Social Indicators Research, April 2020. https://doi.org/10.1007/s11205-020-02324-9.
preprint version

Multidimensional clustering of EU regions. A contribution to orient public policies in reducing regional disparities

Pasquale Pavone[1], Francesco Pagliacci[2],
Margherita Russo[3], Simone Righi[4] and Anna Giorgi[5]

[1] Institute of Economics. Department of Excellence, Economics and Management in the Era of Data Science (EMbeDS). School of Advanced Studies - Pisa, pasquale.pavone@santannapisa.it

[2] Dipartmento Territorio e Sistemi Agro-forestali, Università di Padova, Italy, and CAPP - Research Centre for the Analysis of Public Policies, francesco.pagliacci@unipd.it

[3] Department Economia Marco Biagi, Università di Modena e Reggio Emilia, Italy, and CAPP - Research Centre for the Analysis of Public Policies, margherita.russo@unimore.it

[4] Department of Computer Science, UCL, United Kingdom, and CAPP - Research Centre for the Analysis of Public Policies, s.righi@ucl.ac.uk

[5] Leader AG1 EUSALP Lombardy Region representative, and Gesdimont research centre, University of Milan, Milano, Italy, anna.giorgi@unimi.it

ABSTRACT

This paper applies multidimensional clustering of EU-28 regions with regard to their specialisation strategies and socioeconomic characteristics. It builds on an original dataset.

Several academic studies discuss the relevant issues to be addressed by innovation and regional development policies, but so far no systematic analysis has linked the different aspects of EU regions research and innovation strategies (RIS3) and their socio-economic characteristics. This paper intends to fill this gap, with the aim to provide clues for more effective regional and innovation policies.

In the data set analysed in this paper, the socioeconomic and demographic classification associates each region to one categorical variable (with 19 categories), while the classification of the RIS3 priorities clustering was performed separately on "descriptions" (21 Boolean categories) and "codes" (11 Boolean Categories) of regions' RIS3. The cluster analysis, implemented on the results of the correspondence analysis on the three sets of categories, returns 9 groups of regions that are similar in terms of priorities and socioeconomic characteristics. Each group has different characteristics that revolve mainly around the concepts of selectivity (group's ability to represent a category) and homogeneity (similarity in the group with respect to one category) with respect to the different classifications on which the analysis is based.

Policy implications showed in this paper are discussed as a contribution to the current debate on post-2020 European Cohesion Policy, which aims at orienting public policies toward the reduction of regional disparities and to the enhance complementarities and synergies within macro-regions.

KEYWORDS: regional smart research and innovation strategies, multi-dimensional analysis, clustering, European regions, sustainable development

JEL CODES: R58-Regional Development Planning and Policy; Q5-Environmental Economics Q58-Government Policy; C38-Classification Methods, Cluster Analysis, Principal Components, Factor Models

ACKNOWLEDGEMENTS. This work is part of the Work Package Nr: T-3 "Enhancing shared Alpine Governance project" of the Project "Implementing Alpine Governance Mechanism of the European Strategy for the Alpine Region" (AlpGov) of the Interreg Alpine Space Programme - Priority 4 (Well-Governed Alpine Space), SO4.1 (Increase the application of multilevel and transnational governance in the Alpine Space). A preliminary version of this paper has been presented at the workshop "Promoting open innovation in the EUSALP macro-region: experiences from the Alpine regions", organised by Action Group 1 in the Eusalp Annual Forum, 21st November 2018, Innsbruck, Austria. The authors wish to thank the participants and Mr. Jean-Pierre Halkin, DG Regional and Urban Policy – Head of Unit D.1, for their comments.

1. Introduction

The current debate on post-2020 European Cohesion Policy confirms the need for public policies targeting the reduction of regional disparities and the enhancement of complementarities and synergies within macro-regions. Such interventions, supported by the European Structural & Investment Funds, are key instruments for the implementation of EU policies and programmes, aimed at fostering the cohesion and competitiveness across larger EU spaces, encompassing neighbouring member and non-member States (European Commission, 2016)[1]. To this end, regions are encouraged to share their best practices, to learn from each other and to exploit the opportunities for joint actions, through dedicated tools created by the European Commission. A specific dimension of such leverages is the set of strategic priorities that regions have outlined in their smart specialisation on research and innovation. The concept stems from academic work on the key drivers for bottom-up policies aiming at structural changes that are needed to improve job opportunities and welfare of territories (Foray *et al*., 2009; Barca, 2009; Foray, 2018). In the programming period 2014-2020, the European Commission has adopted the Research and Innovation Smart Specialisation Strategy (RIS3) as an ex-ante conditionality for access of regions to European Regional Development Funds (ERDFs). Such policies are built on specific guidelines and on a very detailed process of implementation (European Commission 2012, 2017; Foray *et al*. 2012; McCann and Ortega, 2015). They identify "strategic areas for intervention, based both on the analysis of the strengths and potential of the regional economies and on a process of entrepreneurial discovery with wide stakeholder involvement. It embraces a broad view of innovation that goes beyond research-oriented and technology-based activities, and requires a sound intervention strategy supported by effective monitoring mechanisms" (European Commission, 2017, p. 11).

Although over 65 billion EUR of ERDFs have been allocated to such policies their impact has not been scrutinised yet and no effective monitoring tool has been implemented[2]. In addition, no systematic information on the list of projects implemented under the various regions' RIS3 priorities is available[3]. For regions aiming at learning from other regions' practices on RIS3, information on regional strategies and goals is shared through online platforms, such as the S3 platform run by EC-JRC. Other loci of interaction among regions are those supported by the EU Interreg programmes[4], the

1 Since 2009, four macro-regions have been implemented: EUSBSR, for the Baltic Sea Region (2009); EUSDR, for the Danube Region (2011); EUSAIR, for the Adriatic and Ionian Region (2014); EUSALP, for the Alpine Region (2015). They comprehensively involve 19 EU Member States and 8 non-EU countries, also with some territorial overlaps (European Commission, 2016).

2 "The long-term impact of implementation of smart specialisation strategies in terms of increased innovation, job creation and improved productivity will require a number of years and will be examined as part of the ongoing and ex-post evaluation of Cohesion Policy programmes" (European Commission, 2017, p. 19).

3 Gianelle *et al*. (2017) present a preliminary analysis on Italy and Poland, grounded on an expert classification of RIS3 priorities.

4 https://www.interregeurope.eu/

Interact Initiatives[5], and the macro-regions strategies[6]. National programmes too, provide fora to cross-region cross-country comparison of structural features and policy measures on diverse domains[7].

Several academic studies provide analytical frameworks to support public decision making on subject such as income disparities (Iammarino *et al*., 2018) or quality of institutions (Charron *et al*., 2014). However, no systematic analysis has linked jointly the different aspects of EU regions specialisation strategies and their socio-economic characteristics. This paper aims to fill this gap by applying a multidimensional clustering of EU-28 regions in order to provide clues for more effective regional policies. The clustering proposed in the paper builds on an original dataset, where the EU-28 regions are classified according to their socioeconomic features (Pagliacci *et al*., 2019), and to the strategic features of their research and innovation smart specialisations strategy (RIS3) (Pavone *et al*., 2019). In the first classification, each region is associated to one categorical variable (with 19 modalities) based on a multidimensional analysis (PCA and CA) of a large dataset, and it provides a perspective focused on regional heterogeneity across EU regions. In the second classification, two clustering of "descriptions" and "codes" of RIS3s' priorities were considered (respectively made of 21 and 11 Boolean categories). This comparative perspective is made possible by a non-supervised boolean textual classification of priorities using information on RIS3 from the Eye@RIS3 platform (European Commission – Joint Research Center JRC).

The paper is structured as follows. Section 2 describes the methods used to obtain a multidimensional classification and the dataset built on the classification of socioeconomic features of EU-28 regions and classification of priorities pointed out in their smart specialisation strategies. Section 3 returns the main results. Section 4 builds on the results of the analysis and discusses their implications for policy and possible future strands of this research.

2. Data and methods

The data analysed in this paper results from the merging of two main datasets[8]. First of all, we use the classification of regions according to their socioeconomic features of Pagliacci *et al*. (2019). A socio-economic categorical variable is defined classifying the 208 territorial entities in EU-28 regions in 19 categories. Secondly, with regard to smart specialisation strategies, we use the classification defined by Pavone *et al*. (2019). There, the RIS3 priorities of 216 EU-28 territorial entities are summarised in two multi-class categorical variables: *Description* (21 categories) and *Codes* (11 categories). These two categorisations derive from an automatic classification of the priorities specified by each region in terms of free text of descriptions and of codes, which belong to three domains:

[5] http://www.interact-eu.net/

[6] https://ec.europa.eu/regional_policy/it/policy/cooperation/macro-regional-strategies/

[7] Example of national fora is the FONA project, in Germany, on sustainable science, technology and innovation for a sustainable society (www.fona.de)

[8] Data are available online at http://hdl.handle.net/11380/1177861, doi: 10.25431/11380_1177861

scientific, economic, and policy objectives[9]. In the dataset, each record refers to a priority defined by the region with a free text description and with a series of codes in the three domains. Each region could specify one or more priorities. The automatic analysis of the two corpora (description and codes) has allowed the classification of priorities in 21 topics for descriptions and 11 groups for codes. The results of the three classifications can be cross-referenced by using the online tool created ad hoc for such cross-tabulation. Developed within the AlpGov project to map R&I in the Alpine regions, the tool is implemented to query the classifications of all the EU regions. Through an effective visualisation of maps and data[10], it allows policy makers, researchers and public to query specific combinations of interest, focusing on the most detailed identification of groups of regions along the three categorisations: of economic characteristics, and of RIS3' priorities descriptions and codes.

Merging the two datasets, in this paper we study the multidimensional classification of 191 territorial entities according to the three above mentioned categorical variables.

The state of the art in clustering is provided by a huge literature (Jain, 2010), developed in a variety of scientific fields with different languages and focusing on the most diverse problems: clustering heterogeneous data, definition of parameters and initialisations (such as the times of iterations in K-means, e.g., MacQueen, 1967) and the threshold in hierarchical clustering (Jain 1988), as well as the problem of defining the optimal number of groups. Research is increasingly focusing on combining multiple clustering of the same dataset to produce a better single one clustering (Boulis & Ostendorf, 2004).

Without going into the merits of what could be the best method of classification, we put forward a grouping of regions according to their similarity in terms of their socio-economic characteristics and their RIS3 priorities. This enable comparing policy strategies in EU by implementing a factor analysis and a cluster analysis, applied on the matrix *Regions* × *Categorical variables*. Given that our case study comprises only one univocal categorical variable (19 regions' socio-economic and demographic categories) and two multi-class categorical variables (*Codes* and *Descriptions* of regions' RIS3's priorities, respectively with 11 and 19 categories), we directly apply a Correspondence Analysis (Benzecri 1992, Greenacre 2007) to the Boolean matrix *Regions* × *Categories* (191×51), in which the totals of rows depends on the number of categories in which each region has been classified. Usually, a matrix Units × Categorical variables (univocal classification) is studied through a multiple correspondences analysis that transforms the matrix Units × Variables (m×s) into a Boolean matrix Units × Categories (m×n). This latter matrix is considered as a particular frequency table which has the total of rows equal to the number of categorical variables considered in the analysis, while the total of columns is equal to the frequency of each category in the *m* units considered (Bolasco, 1999). Then a correspondence analysis is applied, after transforming the Boolean data into row and column profiles, looking for their reproduction in factorial subspaces

[9] Dataset downloaded on 1st October 2018 from Eye@RIS3 platform, EC-JRC.

[10] Available at https://www.alpine-region.eu/actions/mapping-eusalp-regions-governance-concerning-ri-sector

according to the criterion of the best orthogonal projections. In the present analysis, given a multiple categorization in two out of three dimensions, we adopt a Correspondence Analysis on the Boolean matrix. The factors highlight the configuration of the profiles in a graphical context. The interpretation of each factor through the analysis of the nodes' polarization sheds light on the association structure among regions' profiles[11]. Then a hierarchical agglomerative clustering based on Ward's aggregation method, with Euclidean distance, is applied on the results of the Correspondence Analysis on the dataset of regions.

3. Results

The correspondence analysis is applied to the Boolean matrix *Regions* × *Categories*. In this matrix, each region is classified according to a socio-economic class and to the set of categories of codes and categories of descriptions. Results of such an analysis are presented in Figure 1 and Figure 2, with regard to the distribution on *f1f2* plane, respectively, of the 51 categories and of the 191 regions. Annex 1 lists the coordinates of the categories on the first four factors: these figures allow to interpret the existing polarizations in each factor. Building on this information, by analysing Figure 1, we observe that the first factor polarises information on the specialisation of the regional economy, from services (left) to manufacturing (right), while the second factor polarises information on income, from low income (bottom) to high income (top). Figure 2 shows the distribution of the regions relative to the differences highlighted in Figure 1. Therefore, from left to right there are regions more characterised by the production of services vs. the production of goods, while from bottom to top there are regions characterised by a low income vs. a high income.

Figure 1 - Distribution on factorial plane *f1f2* of the 51 categories

Figure 2 - Distribution on the factorial plane *f1f2* of the 191 regions

In the clustering process applied to such results, each factor represents only a part of the overall set of information and different results can be obtained, according to the number of factors considered. The selection of the most appropriate number of factors can be derived by observing the boxplot of coordinates of regions in each factor[12]. Figure

[11] Among the planes generated by the pairs of factorial axes, the one identified by the first two has the most relevant share of the overall inertia and therefore reproduces with less distortion the actual distances between the points of the cloud.

[12] In general, in a correspondence analysis of a medium-large matrix, such as the one under analysis, the rate of inertia is always very low, then it allows the ranking of the factors but it is not very effective in guiding the selection of the number of factors to be considered for the clustering procedure. Histogram of the percentage of inertia of the first 50 factors is plotted in Annex 2.

3 presents the regions coordinates of the ten factors, they show different projections of the cloud of points and highlight outliers.

Figure 3 - Regions coordinates on the first ten factors

In particular, the 5th factor singles out only the difference between one region (in the case in this example, the Brussels region - BE01) and all the others. The same holds true for the 10th factor (in this case, the Luxembourg region - LU00). When five factors are considered, one single cluster results with only this outlier and, by increasing the number of factors under analysis, other outliers emerge as single clusters. Therefore, in order to avoid the influence of these outlier regions within the clustering process, without excluding them from the analysis, we proceed to carry out a cluster analysis considering, for the aggregation criteria, only the coordinates related to the first four factors. By analysing the resulting dendrogram[13] (Figure 4), nine groups of regions have been selected. According to the Calinski and Harabasz index, the optimal number of cluster is five, but in order to single out significant aggregations of regions in terms of dimensions that are relevant for our analysis we adopted a greater number of clusters. The choice of the 5 clusters, although optimal from a statistical point of view, leads to an excessively broad and not relevant aggregation with regard to the economic analysis. For example, with the 5-clusters classification we obtain a first cluster that represents 46% of the information and groups 45% of the regions: with regard to its characteristic features, this cluster has the same RIS3 priorities (Manufacturing, Agro-food and Sustainable Energy) associated to very heterogeneous socio-economic conditions. Therefore, the choice of the greater number of clusters aims at obtaining groups with more homogeneous socio-economic characteristics for the various priorities. We have adopted a classification in nine clusters that will be detailed below and summarised in the table embedded in figure 7.

Figure 4 – Dendrogram and Calinski and Harabasz index

Figures 5 and 6 show the distribution of regions and groups, respectively on the *f1f2* plane and *f3f4* plane.

Figure 5 - Distribution on *f1f2* plane of the 191 regions and nine partitions

legend: black dots: regions; yellow circles: clusters, with size proportional to their absolute weight

Figure 6 - Distribution on *f3f4* plane of the 191 regions and nine partitions

legend: black dots: regions; yellow circles: clusters, with size proportional to their absolute weight

For each of the nine clusters, Table 1 lists the characteristic categories, which are defined as those with a test-value greater than 2.1[14] (they are ranked in decreasing order

[13] For each group, the percentage values indicate its relative weight, in terms of the number of categories.

[14] Test-value for qualitative categorical variable is a statistical criterion associated with the comparison of two portions within the framework of a hypergeometric law. The test-value = 2.1 corresponds to a bilateral test probability $\alpha/2$ of less than 2.5%.

of their test-value, column 3). The weight of those categories, i.e. the number of times the category occurs in the dataset, is shown in absolute and relative terms, respectively in columns 4 and 5. The ratio of each category in the cluster to all categories in the cluster (columns 6) highlights the extent to which the category is characteristic.

Table 1 - Characteristic categories of the nine clusters of regions

cluster ID and label of characteristic categories	(1) # reg.s in the cluster	(2) ID of characteristic frequencies	(3) Test-value	(4) Weight in the dataset	(5) % of frequency in the dataset	(6) Ratio of category in the Cluster to all modes in the Cluster	selectivity (7) % of the category in the Cluster **SELECTIVITY**	homogeneity (8) % of regions with the category in the Cluster **HOMOGENEITY**
Cluster 1	31							
High-income; low-population density; tourism		SocEc-2	5.86	14	0.70	4.38	85.71	38.71
Sustainable Energy		Descr-23	2.41	108	5.36	8.76	22.22	77.42
						13.14		
Cluster 2	31							
Very low-income; manufacturing; no foreigners; highly educated		SocEc-1	6.13	18	0.89	4.66	83.33	48.39
Manufacturing		Descr-17	4.52	55	2.73	7.14	41.82	74.19
Agrofood		Descr-3	2.87	84	4.17	7.45	28.57	77.42
Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density		SocEc-6	2.65	3	0.15	0.93	100.00	9.68
Fashion		Descr-6	2.44	9	0.45	1.55	55.56	16.13
						21.74		
Cluster 3	25							
Medium-income; employm.&popul. imbalances; manufacturing: textile, basic metal, transport; very-low ed.		SocEc-9	2.49	12	0.60	1.85	50.00	24.00
Urban regions; high-income; poorer employment conditions; touristic		SocEc-7	2.43	9	0.45	1.54	55.56	20.00
						3.40		
Cluster 4	14							
Very-low income; agriculture; sparsely populated; very high unemployment; traditional services (G-I)		SocEc-11	5.14	13	0.65	6.61	61.54	57.14
Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very-low educated		SocEc-13	4.46	6	0.30	4.13	83.33	35.71
Tourism		Descr-8	4.42	59	2.93	11.57	23.73	100.00
Creative industry, Tourism & cultural and recreative services		COD-1	2.92	88	4.37	10.74	14.77	92.86
Agrofood		Descr-3	2.69	84	4.17	9.92	14.29	85.71
						42.98		
Cluster 5	14							
High-income; sparsely populated; public sector; highly educated		SocEc-3	5.37	31	1.54	10.43	38.71	85.71
Social innovation & education		COD-2	4.58	36	1.79	9.57	30.56	78.57
Growth & Welfare		Descr-12	4.45	25	1.24	7.83	36.00	64.29
Bioeconomy		Descr-11	3.62	45	2.23	8.70	22.22	71.43
						36.52		
Cluster 6	5							
Very-high income; large urban regions; high-employment; highly educated		SocEc-4	3.95	5	0.25	9.09	60.00	60.00
Growth & Welfare		Descr-12	3.24	25	1.24	12.12	16.00	80.00
Social innovation & education		COD-2	2.82	36	1.79	12.12	11.11	80.00
						33.33		
Cluster 7	18							
Marine & Maritime		Descr-20	3.12	31	1.54	4.65	32.26	55.56
						4.65		
Cluster 8	28							
High-income; high-employment; low-manufacturing; services & public sector		SocEc-15	5.93	24	1.19	5.43	70.83	60.71
Optics		Descr-13	3.75	5	0.25	1.60	100.00	17.86
Transport & Logistics		Descr-19	3.54	45	2.23	5.43	37.78	60.71
Energy Production		Descr-22	3.09	34	1.69	4.15	38.24	46.43
Transport & logistics		COD-9	2.66	52	2.58	5.11	30.77	57.14
						21.73		
CLUSTER 9	25							
Very-high income; manufacturing; population imbalances		SocEc-10	5.70	14	0.70	4.04	85.71	48.00
Healthy Food		Descr-4	5.52	17	0.84	4.38	76.47	52.00
ICT & Tourism		Descr-7	4.39	27	1.34	4.71	51.85	56.00
Life Science		Descr-2	2.82	57	2.83	5.72	29.82	68.00
Low-income; high-density; high unemployment; agriculture; food & drinks; very-low educated		SocEc-12	2.80	8	0.40	1.68	62.50	20.00
Aeronautics, Aerospace & Automotive industry		COD-10	2.36	26	1.29	3.03	34.62	36.00
						23.57		

We observe that not all the codes are characteristic categories associated to the nine clusters: by selecting categories according to their test-value we are focusing only on those presenting a value that is significantly above the average occurrence among the regions in the cluster.

In general, with regard to the three sets of categories under analysis, Table 1 returns that, in seven out of nine cases, the clusters are characterised by a mix of socio-economic categories and classes of priorities. In the case of cluster #3, there are only socio-economic aspects as characteristic categories (being the most barycentric cluster), while in cluster #7 there is only one priority as characteristic category: this happens because none of the other categories of the regions grouped in this cluster are - on average - significantly higher than the average of their occurrence in the whole dataset. The nine clusters are now described with regard to the selectivity/homogeneity of their characteristic categories. These two elements are of fundamental importance for understanding and interpreting each group. *Selectivity* represents the group's ability to represent a category. It indicates the percentage of category in the cluster compared to the entire dataset. *Homogeneity*, on the other hand, represents the similarity in the group with respect to one category, it indicates the percentage of regions with the same category in the cluster.

Cluster #1, encompassing 31 regions, is characterised by the socio economic class *High-income; low-population density; tourism* (with 85.71% occurrences in the cluster, which are associated to 38.71% of regions) and the description priority *Sustainable Energy* (77.42% of regions). The first characteristic category represents an element of selectivity of the category in the cluster, while the second one represents an element of homogeneity within the group.

Cluster #2 comprises 31 regions and it is characterised by two distinct socio-economic classes (both characterised by very low income), and description of priorities associated to *Manufacturing* (74.2% of regions), *Agrofood* (77.4% of regions) and *Fashion* (present at 55.6% in the cluster). Socio economic classes represent the selectivity features, while *Manufacturing* and *Agrofood* represent the homogeneity character of this group.

Cluster #3 encompasses 25 regions and the only distinctive element of this group are socioeconomic conditions: *Medium-income; employment & population imbalances; manufacturing: textile, basic metal, transport; very poorly educated* (present at 50% in the cluster and referred to 24% of regions) and *Urban regions; high-income; poorer employment conditions; touristic* (present at 55. 6% in the cluster and referred to 20% of regions): both characters show critical socioeconomic conditions.

Cluster #4 (with 14 regions) is characterised by regions with a low and very low income (respectively 83.3% and 61.5% of occurrences in the cluster, respectively referred to 35.7% and 57.1% of regions). The priorities' descriptions refer to *Tourism* (100% of regions), *Creative industry* (92.9% of regions) and *Agrofood* (85.79% of regions). Also in this case, the socio-economic conditions represent the selectivity features, while priorities' descriptions are the homogeneity character within the group.

Cluster #5, (with 14 regions), is characterised by the socio-economic class *High-income; sparsely populated; public sector; highly educated* (85.7% of regions) and priorities' descriptions referred to: *Social innovation & education* (78.6% of regions); *Growth & Welfare* (64.3% of regions); *Bio economy* (71.4% of regions). In this case all

the characteristic categories represent the homogeneity character linking the regions in this cluster.

Cluster #6, (with just 5 regions) differs from cluster #5 because of its socio-economic features, characterised by *Very-high income; large urban regions; high-employment; highly educated* (with 60% of occurrences in the cluster associated with three regions).

Cluster #7 encompasses 18 regions with just one characteristic category: i.e. the marine and maritime priority (55.6% of the regions); other categories associated to regions in the cluster are not significantly higher than the average of the whole dataset.

Cluster #8 comprises 28 regions and it is characterised by the socio economic class *High-income; high-employment; low-manufacturing; services & public sector* (with 70.83% occurrences in the cluster, referring to 60.7% of regions) and by the priority descriptions: *Optics* (with 100% occurrences in the cluster and referred to 17.9% of regions); *Transport & Logistics* (60.7% of regions); *Energy Production* (46.4% of regions). *Optics* represent a specific element, while the most homogeneous elements are the socio-economic class and *Transport & Logistics* description.

Cluster #9 is composed of 25 regions and it is characterised by two different socio-economic classes: *Very-high income; manufacturing; population imbalances* (with 85.71% occurrences in the cluster, referred to 48% of regions) and *Low-income; high-density; high unemployment; agriculture; food & drinks; very poorly educated* (62.5% of occurrences in the cluster, referred to 20% of regions). What unites regions with such different socioeconomic conditions is the set of characteristic categories of description: *Healthy Food* (present at 76.5% in the cluster and referred to 52% of regions); *ICT & Tourism* (present at 51.8% in the cluster and referred to 56% of regions); *Life Science* (68% of regions); *Aeronautics, Aerospace & Automotive industry* (36% of regions). Cluster 9 has as selectivity elements both socio-economic classes and *Healthy Food* priority, while there are no very high values of homogeneity (*Life Science*, referred to 68% of regions, is the highest value).

Figure 7 maps the nine clusters, with the table in the right panel summarising the homogeneity and selectivity elements characterising the nine set of clusters under analysis. It is clear from the map that the different clusters do not just capture geographical proximity, but rather the similarity in the status (socio-economic and demographics elements) and areas of specialization.

Figure 7 - Maps of clusters of regions, by socioeconomic features and RIS3s' priorities: summary of selectivity and homogeneity characteristic categories

4. Discussion and conclusions

In this paper, we aim at interpreting the overall framework of interconnected structural socioeconomic and demographic features and policy programmes on smart specialisation strategy in the EU. By identifying clusters of EU regions, we provide policy makers with a more systematic and informed tool they can use to learn from other regions, when they focus on the projects implemented within the various priorities.

Clustering of multidimensional categorisation is a multifaceted issue that must be addressed with the awareness that various methods of clustering are also affected by the data under analysis, such as: the overall number of observations, the number and type of variables (categorical, non-categorical and mixed variables, multiple vs single categorisations), the distribution of observation along the various dimensions under analysis, and missing data. In the analysis presented in this paper, we merge two data sets on EU regions. They summarise information on two interrelated sets of issues: respectively, the structural features of regions and the RIS3 priorities defined by their policy programmes. Each dataset is built by using clustering techniques applied to different types of variables: numerical, for data on the 19 socioeconomic and demographic features, considered by Pagliacci *et al*. (2019), and texts, for RIS3's priorities categorised in the automatic text analysis elaborated by Pavone *et al*. (2019). In each passage of clustering, transparent, i.e. accountable, decisions, have been taken: from the general one of defining the number of clusters, to the selection of the principal components, identification of the socioeconomic categories as well as of the number of factors to be used in clustering the groups of co-occurrences in the multidimensional space of priorities' descriptions and priorities' codes. While the process of progressive reduction of multiple categories produces some loss of information, it makes it possible to single out common or singular features that otherwise would not be observable, and to use them for policy analysis. The value added by the multidimensional analysis of both socioeconomic dimensions and priorities of smart specialisation lies precisely in that.

The results provided by cluster analysis on the results of the correspondence analysis support a complementary indication on the comparative analysis of the EU regions. In the grouping of regions obtained, it is possible to highlight the elements of homogeneity and the elements of selectivity within each of the nine groups: the former are the characteristics common to most of the regions of a group, while the latter are those occurring mainly within a group.

Policy implications emerging from the analysis presented in this paper may be considered at different levels. In particular, macro-regions that aim at designing more focused strategies may leverage on complementarities and synergies across regions each of them encompasses: these clearly emerge from homogeneous features and selectivity characters of priorities identified in the cluster analysis.

References

Barca, F. (2009). An Agenda for a Reformed Cohesion Policy. A Place-Based Approach to Meeting European Union Challenges and Expectations, Independent Report prepared at the request of Danuta Hübner, Commissioner for Regional Policy.

Benzecri, J. P. (1992). Correspondence Analysis Handbook, Dekker, New York.

Bolasco, S. (1999). Analisi multidimensionale dei dati [multidimensional analysis of data]. Roma: Carocci.

Bohlin, Ludvig, *et al*. (2014). Community detection and visualization of networks with the map equation framework, Measuring Scholarly Impact. Springer, Cham, 2014. 3-34.

Boulis C., and Ostendorf, M. (2004). Combining multiple clustering systems. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 63-74). Springer, Berlin, Heidelberg.

Charron, N., Dijkstra L., and Lapuente V. (2014). 'Regional Governance Matters: Quality of Government within European Union Member States'. Regional Studies 48 (1): 68–90. https://doi.org/10.1080/00343404.2013.770141.

European Commission (2012). Guide to Research and Innovation Strategies for Smart Specialisations (RIS 3). https://bit.ly/ZOgEpZ

European Commission (2016). Report on the implementation of EU macro-regional strategies. Available at: http://ec.europa.eu/regional_policy/en/information/ publications/reports/2016/ report-on-the-implementation-of-eu-macro-regional-strategies.

European Commission (2017). Strengthening Innovation in Europe's Regions: Strategies for Resilient, Inclusive and Sustainable Growth. Commission Staff Working Document Accompanying the Document Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions. Brussels, Publication Office. http://ec.europa.eu/regional_policy/sources /docoffic/2014/com_2017_376_2_en.pdf.

Foray D., David, P.A., and Hall B. (2009). Smart Specialisation: The Concept, *Knowledge for Growth Expert Group*.

Foray, D., Goddard, J., Morgan, K., Goenaga Beldarrain, X., Landabaso, M., Neuwelaars, C., & Ortega-Argilés, R. (2012). Guide to research and innovation strategies for smart specialisation (RIS3), S3 Smart Specialisation Platform. Seville: IPTS Institute for Prospective Technological Studies, Joint Research Centre of the European Commission. Available at: http://s3platform.jrc.ec.europa. eu/en/c/ document_library/get_file?uuid = e50397e3-f2b1-4086-8608-7b86e69e8553& groupId = 10157

Foray, D. (2018). Smart Specialisation Strategies and Industrial Modernisation in European Regions—Theory and Practice'. Cambridge Journal of Economics, October. https://doi.org/10.1093/cje/bey022.

Gianelle, C., Guzzo F., and Mieszkowski K. (2017). Smart Specialisation at Work: Analysis of the Calls Launched under ERDF Operational Programmes, 11/2017. JRC Technical Reports, S3 Working Paper Series. Seville: European Commission: Joint Research Centre (JRC), the European Commission's science and knowledge service.

Greenacre, M. J. (2007). Correspondence analysis in practice. London: chapman & Hall.

Jain A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Kassambara, A. (2017). Practical Guide to Principal Component Methods in R. Create Space Independent Publishing Platform. United States.

Iammarino, S., Rodriguez-Pose, A., Storper, M. (2018). Regional inequality in Europe: evidence, theory and policy implications, *Journal of Economic Geography*. https://doi.org/10.1093/ jeg/lby021.

McCann, P. & Ortega-Argilés, R. (2015). Smart Specialization, Regional Growth and Applications to European Union Cohesion Policy. Regional Studies 49(8): 1291-1302.

Navarro J.P. J., and Uihlein A. (2016). Regional Mapping. Science for Policy Report by the Joint Research Centre, 72.

Pagliacci, F., Pavone, P., Russo, M., Giorgi (2019). A. Regional structural heterogeneity: Evidence and policy implications for RIS3 in macro-regional strategies, *Regional Studies*, doi 10.1080/00343404.2019.1635689.

Pavone, P., Pagliacci, F., Russo, M., Giorgi, A. (2019). R&I smart specialisation strategies: classification of EU regions' priorities. Results from automatic text analysis. DEMB Working Paper Series n.148 http://merlino.unimo.it/campusone/web_dep/wpdemb/0148.pdf

Russo M., Pagliacci F., Pavone P., Giorgi, A. (2019). RIS3 in macro-regional strategies: tools to design and monitor integrated territorial development paths. DEMB Working Paper Series n.145. ISSN: 2281-440X online. http://merlino.unimo.it/campusone/web_dep/wpdemb /0145.pdf

Annex 1 – Coordinates of categories referred to Socioeconomic classification, Priority Description Classification and Priority Codes Classification, on the first 4 Factors

Legend: colours of Label class and Type highlight the three classifications: socioeconomic classification (SEc), RIS3 prioritities descriptions (desc) and codes (cod). Coordinates of categories in each factor are coloured according the maximum-minimum value in each column. Relative weight of the category refers to the whole dataset

Label Class	Type	Relative Weight	Factor 1
Water jet cutting	desc	0.05	1.22
Very-high income; capital city-regions; diversified services	SEc	0.05	1.09
Medium-income; high-employment; manufacturing & private services	SEc	0.30	1.02
Very-high income; manufacturing; population imbalances	SEc	0.70	0.89
Low-income; high-density; high unemployment; agriculture; food & drinks; very-low educated	SEc	0.40	0.74
Healthy Food	desc	0.84	0.72
Transport & Logistics	desc	2.23	0.64
Life Science	desc	2.83	0.52
Aeronautics, Aerospace & Automotive industry	cod	1.29	0.50
High-income; low-population density; tourism	SEc	0.70	0.47
Creative industry	desc	0.79	0.44
Transport & logistics	cod	2.58	0.39
Very low-income; manufacturing; no foreigners; highly educated	SEc	0.89	0.37
Medium-income; employment imbalances; low-manufacturing; services & public sector	SEc	0.79	0.34
ICT & Tourism	desc	1.34	0.32
Mechatronics	desc	1.94	0.30
Automotive & Aerospace	desc	3.18	0.28
Optics, photonics	desc	0.25	0.26
New materials	desc	1.59	0.25
Energy Production	desc	1.69	0.24
High-income; high-employment; low-manufacturing; services & public sector	SEc	1.19	0.24
Marine & Maritime	desc	1.54	0.24
Very-high income; high-density city-regions; high-employement; highly educated; touristic	SEc	0.25	0.21
Medium-income; high-employment; highly educated; manufacturing: mining & quarrying	SEc	0.20	0.20
Manufacturing	desc	2.73	0.16
Fashion	desc	0.45	0.15
Health & Life Science	cod	6.50	0.15
Urban regions; high-income; poorer employment conditions; touristic	SEc	0.45	0.13
Blue Economy	cod	0.94	0.12
Digital &ICT	desc	5.16	0.11
Energy Production, Efficiency & Sustainability	cod	4.52	0.10
Sustainable Energy	desc	5.36	0.09
Manufacturing	cod	6.70	0.09
Health	desc	3.57	0.08
Agrofood, forestry and tobacco	cod	5.46	0.02
Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density	SEc	0.15	-0.02
ICT & digital transformation	cod	4.92	-0.05
Bioeconomy & Waste collection, treatment etc	cod	5.31	-0.14
Agrofood	desc	4.17	-0.15
Creative industry, Tourism & cultural and recreative services	cod	4.37	-0.25
Medium-income; employment & population imbalances; manufacturing: textile, basic metal, tranport; very-low educated	SEc	0.60	-0.39
Low-income; high-employment; manufacturing; no foreigners; very highly educated	SEc	0.05	-0.51
Tourism	desc	2.93	-0.52
Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very-low educated	SEc	0.30	-0.98
Bioeconomy	desc	2.23	-1.12
Very-low income; agriculture; sparsely populated; very high unemployment; traditional services (G-I)	SEc	0.65	-1.19
High-income; sparsely populated; public sector; highly educated	SEc	1.54	-1.25
SEcial innovation & education	cod	1.79	-1.38
Growth & Welfare	desc	1.24	-1.85
Very-high income; financial centres; foreigners	SEc	0.05	-2.26
Very-high income; large urban regions; high-employment; highly educated	SEc	0.25	-2.35

Label Class	Type	Relative Weight	Factor 2
Very-high income; financial centres; foreigners	SEc	0.05	2.88
Very-high income; capital city-regions; diversified services	SEc	0.05	2.58
Very-high income; large urban regions; high-employment; highly educated	SEc	0.25	2.50
Growth & Welfare	desc	1.24	0.96
Very-high income; high-density city-regions; high-employement; highly educated; touristic	SEc	0.25	0.83
Optics, photonics	desc	0.25	0.67
SEcial innovation & education	cod	1.79	0.67
Very-high income; manufacturing; population imbalances	SEc	0.70	0.65
High-income; low-population density; tourism	SEc	0.70	0.52
Medium-income; high-employment; highly educated; manufacturing: mining & quarrying	SEc	0.20	0.51
Water jet cutting	desc	0.05	0.50
Medium-income; high-employment; manufacturing & private services	SEc	0.30	0.41
Healthy Food	desc	0.84	0.36
Mechatronics	desc	1.94	0.34
Life Science	desc	2.83	0.34
Medium-income; employment imbalances; low-manufacturing; services & public sector	SEc	0.79	0.32
High-income; high-employment; low-manufacturing; services & public sector	SEc	1.19	0.31
High-income; sparsely populated; public sector; highly educated	SEc	1.54	0.30
Transport & Logistics	desc	2.23	0.28
Transport & logistics	cod	2.58	0.25
Digital &ICT	desc	5.16	0.23
ICT & digital transformation	cod	4.92	0.23
Health & Life Science	cod	6.50	0.22
Bioeconomy	desc	2.23	0.15
Health	desc	3.57	0.15
Bioeconomy & Waste collection, treatment etc	cod	5.31	0.14
New materials	desc	1.59	0.13
Creative industry	desc	0.79	0.13
Aeronautics, Aerospace & Automotive industry	cod	1.29	0.12
Manufacturing	cod	6.70	0.10
Energy Production	desc	1.69	0.10
Medium-income; employment & population imbalances; manufacturing: textile, basic metal, tranport; very-low educated	SEc	0.60	0.08
Automotive & Aerospace	desc	3.18	0.03
Sustainable Energy	desc	5.36	0.02
ICT & Tourism	desc	1.34	0.00
Energy Production, Efficiency & Sustainability	cod	4.52	-0.03
Manufacturing	desc	2.73	-0.28
Creative industry, Tourism & cultural and recreative services	cod	4.37	-0.38
Urban regions; high-income; poorer employment conditions; touristic	SEc	0.45	-0.40
Low-income; high-density; high unemployment; agriculture; food & drinks; very-low educated	SEc	0.40	-0.42
Agrofood, forestry and tobacco	cod	5.46	-0.48
Very low-income; manufacturing; no foreigners; highly educated	SEc	0.89	-0.48
Marine & Maritime	desc	1.54	-0.59
Agrofood	desc	4.17	-0.68
Low-income; high-employment; manufacturing; no foreigners; very highly educated	SEc	0.05	-0.72
Tourism	desc	2.93	-0.89
Fashion	desc	0.45	-1.06
Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density	SEc	0.15	-1.11
Blue Economy	cod	0.94	-1.20
Very-low income; agriculture; sparsely populated; very high unemployment; traditional services (G-I)	SEc	0.65	-1.77
Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very-low educated	SEc	0.30	-2.46

Label Class	Type	Relative Weight	Factor 3
Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density	SEc	0.15	1.26
Very low-income; manufacturing; no foreigners; highly educated	SEc	0.89	1.01
High-income; sparsely populated; public sector; highly educated	SEc	1.54	0.92
Fashion	desc	0.45	0.91
High-income; low-population density; tourism	SEc	0.70	0.88
Manufacturing	desc	2.73	0.78
Water jet cutting	desc	0.05	0.72
New materials	desc	1.59	0.68
Low-income; high-employment; manufacturing; no foreigners; very highly educated	SEc	0.05	0.56
Optics, photonics	desc	0.25	0.49
Manufacturing	cod	6.70	0.36
Mechatronics	desc	1.94	0.35
Medium-income; employment imbalances; low-manufacturing; services & public sector	SEc	0.79	0.26
Agrofood	desc	4.17	0.22
Sustainable Energy	desc	5.36	0.18
ICT & digital transformation	cod	4.92	0.17
Health	desc	3.57	0.15
Bioeconomy & Waste collection, treatment etc	cod	5.31	0.14
Very-low income; agriculture; sparsely populated; very high unemployment; traditional services (G-I)	SEc	0.65	0.10
Digital &ICT	desc	5.16	0.08
Transport & Logistics	desc	2.23	0.06
Bioeconomy	desc	2.23	0.03
Automotive & Aerospace	desc	3.18	0.02
Agrofood, forestry and tobacco	cod	5.46	0.01
High-income; high-employment; low-manufacturing; services & public sector	SEc	1.19	0.01
Medium-income; employment & population imbalances; manufacturing: textile, basic metal, tranport; very-low educated	SEc	0.60	-0.03
Transport & logistics	cod	2.58	-0.04
Energy Production, Efficiency & Sustainability	cod	4.52	-0.07
Energy Production	desc	1.69	-0.11
Very-high income; capital city-regions; diversified services	SEc	0.05	-0.12
Health & Life Science	cod	6.50	-0.14
Tourism	desc	2.93	-0.21
Medium-income; high-employment; manufacturing & private services	SEc	0.30	-0.23
Creative industry	desc	0.79	-0.32
Very-high income; large urban regions; high-employment; highly educated	SEc	0.25	-0.35
SEcial innovation & education	cod	1.79	-0.42
Life Science	desc	2.83	-0.47
Creative industry, Tourism & cultural and recreative services	cod	4.37	-0.47
Marine & Maritime	desc	1.54	-0.53
Growth & Welfare	desc	1.24	-0.67
Aeronautics, Aerospace & Automotive industry	cod	1.29	-0.69
Blue Economy	cod	0.94	-0.70
Very-high income; manufacturing; population imbalances	SEc	0.70	-0.85
Very-high income; high-density city-regions; high-employement; highly educated; touristic	SEc	0.25	-0.86
Urban regions; high-income; poorer employment conditions; touristic	SEc	0.45	-0.92
Medium-income; high-employment; highly educated; manufacturing: mining & quarrying	SEc	0.20	-0.96
ICT & Tourism	desc	1.34	-1.18
Healthy Food	desc	0.84	-1.28
Low-income; high-density; high unemployment; agriculture; food & drinks; very-low educated	SEc	0.40	-1.39
Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very-low educated	SEc	0.30	-1.53
Very-high income; financial centres; foreigners	SEc	0.05	-2.10

Label Class	Type	Relative Weight	Factor 4
Water jet cutting	desc	0.05	3.20
High-income; high-employment; low-manufacturing; services & public sector	SEc	1.19	1.68
Optics, photonics	desc	0.25	1.66
Very-high income; high-density city-regions; high-employement; highly educated; touristic	SEc	0.25	1.24
Energy Production	desc	1.69	0.91
Medium-income; high-employment; manufacturing & private services	SEc	0.30	0.63
Creative industry	desc	0.79	0.60
Growth & Welfare	desc	1.24	0.58
Transport & Logistics	desc	2.23	0.54
Low-income; high-employment; manufacturing; no foreigners; very highly educated	SEc	0.05	0.52
Blue Economy	cod	0.94	0.52
Marine & Maritime	desc	1.54	0.48
Transport & logistics	cod	2.58	0.44
Manufacturing	desc	2.73	0.39
Energy Production, Efficiency & Sustainability	cod	4.52	0.36
Medium-income; high-employment; highly educated; manufacturing: mining & quarrying	SEc	0.20	0.25
Life Science	desc	2.83	0.17
Bioeconomy	desc	2.23	0.13
Very low-income; manufacturing; no foreigners; highly educated	SEc	0.89	0.09
Health & Life Science	cod	6.50	0.07
Agrofood	desc	4.17	0.03
Tourism	desc	2.93	0.01
SEcial innovation & education	cod	1.79	0.01
Creative industry, Tourism & cultural and recreative services	cod	4.37	-0.04
Manufacturing	cod	6.70	-0.06
Fashion	desc	0.45	-0.07
New materials	desc	1.59	-0.08
Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very-low educated	SEc	0.30	-0.08
Aeronautics, Aerospace & Automotive industry	cod	1.29	-0.10
High-income; sparsely populated; public sector; highly educated	SEc	1.54	-0.11
Medium-income; employment imbalances; low-manufacturing; services & public sector	SEc	0.79	-0.13
Agrofood, forestry and tobacco	cod	5.46	-0.16
Health	desc	3.57	-0.17
Urban regions; high-income; poorer employment conditions; touristic	SEc	0.45	-0.19
Mechatronics	desc	1.94	-0.20
ICT & digital transformation	cod	4.92	-0.20
Digital &ICT	desc	5.16	-0.20
Sustainable Energy	desc	5.36	-0.21
Very-low income; agriculture; sparsely populated; very high unemployment; traditional services (G-I)	SEc	0.65	-0.22
Bioeconomy & Waste collection, treatment etc	cod	5.31	-0.23
Very-high income; large urban regions; high-employment; highly educated	SEc	0.25	-0.30
Automotive & Aerospace	desc	3.18	-0.48
Very-high income; financial centres; foreigners	SEc	0.05	-0.51
Low-income; high-density; high unemployment; agriculture; food & drinks; very-low educated	SEc	0.40	-0.54
Very-high income; manufacturing; population imbalances	SEc	0.70	-0.65
ICT & Tourism	desc	1.34	-0.67
Medium-income; employment & population imbalances; manufacturing: textile, basic metal, tranport; very-low educated	SEc	0.60	-0.83
Healthy Food	desc	0.84	-1.15
High-income; low-population density; tourism	SEc	0.70	-1.21
Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density	SEc	0.15	-1.89
Very-high income; capital city-regions; diversified services	SEc	0.05	-3.37

Annex 2 – Histogram of the percentage inertia of the first 50 Factors