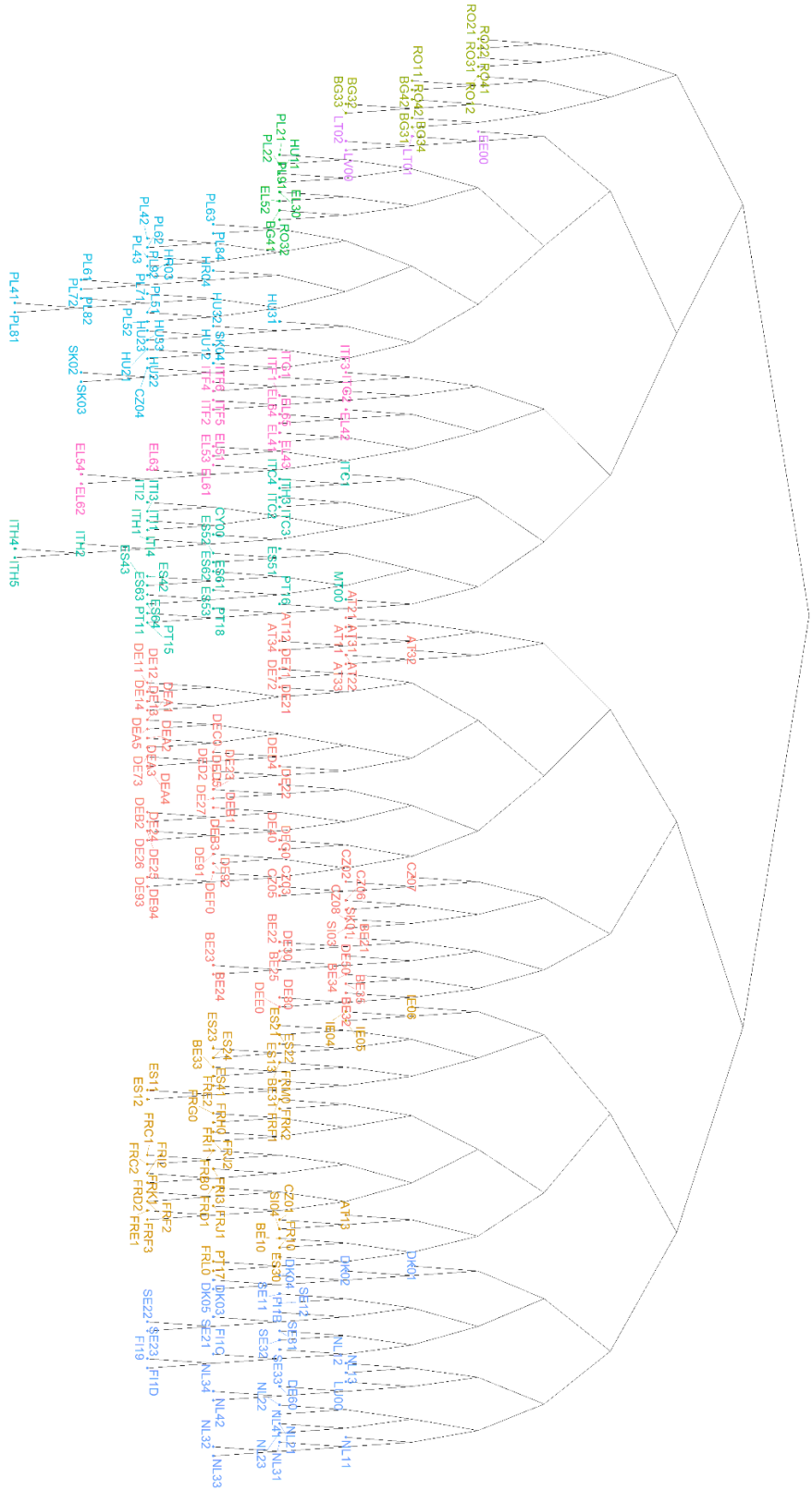
Appendix 3. Dendrogram of the clustering model coloured by cluster membership at 2 clusters, only cluster 1 is shown



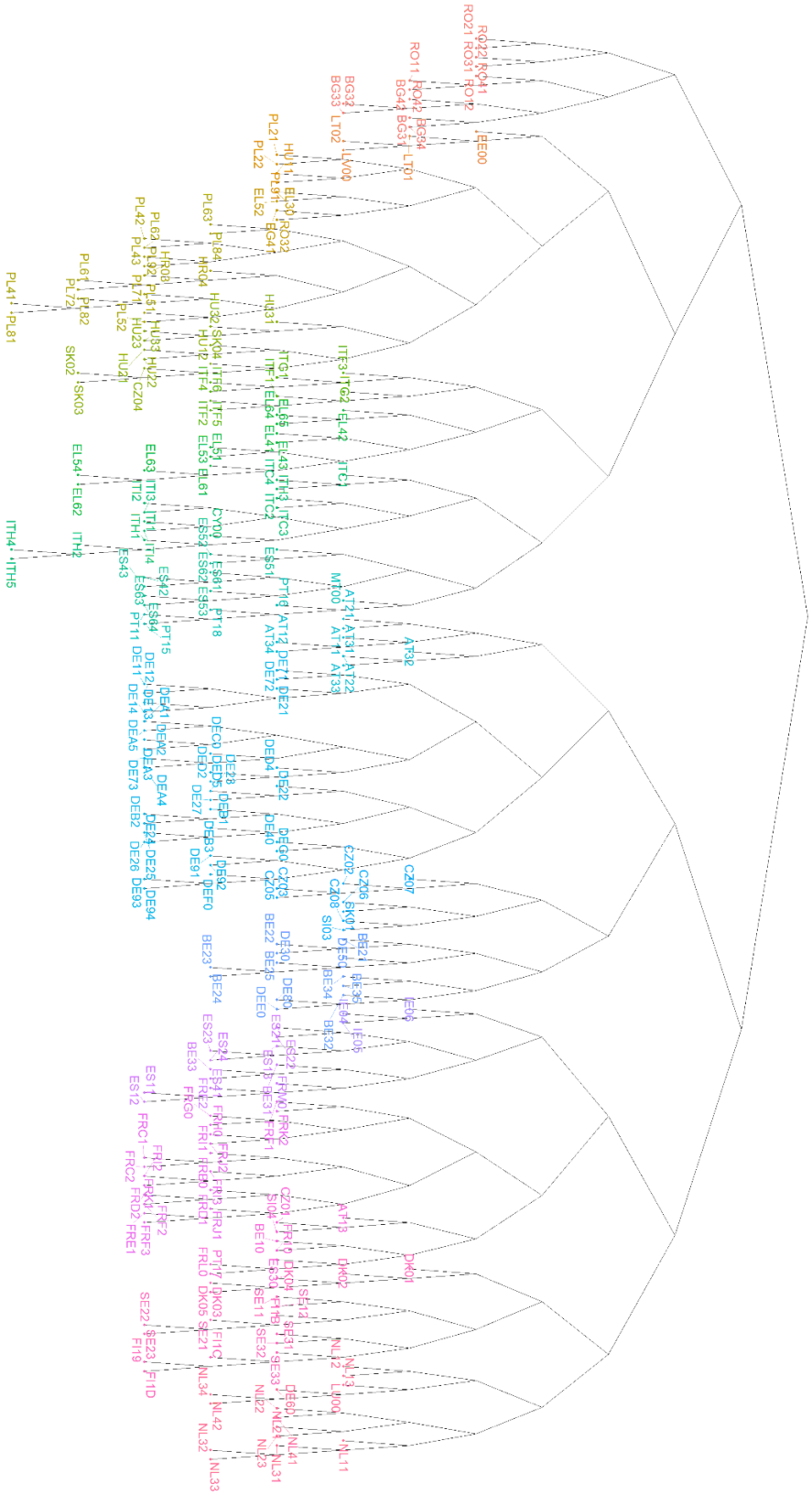
Appendix 4. Dendrogram of the clustering model coloured by cluster membership at 2 clusters, only cluster 2 is shown



Appendix 5. Dendrogram of the clustering model coloured by cluster membership at 9 clusters



Appendix 6. Dendrogram of the clustering model coloured by cluster membership at 21 clusters



Appendix 7. NUTS-2 ID to region name conversion

*Each region name is followed by three numbers that represent cluster membership when there are 2 clusters, when there are 9 clusters and when there are 21 clusters

NUTS ID	Region name*
AT11	Burgenland (1, 1, 1)
AT12	Niederösterreich (1, 1, 1)
AT13	Wien (1, 2, 2)
AT21	Kärnten (1, 1, 1)
AT22	Steiermark (1, 1, 1)
AT31	Oberösterreich (1, 1, 1)
AT32	Salzburg (1, 1, 1)
AT33	Tirol (1, 1, 1)
AT34	Vorarlberg (1, 1, 1)
BE10	Rég. de Bruxelles-Cap./Brussels Hfst. Gew. (1, 2, 2)
BE21	Antwerpen (1, 1, 3)
BE22	Limburg (1, 1, 3)
BE23	Oost-Vlaanderen (1, 1, 3)
BE24	Vlaams-Brabant (1, 1, 3)
BE25	West-Vlaanderen (1, 1, 3)
BE31	Brabant Wallon (1, 2, 4)
BE32	Hainaut (1, 1, 3)
BE33	Liège (1, 2, 4)
BE34	Luxembourg (1, 1, 3)
BE35	Namur (1, 1, 3)
BG31	Severozapaden (2, 3, 5)
BG32	Severen tsentralen (2, 3, 5)
BG33	Severoiztochen (2, 3, 5)
BG34	Yugoiztochen (2, 3, 5)
BG41	Yugozapaden (2, 4, 6)
BG42	Yuzhen tsentralen (2, 3, 5)
CY00	Kýpros (2, 5, 7)
CZ01	Praha (1, 2, 2)
CZ02	Střední Čechy (1, 1, 8)
CZ03	Jihozápad (1, 1, 8)
CZ04	Severozápad (2, 6, 9)
CZ05	Severovýchod (1, 1, 8)
CZ06	Jihovýchod (1, 1, 8)
CZ07	Střední Morava (1, 1, 8)
CZ08	Moravskoslezsko (1, 1, 8)
DE11	Stuttgart (1, 1, 10)
DE12	Karlsruhe (1, 1, 10)
DE13	Freiburg (1, 1, 10)
DE14	Tübingen (1, 1, 10)
DE21	Oberbayern (1, 1, 10)
DE22	Niederbayern (1, 1, 10)
DE23	Oberpfalz (1, 1, 10)
DE24	Oberfranken (1, 1, 10)
DE25	Mittelfranken (1, 1, 10)
DE26	Unterfranken (1, 1, 10)
DE27	Schwaben (1, 1, 10)

NUTS ID	Region name*
DE30	Berlin (1, 1, 3)
DE40	Brandenburg (1, 1, 10)
DE50	Bremen (1, 1, 3)
DE60	Hamburg (1, 7, 11)
DE71	Darmstadt (1, 1, 10)
DE72	Gießen (1, 1, 10)
DE73	Kassel (1, 1, 10)
DE80	Mecklenburg-Vorpommern (1, 1, 3)
DE91	Braunschweig (1, 1, 10)
DE92	Hannover (1, 1, 10)
DE93	Lüneburg (1, 1, 10)
DE94	Weser-Ems (1, 1, 10)
DEA1	Düsseldorf (1, 1, 10)
DEA2	Köln (1, 1, 10)
DEA3	Münster (1, 1, 10)
DEA4	Detmold (1, 1, 10)
DEA5	Arnsberg (1, 1, 10)
DEB1	Koblenz (1, 1, 10)
DEB2	Trier (1, 1, 10)
DEB3	Rheinhessen-Pfalz (1, 1, 10)
DEC0	Saarland (1, 1, 10)
DED2	Dresden (1, 1, 10)
DED4	Chemnitz (1, 1, 10)
DED5	Leipzig (1, 1, 10)
DEE0	Sachsen-Anhalt (1, 1, 3)

NUTS ID	Region name*
DEF0	Schleswig-Holstein (1, 1, 10)
DEG0	Thüringen (1, 1, 10)
DK01	Hovedstaden (1, 7, 12)
DK02	Sjælland (1, 7, 12)
DK03	Syddanmark (1, 7, 12)
DK04	Midtjylland (1, 7, 12)
DK05	Nordjylland (1, 7, 12)
EE00	Eesti (2, 8, 13)
EL30	Attiki (2, 4, 6)
EL41	Voreio Aigaio (2, 9, 14)
EL42	Notio Aigaio (2, 9, 14)
EL43	Kriti (2, 9, 14)
EL51	Anatoliki Makedonia, Thraki (2, 9, 14)
EL52	Kentriki Makedonia (2, 4, 6)
EL53	Dytiki Makedonia (2, 9, 14)
EL54	Ipeiros (2, 9, 14)
EL61	Thessalia (2, 9, 14)
EL62	Ionia Nisia (2, 9, 14)
EL63	Dytiki Ellada (2, 9, 14)
EL64	Sterea Ellada (2, 9, 14)
EL65	Peloponnisos (2, 9, 14)
ES11	Galicia (1, 2, 4)
ES12	Principado de Asturias (1, 2, 4)
ES13	Cantabria (1, 2, 4)

NUTS ID	Region name*
ES21	País Vasco (1, 2, 4)
ES22	Comunidad Foral de Navarra (1, 2, 4)
ES23	La Rioja (1, 2, 4)
ES24	Aragón (1, 2, 4)
ES30	Comunidad de Madrid (1, 2, 2)
ES41	Castilla y León (1, 2, 4)
ES42	Castilla-La Mancha (2, 5, 15)
ES43	Extremadura (2, 5, 15)
ES51	Cataluña (2, 5, 15)
ES52	Comunidad Valenciana (2, 5, 15)
ES53	Illes Balears (2, 5, 15)
ES61	Andalucía (2, 5, 15)
ES62	Región de Murcia (2, 5, 15)
ES63	Ciudad Autónoma de Ceuta (2, 5, 15)
ES64	Ciudad Autónoma de Melilla (2, 5, 15)
FI19	Länsi-Suomi (1, 7, 12)
FI1B	Helsinki-Uusimaa (1, 7, 12)
FI1C	Etelä-Suomi (1, 7, 12)
FI1D	Pohjois- ja Itä-Suomi (1, 7, 12)
FR10	Île de France (1, 2, 2)
FRB0	Centre - Val de Loire (1, 2, 16)
FRC1	Bourgogne (1, 2, 16)
FRC2	Franche-Comté (1, 2, 16)
FRD1	Basse-Normandie (1, 2, 16)
FRD2	Haute-Normandie (1, 2, 16)
FRE1	Nord-Pas de Calais (1, 2, 16)
FRE2	Picardie (1, 2, 4)
FRF1	Alsace (1, 2, 16)
FRF2	Champagne-Ardenne (1, 2, 16)
FRF3	Lorraine (1, 2, 16)
FRG0	Pays de la Loire (1, 2, 16)
FRH0	Bretagne (1, 2, 16)
FRI1	Aquitaine (1, 2, 16)
FRI2	Limousin (1, 2, 16)
FRI3	Poitou-Charentes (1, 2, 16)
FRJ1	Languedoc-Roussillon (1, 2, 16)
FRJ2	Midi-Pyrénées (1, 2, 16)
FRK1	Auvergne (1, 2, 16)
FRK2	Rhône-Alpes (1, 2, 16)
FRL0	Provence-Alpes-Côte d’Azur (1, 2, 2)
FRM0	Corse (1, 2, 4)
HR03	Jadranska Hrvatska (2, 6, 17)

NUTS ID	Region name*
HR04	Kontinentalna Hrvatska (2, 6, 17)
HU11	Budapest (2, 4, 18)
HU12	Pest (2, 6, 9)
HU21	Közép-Dunántúl (2, 6, 9)
HU22	Nyugat-Dunántúl (2, 6, 9)
HU23	Dél-Dunántúl (2, 6, 9)
HU31	Észak-Magyarország (2, 6, 9)
HU32	Észak-Alföld (2, 6, 9)
HU33	Dél-Alföld (2, 6, 9)
IE04	Northern and Western (1, 2, 19)
IE05	Southern (1, 2, 19)
IE06	Eastern and Midland (1, 2, 19)
ITC1	Piemonte (2, 5, 7)
ITC2	Valle d'Aosta/Vallée d'Aoste (2, 5, 7)
ITC3	Liguria (2, 5, 7)
ITC4	Lombardia (2, 5, 7)
ITF1	Abruzzo (2, 9, 20)
ITF2	Molise (2, 9, 20)
ITF3	Campania (2, 9, 20)
ITF4	Puglia (2, 9, 20)
ITF5	Basilicata (2, 9, 20)
ITF6	Calabria (2, 9, 20)
ITG1	Sicilia (2, 9, 20)
ITG2	Sardegna (2, 9, 20)
ITH1	Prov. Autonoma di Bolzano/Bozen (2, 5, 7)
ITH2	Provincia Autonoma di Trento (2, 5, 7)
ITH3	Veneto (2, 5, 7)
ITH4	Friuli-Venezia Giulia (2, 5, 7)
ITH5	Emilia-Romagna (2, 5, 7)
ITI1	Toscana (2, 5, 7)
ITI2	Umbria (2, 5, 7)
ITI3	Marche (2, 5, 7)
ITI4	Lazio (2, 5, 7)
LT01	Sostinės regionas (2, 8, 13)
LT02	Vidurio ir vakarų Lietuvos regionas (2, 8, 13)
LU00	Luxembourg (1, 7, 11)
LV00	Latvija (2, 8, 13)
MT00	Malta (2, 5, 21)
NL11	Groningen (1, 7, 11)
NL12	Friesland (1, 7, 11)
NL13	Drenthe (1, 7, 11)
NL21	Overijssel (1, 7, 11)
NL22	Gelderland (1, 7, 11)
NL23	Flevoland (1, 7, 11)
NL31	Utrecht (1, 7, 11)
NL32	Noord-Holland (1, 7, 11)

NUTS ID	Region name*
NL33	Zuid-Holland (1, 7, 11)
NL34	Zeeland (1, 7, 11)
NL41	Noord-Brabant (1, 7, 11)
NL42	Limburg (1, 7, 11)
PL21	Małopolskie (2, 4, 18)
PL22	Śląskie (2, 4, 18)
PL41	Wielkopolskie (2, 6, 17)
PL42	Zachodniopomorskie (2, 6, 17)
PL43	Lubuskie (2, 6, 17)
PL51	Dolnośląskie (2, 6, 17)
PL52	Opolskie (2, 6, 17)
PL61	Kujawsko-pomorskie (2, 6, 17)
PL62	Warmińsko-mazurskie (2, 6, 17)
PL63	Pomorskie (2, 6, 17)
PL71	Łódzkie (2, 6, 17)
PL72	Świętokrzyskie (2, 6, 17)
PL81	Lubelskie (2, 6, 17)
PL82	Podkarpackie (2, 6, 17)
PL84	Podlaskie (2, 6, 17)
PL91	Warszawski stołeczny (2, 4, 18)
PL92	Mazowiecki regionalny (2, 6, 17)
PT11	Norte (2, 5, 21)
PT15	Algarve (2, 5, 21)
PT16	Centro (2, 5, 21)
PT17	Área Metr. de Lisboa (1, 2, 2)
PT18	Alentejo (2, 5, 21)
RO11	Nord-Vest (2, 3, 5)
RO12	Centru (2, 3, 5)
RO21	Nord-Est (2, 3, 5)
RO22	Sud-Est (2, 3, 5)
RO31	Sud - Muntenia (2, 3, 5)
RO32	Bucureşti - Ilfov (2, 4, 6)
RO41	Sud-Vest Oltenia (2, 3, 5)
RO42	Vest (2, 3, 5)
SE11	Stockholm (1, 7, 12)
SE12	Östra Mellansverige (1, 7, 12)
SE21	Småland med öarna (1, 7, 12)
SE22	Sydsverige (1, 7, 12)
SE23	Västsverige (1, 7, 12)
SE31	Norra Mellansverige (1, 7, 12)
SE32	Mellersta Norrland (1, 7, 12)
SE33	Övre Norrland (1, 7, 12)
SI03	Vzhodna Slovenija (1, 1, 8)
SI04	Zahodna Slovenija (1, 2, 2)
SK01	Bratislavský kraj (1, 1, 8)

NUTS ID	Region name*
SK02	Západné Slovensko (2, 6, 9)
SK03	Stredné Slovensko (2, 6, 9)
SK04	Východné Slovensko (2, 6, 9)

*Each region name is followed by three numbers that represent cluster membership when there are 2 clusters, when there are 9 clusters and when there are 21 clusters