UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Economia Marco Biagi

DEMB Working Paper Series

N. 142

Detecting multidimensional clustering across EU regions. *Focus on R&I smart specialisation strategies and on socio-economic and demographic conditions*

Margherita Russo*, Pasquale Pavone**, Francesco Pagliacci***
Simone Righi****, Anna Giorgi*****

March 2019

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

** CAPP - Research Centre for the Analysis of Public Policies
E-mail: pasquale.pavone@unimore.it

*** Department Territorio e Sistemi Agro-forestali, Università di Padova, Italy, and CAPP - Research Centre for the Analysis of Public Policies
E-mail: francesco.pagliacci@unipd.it

**** Department of Computer Science, UCL, United Kingdom, and CAPP - Research Centre for the Analysis of Public Policies
E-mail: s.righi@ucl.ac.uk

***** Leader AG1 EUSALP Lombardy Region representative, and Gesdimont research centre, University of Milan, Milano, Italy
E-mail: anna.giorgi@unimi.it

ISSN: 2281-440X online

DEMB Working Paper Series n. 142

Detecting multidimensional clustering across EU regions.
Focus on R&I smart specialisation strategies and on socio-economic and demographic conditions

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

13 March 2019

[1] 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

[2] CAPP - Research Centre for the Analysis of Public Policies, pasquale.pavone@unimore.it

[3] Department Territorio e Sistemi Agro-forestali, Università di Padova, Italy, and CAPP - Research Centre for the Analysis of Public Policies, francesco.pagliacci@unipd.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 to identify similar specialisation strategies and socioeconomic characteristics. It builds on an original dataset where the EU-28 regions are classified according to their socioeconomic and demographic features and to the strategic priorities outlined in their research and innovation smart specialisations strategy (RIS3). The socioeconomic and demographic classification associates each region to one categorical variable (with 19 modalities), while the classification of the RIS3 priorities clustering was performed separately on "descriptions" (21 Boolean categories) and "codes" (11 Boolean Categories) of regions' RIS3.

Three techniques of clustering have been applied: Infomap multilayer algorithm, Correspondence Analysis plus Cluster Analysis and cross tabulation. The most effective clustering, in terms of both the characteristics of the data and the emerging results, is that obtained on the results of the Correspondence Analysis. By contrast, due to the very dense network induced by the data characteristics, the Infomap algorithm does not produce significant results. Finally, cross tabulation is the most detailed tool to identify groups of regions with similar characteristics. In particular, in the paper we present an application of cross tabulation to focus on the regions investing in sustainable development priorities. Policy implications of methods implemented 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 the enhancement of complementarities and synergies within macroregions.

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 further interventions of public policies targeting the reduction of regional disparities and the enhancement of complementarities and synergies within macroregions, namely a key instrument for the implementation of EU policies and programmes, whose main aim is fostering a greater cohesion and competitiveness across larger EU spaces, encompassing neighbouring member and non-member States, endorsed by the European Council and supported by the ESIF, among others (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 (RIS3). The concept of smart specialisation on research and innovation stems from academic work on the key drivers for bottom-up policies aiming at structural changes that are needed to enhance 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 RIS3 as an ex-ante conditionality for access of regions to European Regional Development Funds. Such policies are built on specific guidelines and 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 EUR 65 billion of ERDF have been allocated to such policies, they are not yet under scrutiny for the actual impact they have produced nor for the effective monitoring that was supposed to be implemented (as a crucial tool of that policy)[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, a forum to support regions with information and tools for bottom up coordination. Other loci of interaction among regions

[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.

are those supported by the EU Interreg programmes[4], the 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].

In general, several analyses provide analytical frameworks to discuss relevant issues to be addressed by public policies, such as income disparities (Iammarino *et al*., 2018) or quality of institutions (Charron *et al*., 2014), but so far no systematic analysis has focused together on the different aspects of EU regions specialization strategies and on their socio-economic characteristics. This paper intends to fill this gap by applying a multidimensional clustering of EU-28 regions to identify similar specialisation strategies and socioeconomic characteristics. A clustering based on these aspects can be expected to provide clues for more effective regional policies. The clustering proposed in the paper builds on an original dataset created by the research team, where the EU-28 regions are classified according to their socioeconomic features (Pagliacci *et al*., 2018) and to the strategic features of their research and innovation smart specialisations strategy (RIS3) (Pavone *et al*., 2018). In the former 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 classification of RIS3, 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 textual classification of priorities using information on RIS3 made available on line in the platform Eye@RIS3 (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. Methods and data

One of the general objectives of the analysis of complex phenomena concerns the possibility of defining classes of elements from a plurality of interconnected elementary measurements. Techniques of automatic classification allow the organizing of objects into groups which have similar members based on specific criteria for evaluating their similarity.

The dataset analysed in this paper results from the merging of two main datasets, developed in previous papers. First of all, we use the classification, provided in Pagliacci *et al*. (2018), of regions categorised according to their socioeconomic features: with

[4] https://www.interregeurope.eu/
[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)

regard to 208 territorial entities in EU-28 regions, a socio-economic disjunctive categorical variable is defined, with 19 categories. Secondly, with regard to smart specialisation strategies, we use the classification defined by Pavone *et al*. (2018): with regard to 216 territorial entities, in EU-28, priorities of RIS3 are summarised in two multi-class categorical variables, respectively, Description (21 categories) and Codes (11 categories)[8]. Merging the two datasets, in this paper we study the multidimensional classification of 191 territorial entities according to the three categorical variables.

In order to provide multidimensional clustering of regions, we suggest two methods[9]. The state of the art in clustering is provided by a literature in continuous and rapid growth (Jain, 2010, Duda *et al*., 2012), developed in a variety of scientific fields with different languages and focusing on the most diverse problems. Clustering heterogeneous data; definition of parameters and initializations (such as the times of iterations in K-means (MacQueen, 1967) and the threshold in hierarchical clustering [Jain 1988]), as well as the problem of defining the optimal number of groups. In this last direction, research is increasingly focusing on combining multiple clustering of the same dataset to produce a better single one (Boulis & Ostendorf, 2004).

In order to obtain groups of regions based on the similarity of their profiles, it is possible to carry out, in order, 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 (regions' socio-economic and demographic category) and two multi-class categorical variables (respectively, *Codes* and *Descriptions* of regions' RIS3's priorities), we directly apply a Correspondence Analysis to the Boolean matrix *Regions* × *Modes* (191×51), in which the totals of rows depend on the number of categories in which each region has been classified[10]. Consequently, two regions are considered more similar to each other – and thus closer together on a factorial plan – if

[8] With regard to the two multi-class categorisations of regions, they derive from an automatic classification of the priorities specified by each region in terms of free text of descriptions and of codes belonging to three domains: scientific domain, economic domain and policy objectives. Dataset downloaded on 01 October 2018 from Eye@RIS3 platform, EC-JRC. 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, respectively) has allowed recognition and classification of the priorities in 21 topic groups of description and 11 topic groups of codes. In assigning each classification of priorities to the different regions, we obtain a multiclass matrix of *Regions* × *Modes* of Description and Codes.

[9] A third method, based on a network representation of the different data classifications and on the application of the Infomap multilayer algorithm, turns out to be unsuitable for our application. Its results are discussed in Annex 3, for completeness.

[10] Usually, a matrix unit × categorical variables (univocal classification) is studied through a multiple correspondences analysis that transforms the matrix unit × variables (m×s) into a Boolean matrix unit × categories (m×n). This last 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 according to the criterion of the best orthogonal projections.

they fall in the same categories more than the variables considered[11]. With the Correspondence Analysis, the factors highlight the configuration of the profiles in a graphic context. The interpretation of each factor through the analysis of the nodes' polarization, sheds light upon the association structure among regions' profiles[12].

The second method presented in this paper implements cross tabulation on the three classifications, by combining, for each region, the set of categories (Socio economic, Codes of priority, Description of priority) in which it has been categorized. For any given socioeconomic class, two different paths to create cross tabulations can be followed, given the presence of two multiclass variables. Starting from the matrix *Records* × *Classifications*, the cross-references between descriptions and codes concern the priority in which these two classifications coexist. Alternatively, in the cross tabulation built from the matrix *Regions* × *Modes* (Description and Codes), the crossings between description and codes concern their coexistence in the same region, regardless of whether they refer to the same priority. Depending on which unit of analysis is selected, record or region, it is possible to create different contingency matrices between pairs of priority classifications, keeping the third classification fixed (the socioeconomic class, in our analysis).

In this paper, we elaborate cross tabulations by focusing on records as the unit of analysis, and associating to each record the socioeconomic category of the region. In this way, each region is represented in the table for every combination of the description and code classification of its records. In the Correspondence Analysis, the unit of analysis is the region: thus, for each region, any code category and any description category is considered, regardless of their co-occurrence in the same record.

3. Results

Correspondence Analysis and Cluster Analysis

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 plan, respectively, of the 51 modes and of the 191 regions. By analysing Figure 1, we observe that the first factor polarises information on the type of production, 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 characterized by the production of services vs. the production of goods,

[11] At the extreme, two points are superimposed on the factorial plan if they assume the same values for all the variables.

[12] Among the plans 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.

while from bottom to top there are regions characterized by a low income vs. a high income.

Figure 1 - Distribution on f1f2 plan of the 51 modes

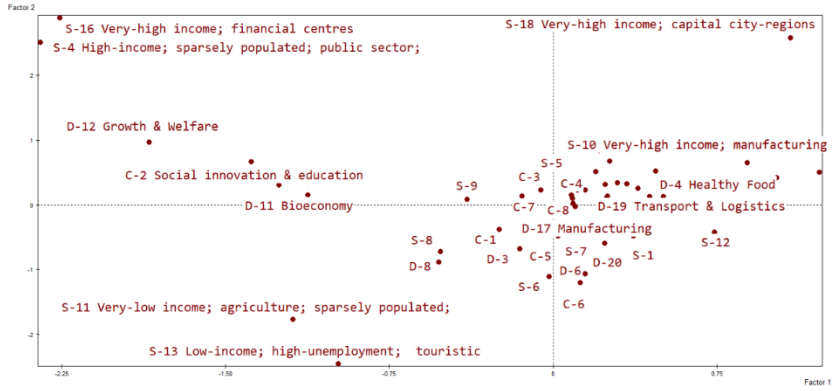
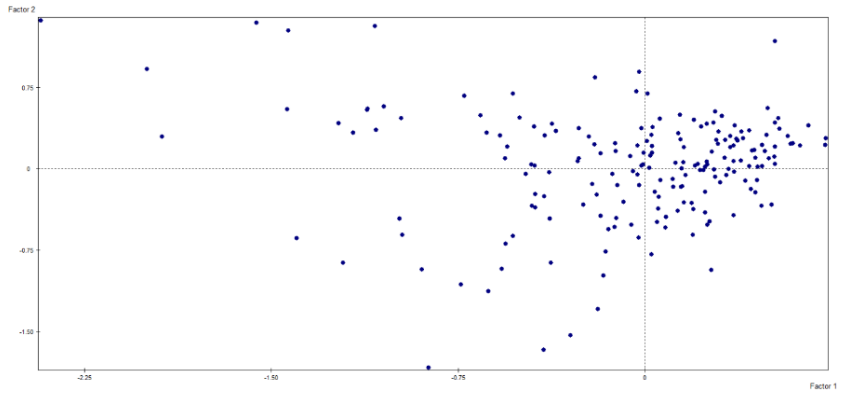


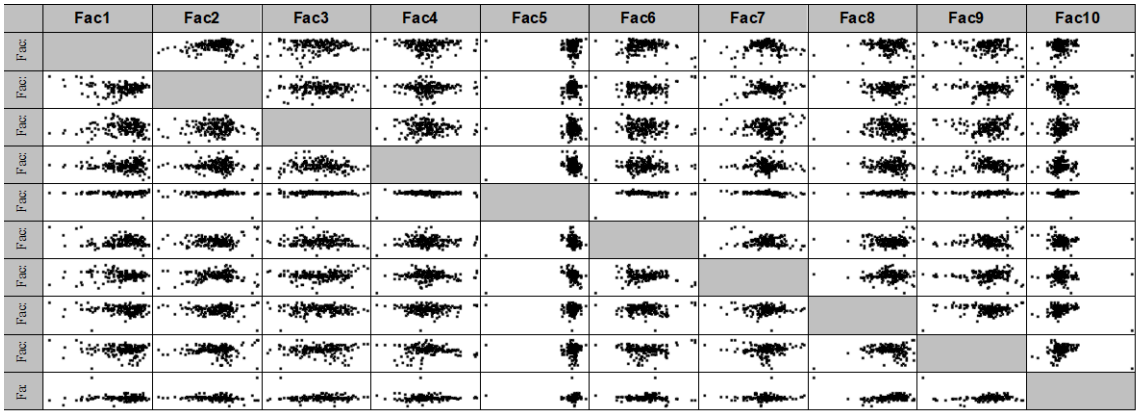
Figure 2 - Distribution on f1f2 plan of the 191 regions



In the clustering process applied to such results, each factor represents only a part of the overall 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 matrix of factorial plans[13]. In particular, Figure 3 presents all possible combinations of the first 10 factors. They show different projections of the cloud of points and highlight outliers. In particular, the 5th factor singles out only the difference between one region (in this case, the Brussels region - BE01) and all the others. The same holds true for 10th factor (in this case, the Luxembourg region - LU00). When five factors are considered, a 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 observing the resulting dendrogram, nine groups of regions emerge.

[13] 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.

Figure 3 - Factorial plans relating to all possible combinations of the first 10 factors

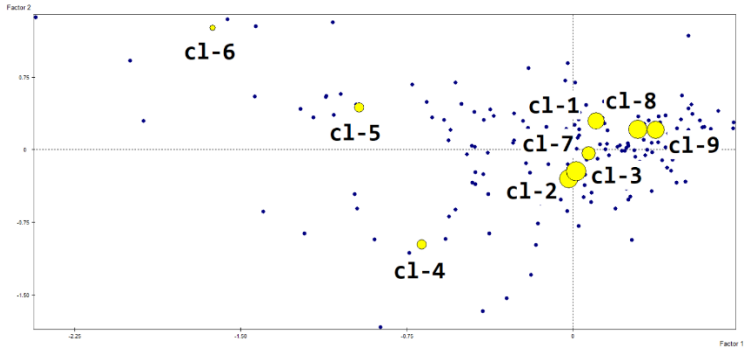


The most polarized groups among them are: clusters #6, #9 and #4 (Figure 4). While the clusters # 2; #3; #7 and #1, are very close to each other on a factorial level and are in a more barycentric position, this information highlights a lesser diversity among them. Each cluster is characterized by some categories that do not represent all of those through which the regions have previously been classified, but only the distinctive features of the different groups.

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 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 mode in the cluster to all modes in the cluster (columns 6) highlights the extent to which the category is characteristic. For the most polarised clusters, i.e. the groups that are furthest from the centre of gravity (clusters #4, #5, #6 and #9, in Figure 4), the weight of characteristic categories is relatively higher, respectively 42.98%, 36.52% 33.33%, and 23.57% (see the total value in bold, in Table 1, column 6).

Figure 4 - Distribution on *f1f2* plan of the 191 regions and nine partitions

legend: black dots: regions; yellow circles: clusters, with size proportional to the number of regions in the cluster



[14] Test-value for qualitative variable modes is a statistical criterion associated with the comparison of two portions within the framework of a hypergeometric law approximated by a standardized normal law. The test-value = 2.1 corresponds to a bilateral test probability α/2 of less than 2.5%.

Table 1 - Characteristic categories of the nine clusters of regions

cluster ID and label of characteristic frequencies	(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 mode in the Cluster to all modes in the Cluster	(7) % of the mode in the Cluster **SELECTIVITY**	(8) % of regions with the mode 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, tranport; very poorly 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 poorly 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 poorly 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 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 characterized 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, 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 will be now described with regard to the selectivity/homogeneity of their characteristic categories.

Cluster #1, encompassing 31 regions, is characterized 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 mode in the cluster, while the second one represents an element of homogeneity within the group.

Cluster #2 comprises 31 regions and it is characterized by two distinct socio-economic classes (both characterized 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 characterized 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 characterized 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, characterized 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.

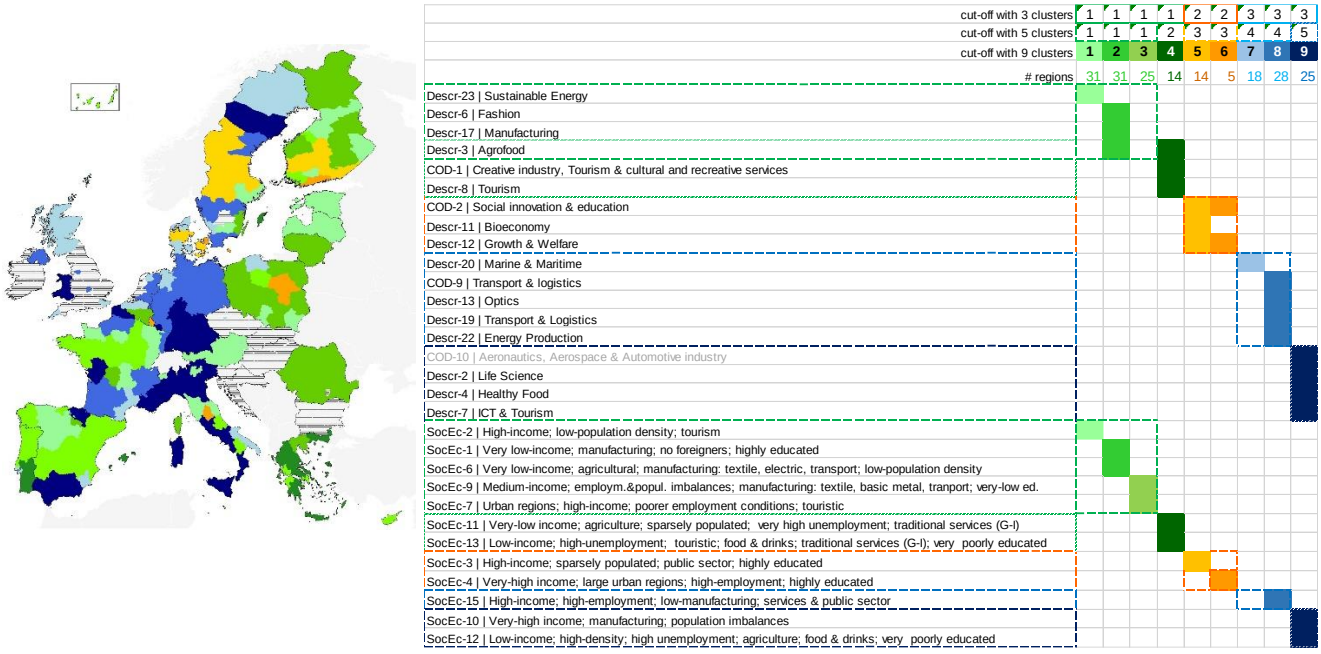
Cluster #8 comprises 28 regions and it is characterized by the socio economic class *High-income; high-employment; low-manufacturing; services & public sector* (with 70.83% occurrences in the cluster, referring to 60.7% of regions) and by the priority descriptions: *Optics* (with 100% occurrences in 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 characterized 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 5 maps the nine clusters, with the table in the right panel summarising the homogeneity and selectivity elements characterizing them. 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 5 - Maps of clusters of regions, by socioeconomic features and RIS3s' priorities: summary of selectivity and homogeneity characteristic categories



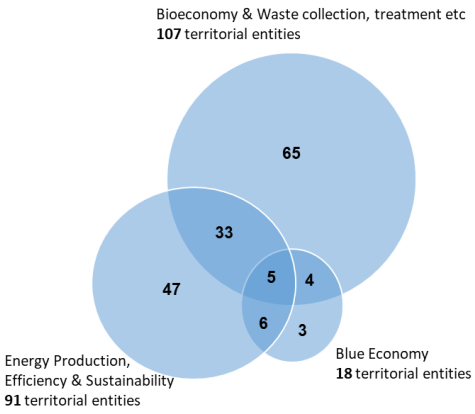
Cross tabulation

Cross tabulation allows the analysis of combinations of the three classifications under analysis. Russo *et al.* (2018) explore cross tabulation with a focus on the EUSALP macroregion. Here, in order to exemplify and explore the resulting combinations, we focus on the three categories of codes (as defined by Pavone *et al.* 2018) belonging to the broad category of policies aiming at supporting sustainable development[15]. To this end, we extracted the group of 163 regions that have explicitly oriented their smart specialisation strategy towards more specific areas that have been classified as sustainable development: *Blue Economy* (18 regions), *Bio Economy* (107 regions), and *Energy Production, Efficiency & Sustainability* (91 regions). The Venn diagram of the three

[15] Annex 1 contains the characteristic dictionaries of codes for the three classes of codes considered in this section. For this and other priorities, it will be possible to browse specific queries on line in the Platform of Knowledge implemented by EUSALP [the first release for this tool is scheduled for April 2019].

codes, represented in Figure 2, highlights that almost 30% of those regions have more than one of the three priorities.

Figure 2 – Venn Diagram of Modes represented in Table 1



The three sections of Table 2 present the list of territorial entities classified by each of the three codes, respectively in the top, the middle and the bottom section. In each section, names of regions are repeated when associated to more than one description (in such cases, a dot is added in front of the region' NUTS identification code). In the header of the columns, the socioeconomic classes are grouped by macro category, according to the three macro groups identified in the cluster analysis - Eastern manufacturing regions, Mediterranean traditional-economy regions, North-Western EU regions - and their subgroups (see Pagliacci *et al*. 2018). The total number of regions by socioeconomic class is listed in the 4th row and for each of the codes under analysis.

As a general result, it is possible to observe that not all the socioeconomic classes nor all the descriptions are associated to the EU regions. Moreover, as we could expect, codes are largely associated to the description in the same area. Indeed, the Code Blue Economy mainly occurs in the Description class *Marine and Maritime*, while in only two regions it is associated to *Agrofood* and *Automotive & Aerospace*[16], respectively.

For what concerns the Code class Bio economy, Table 2 shows that, for 65 regions, this class is associated to the Description *Sustainable Energy*. In this Description class there are regions belonging to all the different socioeconomic classes, with the exceptions of the class "Medium-income; high-employment; manufacturing & private services" (no region of this class has been classified with a priority in bio economy) and the class "Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very poorly educated", which have no characterising priorities in this Description. With regard to this socioeconomic class, the Code Bio economy concerns Tourism and Health priorities (respectively, for the regions ES53 and ES70[17]). Some regions associate the Code Bio economy to more than one description. For example, regions PL21 and PL42, in Poland, have defined the code priority Bio economy in the priority descriptions of Health, of Agrofood, as well as Manufacturing[18].

[16] See Annex 2 for details on free text descriptions associated to this region.
[17] See Annex 2 for details on free text descriptions associated to these regions.
[18] See Annex 2 for details on free text descriptions associated to these regions.

Table 2 – The EU-28 regions with priorities classified in the three selected codes of priorities Bio economy, Blue Economy, Energy Production, Efficiency and Sustainability, by descriptions of priorities (rows) and socio-economic class (columns)

		Socio Economic Classes																			
		North-Western EU regions												Eastern manufacturing regions			Mediterranean traditional-economy regions				
		Very-high income capital city-regions		*Other urban regions*			Very-high income manuf. regions	High-income low-population density regions		Medium-high income regions, services & public sector		High-employment, with advanced services		Eastern manufacturing regions			Regions with traditional economy & empl.imbalances			Tourist. areas; tradit. econ.	
	class id	*S-16*	*S-18*	*S-4*	*S-5*	*S-7*	*S-10*	*S-2*	*S-3*	*S-15*	*S-14*	*S-19*	*S-17*	*S-8*	*S-1*	*S-6*	*S-9*	*S-12*	*S-11*	*S-13*	# reg.s
		Very-high income; financial centres; foreigners	Very-high income; capital city-regions; diversified services	Very-high income; large urban regions; high-employment; highly educated	Very-high income; high-density city-regions; high-employment; highly educated; touristic	High-income; urban regions poorer with employment conditions; touristic	Very-high income; manufacturing; population imbalances	High-income; low-population density; tourism	High-income; sparsely populated; public sector; highly educated	High-income; high-employment; low-manufacturing; services & public sector	Medium-income; employment imbalances; low-manufacturing; services & public sector	Medium-income; high-employment; manufacturing & private services	Medium-income; high-employment; highly educated; manufacturing: mining & quarrying	Low-income; high-employment; manufacturing; no foreigners; very highly educated	Very low-income; manufacturing; no foreigners; highly educated	Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density	Medium-income; employment & population imbalances; manufacturing: textile, basic metal, transport; very poorly educated	Low-income; high-density; high unemployment; agriculture; food & drinks; very poorly educated	Very-low income; agriculture; sparsely populated; very high unemployment; traditional services	Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very poorly educated	
	# of territorial entities in the dataset*	1	1	5	5	9	14	14	31	24	16	6	4	1	18	3	12	8	13	6	191
	#...with sustainable development	**1**	**1**	**5**	**3**	**9**	**12**	**13**	**25**	**20**	**16**	**5**	**4**	**1**	**16**	**3**	**7**	**8**	**8**	**6**	**163**
107 regions with CODE category "Bio economy"																					
Description	**1 Health**								**.FI1C2**						**.PL21**					ES70	**3**
	3 Agrofood			**.PL12**				**.AT12**		DE94	**.FR61**				**.PL42**						**5**
	7 ICT & Tourism																	ITF3			**1**
	8 Tourism															**.RO41**				ES53	**2**
	9//10 Digital &ICT						DE1	ITH1	SE321	DE8							**.ES41** **.ITF5**				**6**
	11 Bioeconomy			FI1B1 **.PL12**					**.FI196** FI1C1	SE231	DK02				PL32 RO22		**.ITF5**		**.EL61** **.EL63** PT18		**11**
	12 Growth & Welfare			DK01				AT33	SE322												**3**
	13 Optics										**.FR61**										**1**
	14 Photonics									**.FR71**	FR24										**2**
	15//18 Mechatronics								SE212									ES52			**2**
	16 Automotive & Aerospace			FR10			ITC3	AT32							PL43	**.RO41**	PT16				**6**
	17 Manufacturing								**.FI1C5** **.FI1D1** **.FI1D2** **.FI1D4**						**.PL21** **.PL42** PL51 PL52						**8**
	19 Transport & Logistics										FR43 **.FR61**										**2**
	20 Marine & Maritime				DE5	PT17			**.FI196**		FR51 FR52										**5**
	22 Energy Production						DE7		FI1C3	BE2 **.DEE** **.FR71**											**5**
	23 Sustainable Energy	LU00	BE1	SE110	DE3	CY00 EL30 ES30 ES51 FR81 ITI4 MT00	ES21 ES22 ES24 ITC1	**.AT12** AT22 EE00 ITC2 ITH2	FI193 **.FI196** **.FI1C2** **.FI1C5** **.FI1D1** **.FI1D2** FI1D3 **.FI1D4** FI1D5 FR63 SE213 SE313	DE4 **.DEE** DEG FR42 NL1 NL12 SE232	BE3 FR26 FR30 FR41 FR43 FR53 **.FR61** FR72		PL22	PL33	LT00 PL34 RO21	RO31 **.RO41** RO42	ES11 ES12 **.ES41**	ES61 ES62 ITF6	EL53 **.EL61** **.EL63** ES42		**65**
	no-description						DE2		FI1D7												**1**
	number of regions	**1**	**1**	**5**	**2**	**8**	**8**	**8**	**18**	**12**	**12**		**1**	**1**	**10**	**3**	**5**	**5**	**5**	**2**	

		Socio Economic Classes																			
		North-Western EU regions												Eastern manufacturing regions			Mediterranean traditional-economy regions				
		Very-high income capital city-regions		*Other urban regions*			Very-high income manuf. regions	High-income low-population density regions		Medium-high income regions, services & public sector		High-employment, with advanced services		Eastern manufacturing regions			Regions with traditional economy & empl.imbalances			Tourist. areas; tradit. econ.	
class id		*S-16*	*S-18*	*S-4*	*S-5*	*S-7*	*S-10*	*S-2*	*S-3*	*S-15*	*S-14*	*S-19*	*S-17*	*S-8*	*S-1*	*S-6*	*S-9*	*S-12*	*S-11*	*S-13*	# reg.s
		Very-high income; financial centres; foreigners	Very-high income; capital city-regions; diversified services	Very-high income; large urban regions; high-employment; highly educated	Very-high income; high-density city-regions; high-employment; highly educated; touristic	High-income; urban regions poorer with employment conditions; touristic	Very-high income; manufacturing; population imbalances	High-income; low-population density; tourism	High-income; sparsely populated; public sector; highly educated	High-income; high-employment; low-manufacturing; services & public sector	Medium-income; employment imbalances; low-manufacturing; services & public sector	Medium-income; high-employment; manufacturing & private services	Medium-income; high-employment; highly educated; manufacturing: mining & quarrying	Low-income; high-employment; manufacturing; no foreigners; very highly educated	Very low-income; manufacturing; no foreigners; highly educated	Very low-income; agricultural; manufacturing: textile, electric, transport; low-population density	Medium-income; employment & population imbalances; manufacturing: textile, basic metal, transport; very poorly educated	Low-income; high-density; high unemployment; agriculture; food & drinks; very poorly educated	Very-low income; agriculture; sparsely populated; very high unemployment; traditional services	Low-income; high-unemployment; touristic; food & drinks; traditional services (G-I); very poorly educated	
# of territorial entities in the dataset*		1	1	5	5	9	14	14	31	24	16	6	4	1	18	3	12	8	13	6	191
#...with sustainable development		**1**	**1**	**5**	**3**	**9**	**12**	**13**	**25**	**20**	**16**	**5**	**4**	**1**	**16**	**3**	**7**	**8**	**8**	**6**	**163**

18 regions with CODE category "Blue economy"

Description	**3**	**Agrofood**																		EL42	**1**
	16	**Automotive & Aerospace**													PL62						**1**
	20	**Marine & Maritime**								DE94 DK05 SE232					PL42 PL63	EL30 FR81 MT00	PT11 PT16	ES62 ITF4 ITG1		EL62 PT15	**15**
		no-description															ES11				**1**
	number of regions									**3**					**3**	**3**	**3**	**3**		**3**	

91 regions with CODE category "Energy Production, Efficiency & Sustainability"

	9//10	**Digital &ICT**																	**.EL63**		**1**
	11	**Bioeconomy**								SE311 SE312	DK04										**3**
	12	**Growth & Welfare**									NL2										**1**
	15//18	**Mechatronics**													RO11						**1**
	16	**Automotive & Aerospace**													PL61				ES43		**2**
	17	**Manufacturing**																	EL52		**1**
	19	**Transport & Logistics**										FR23									**1**
	20	**Marine & Maritime**									SE221	FR51	**.UKK3**	**.UKM**			ES13				**5**
	22	**Energy Production**				DE3 DE5 DE6		ES22 ES24	FR83	**.FI1C5** SE123 SE321	DE8 DE9 DE94 DEB DEE DK03 FR71 NL12 SE232	FR21 FR22 **.FR24**	**.UKK3** UKN	SE331	PL33	PL52			ES62 ITF4	**.EL63** ES42		**30**
	23	**Sustainable Energy**					CY00 ES30 FR81 FR82	DE1 DEA DEC ES21 ITH5	AT11 AT21 AT34 ITH1 LV00	FI1C4 **.FI1C5** FI1D4 FI1D6 SE121 SE125	DK05 FR42 NL1 SE221	**.FR24** FR25 FR41	UKF24 UKJ4 **.UKK3** UKL	SE332 **.UKM**		PL11 PL21 PL31 PL63 RO22		ITF5	ES61 ITF3 ITG1 ITG2	EL53 EL61 **.EL63** EL64	EL42 EL43 ES70 PT15	**51**
		no-description						ITC4										ES11				**2**
	number of regions				**3**	**4**	**8**	**6**	**10**	**16**	**7**	**5**	**3**	**1**	**8**		**3**	**6**	**7**	**4**	

Similarly, with regard to the Code class Energy Production, Efficiency & Sustainability, most regions refer to the descriptions mirroring the code category (respectively, 30 regions have priorities in Energy production and 51 in Sustainable energy). In 15 cases, the categories of Descriptions highlight different areas, which mainly concern manufacturing and transport and logistics.

The result presented in Table 2 might support focused initiatives targeted to all the regions mentioned in Table 2, for a discussion of the specific programmes and initiatives that regions have implemented on those priorities [19]. In particular, macroregions can easily identify which the regions are that share the same socioeconomic characteristics and priorities and focus on potential complementarities, learning, and so on[20].

4. 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. By identifying clusters of EU regions, we provide policy makers with a more systematic and informed tool which 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 categorizations), 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 of data 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 data set is built by using clustering techniques applied to different types of variables: numerical, for data on the 16 socioeconomic and demographic features, considered by Pagliacci *et al*. (2018), and texts, for RIS3's priorities categorised in the automatic text analysis elaborated by Pavone *et al*. (2018). 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.

[19] FONA Forum and Workshop on Sustainable STI and SDGs, on 13-14 May 2019, Berlin, Germany

[20] Indeed, the very detailed picture of specific priorities resulting from the cross tabulation, presented in Table 2, is broader than the information one can obtain from the S3 Platform with regard to the areas of Energy, as one example in the domain of sustainable development.

In summing up the results obtained with the two techniques of clustering applied in this paper, Correspondence Analysis and cross tabulation[21], we focus here on what is missing and what is emerging in these processes of elaboration.

The most effective clustering, in terms of both the characteristics of the data and the emerging results, is that obtained with a Correspondence Analysis. On the contrary, given the very dense network, the Infomap algorithm does not produce significant results. Finally, cross tabulation is the finest-grained tool to identify the groups of regions with similar characteristics. This method will be implemented in the Platform of Knowledge – developed by EUSALP - as a tool to browse information on regions. It will support queries to select regions with given characteristics in terms of priorities of smart specialisation or in terms of socioeconomic features. In the paper, we have presented an application with a focus on the regions investing in sustainable development priorities. Online queries will allow easy access to more detailed information on specific areas of policy interventions, as they are described in the vocabularies associated to the categories of priorities (according to free text descriptions and codes), as well as on the values characterising the socioeconomic and demographic variables in the regions that are grouped.

The results provided by the two methods - factor analysis and cross tabulation - support different and complementary indications on the comparative analysis. In the grouping of regions obtained through factor analysis, 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. Cross tabulation provides very detailed information, for example, for a given domain of policy intervention, like bio economy or blue economy, about the regions orienting their priorities in that direction, informing on what the socioeconomic conditions and the priorities defined by the territorial entities under analysis are. Both methods are grounded on a systematic exploration of the original information through statistical criteria. Ambiguity or misspecification may be controlled for those cases that experts or practitioners do not recognize as appropriate.

Policy implications emerging from the research activity 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: these clearly emerge from homogeneous features and selectivity characters of priorities identified in the cluster analysis. Strategic partnerships within and across macroregions may be outlined by the more focused selection emerging from the cross tabulation: the analysis of priorities concerning sustainable development goals highlight many possible collaborations based on which regions may start a fruitful analysis of practices and results of regions within the selected set of priorities and socio-economic conditions.

[21] As noted in Annex 3, the use of tools from community detection, such as Infomap, did not yield satisfactory results.

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.

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.

Duda R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons. New York.

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.

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. CreateSpace 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 A. (2018). Should we consider regional structural heterogeneity in learning from RIS3? Evidence and policy implications for macro-regional strategies. Deliverable prepared for the EU project "AlpGov", Work Package T-3.1 "Mapping the governance in the research and innovation field of the Eusalp regions", 31th December 2018.

Pavone P., Pagliacci F., Russo M., Giorgi A (2018). Perspectives on RIS3s: a classification of priorities emerging from automatic text analysis. Deliverable prepared for the EU project "AlpGov", Work Package T-3.1 "Mapping the governance in the research and innovation field of the Eusalp regions", 31th December 2018

Russo M., Pagliacci F., Pavone P., Giorgi A., (forthcoming 2019). RIS3 in macro-regional strategies: tools to design and monitor integrated territorial development paths. Paper presented at the ESPON Scientific Conference “Building the Next Generation of Research on Territorial Development”, London, 14 November 2018. Conference Proceedings.

Annex 1 - Characteristic dictionaries of Codes related to sustainable development: Blue Economy, Bioeconomy & Waste collection, treatment, Energy Production, Efficiency & Sustainability

Codes with p-value less than 0.001 are listed in decreasing order of their test-value
Codes:

EcD: Economic Domains, NACE Rev. 2, two-digit codes
ScDom: Scientific Domain, NABS 2007, two-digit codes
PolObj: list of items created by JRC

Clusters' label: assigned by expert reading

Cl-6: 24 records, **Blue Economy**

Code	Label	Test-value
PolOb-B11	Fisheries	10,96
PolOb-B08	Aquaculture	10,79
PolOb-B14	Shipbuilding & ship repair	10,42
PolOb-B10	Coastal & maritime tourism	10,09
PolOb-B12	Marine biotechnology	9,88
PolOb-B15	Transport & logistics (incl highways of the seas)	9,27
PolOb-B09	Blue renewable energy	9,16
PolOb-B13	Offshore mining, oil & gas	8,79
ScDom-01_07	Sea and oceans	8,05
EcDom-A03	Fishing and aquaculture	7,18
EcDom-H50	Water transport	5,69

Cl-7: 157 records, **Bioeconomy & Waste collection, treatment etc**

Code	Label	Test-value
EcDom-E36	Water collection, treatment and supply	11,52
EcDom-E38	Waste collection, treatment and disposal activities; materials recovery	11,33
ScDom-02_14	Protection of soil and groundwater	11,02
ScDom-02_18	The elimination and prevention of pollution	10,51
EcDom-E39	Remediation activities and other waste management services	10,44
ScDom-02_12	Protection of ambient water	10,29
EcDom-E37	Sewerage	9,59
PolOb-J65	Resource efficiency	9,42
PolOb-J71	Waste management	9,11
ScDom-02_13	Protection of atmosphere and climate	8,99
ScDom-02_08	Monitoring facilities for measurement of pollution	8,87
ScDom-02_11	Protection of ambient air	8,83
PolOb-J69	Sustainable land & water use	8,74
PolOb-J63	Eco-innovations	8,66
PolOb-F45	Nature preservation	8,42
ScDom-06_40	Recycling waste	8,23
ScDom-02_17	Solid waste	8,18
PolOb-J70	Sustainable production & consumption	6,79
PolOb-J61	Bioeconomy	6,76
ScDom-02_10	Protection against natural hazards	6,64
EcDom-F41	Construction of buildings	6,59
ScDom-05_32	Energy efficiency	6,44
PolOb-J62	Climate change	6,18
ScDom-05_37	Renewable energy sources	5,94
ScDom-02_09	Noise and vibration	5,87
ScDom-02_15	Protection of species and habitats	5,80
EcDom-F43	Specialised construction activities	5,58
ScDom-05_31	Energy conservation	5,56
PolOb-J68	Sustainable energy & renewables	5,30
ScDom-04_24	Construction and planning of building	4,87
ScDom-01_05	Hydrology	4,51
EcDom-F42	Civil engineering	4,39
ScDom-05_30	CO2 capture and storage	4,28
PolOb-F43	Biodiversity	4,27
ScDom-12_101	Earth and related environmental sciences	4,19
ScDom-05_33	Energy production and distribution efficiency	4,11
ScDom-02_16	Radioactive pollution	4,10
ScDom-04_29	Water supply	3,98
ScDom-01_06	Mineral, oil and natural gas prospecting	3,60
ScDom-01_02	Climatic and meteorological research	3,37
EcDom-B09	Mining support service activities	3,37

Cl-8: 110 records, **Energy Production, Efficiency & Sustainability**

Code	Label	Test-value
EcDom-D35	Electricity, gas, steam and air conditioning supply	18,80
ScDom-05_33	Energy production and distribution efficiency	17,86

PolOb-J68	Sustainable energy & renewables	17,20
ScDom-05_37	Renewable energy sources	15,69
ScDom-05_32	Energy efficiency	14,29
ScDom-05_31	Energy conservation	13,82
ScDom-05_36	Other power and storage technologies	13,65
ScDom-05_34	Hydrogen and fuel gas	12,06
ScDom-05_35	Nuclear fission and fusion	6,92
ScDom-05_30	CO2 capture and storage	6,87
PolOb-D22	Cleaner environment & efficient energy networks and low energy computing	5,19
PolOb-B09	Blue renewable energy	4,75
PolOb-J65	Resource efficiency	4,57
PolOb-J63	Eco-innovations	4,07
EcDom-F43	Specialised construction activities	4,03
PolOb-J62	Climate change	3,76
PolOb-J66	Smart green & integrated transport systems	3,28

Annex 2 - Free texts descriptions by code category and description category

The following are some of the examples cited in the paragraph *Cross-Tabulation* (source: Eye@RIS3, download 1st October 2018)

Code category: Blue Economy

Description: Agrofood

EL42 (Notio Aigaio). Fisheries and aquaculture. Emphasis will be placed on product differentiation, biotechnological applications, links with tourism, biodiversity, quality and certification management, logistics, new methods of processing and preservation (non-thermal), networks and marketing.

Description: Automotive & Aerospace

PL62 (Warminsko-Mazurskie) - Water economy. Transport, sports, manufacturing, tourism, food, machinery, yachts, environment.

Code category: Bio economy

Description: Tourism

ES53 (Illes Balears) - Sustainable Tourism. To promote excellence in tourism related firms and extend the image of sustainable tourism of the Balearic Islands. Also to improve the design, development and commercialization of advanced services and sustainability technologies.

Description: Health

ES70 (Canarias) - Biotechnology

PL21 (Malopolskie). Life sciences. The mix of two value chains: health and quality of life which include products and technologies used in the prevention, diagnosis, treatment and rehabilitation of human and animal diseases and bio-economy comprising semi-finished products and products used in the production of pharmaceuticals, cosmetics, food, materials and energy.

Description: Agrofood

PL42 (Zachodniopomorskie). Eco-friendly packaging. Maximising the biodegradability, flexible and energy efficiency of packaging materials, packaging with nanocomposites, materials with increased external barrier properties, smart, safe and active packaging, design attractiveness of products, packaging ensuring greater food safety and longer shelf life, use of bio-based raw materials in packaging production, circular management of packaging.

Description: Manufacturing

PL21 (Malopolskie). Chemical industry. Programmes to implement new compounds, materials and chemical technologies, including chemical engineering solutions, in areas (9 domains) related to health care, agriculture, food, wood, pulp and paper industries, biological and environmental chemistry, energy, raw materials, waste management, materials for construction and transport, advanced materials and nanotechnologies, sensors.

PL42 (Zachodniopomorskie). Chemical and materials engineering products. Production of standardised materials, products and semi-finished chemical products (including organic and mineral fertilisers) and chemical processing and specialty chemicals, waste management and biomass production, in particular in the context of the use of renewable energy sources.PL21 (Malopolskie). Chemical industry. Programmes to implement new compounds, materials and chemical technologies, including chemical engineering solutions, in areas (9 domains) related to health care, agriculture, food, wood, pulp and paper industries, biological and environmental chemistry, energy, raw materials, waste management, materials for construction and transport, advanced materials and nanotechnologies, sensors.

Annex 3 - Multi layer clustering with Infomap

As a possible mechanism of multidimensional clustering, we considered each of the three classifications as layers of a multi-layer network. For each layer, a node (i.e., a region) is connected to all other regions with the same classification. We then ran extensive attempts to cluster the results based on Infomap Multilayer (De Domenico *et al.*, 2015). Infomap is a method, based on information theory, to detect communities in complex networks by "minimizing the description length of a random walker's movements on a network" (Bohlin *et al*. 2014). However, using this algorithm on the network structure generated by our classifications, the algorithm returns a single community, comprising all nodes. This result is due to the excessive density of the network in some layers, which prevents the identification of separate communities. Indeed, with regard to the layers "codes" and "descriptions", density is, respectively, 0.94 and 0.82. Transforming clusters identified through the multilayer network analysis in interconnected cliques produces dense networks, that are difficult to exploit for community detection through network-based procedures. This method turns out to be unsuitable for this particular application.