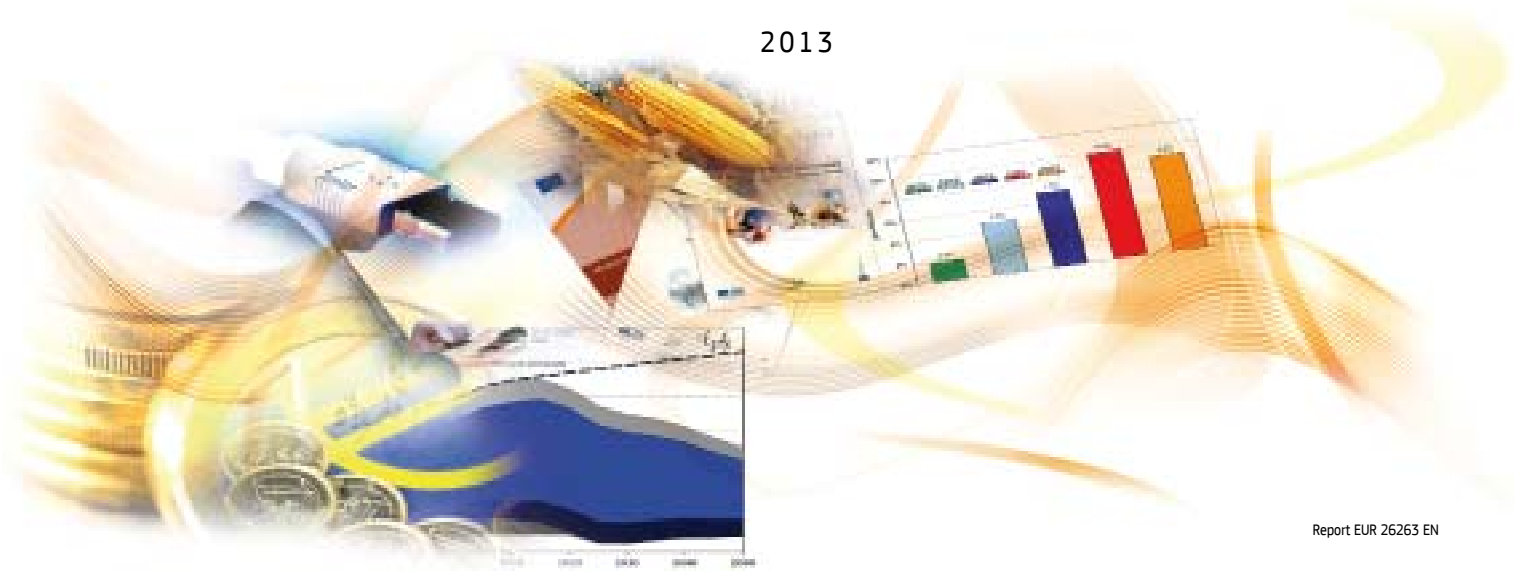
European Commission

JRC SCIENTIFIC AND POLICY REPORTS

A classification of European NUTS3 regions

Meri Raggi, Sébastien Mary, Fabien Santini, Sergio Gomez Y Paloma

2013



Report EUR 26263 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for Prospective Technological Studies

Contact information
Address: Edificio Expo. c/ Inca Garcilaso, 3. E-41092 Seville (Spain)
E-mail: jrc-ipts-secretariat@ec.europa.eu
Tel.: +34 954488318
Fax: +34 954488300

http://ipts.jrc.ec.europa.eu
http://www.jrc.ec.europa.eu

JRC85163

EUR 26263 EN

ISBN 978-92-79-34483-1 (pdf)

ISSN 1831-9424 (online)

doi:10.2791/35200

Luxembourg: Publications Office of the European Union, 2013

Printed in Spain

A CLASSIFICATION OF EUROPEAN NUTS3 REGIONS

Meri Raggi[a], Sébastien Mary[b], Fabien Santini[b], Sergio Gomez y Paloma[b]

[a] University of Bologna, Department of Statistical Sciences, Italy

[b] European Commission, Joint Research Centre, Institute for Prospective and Technological Studies, Spain

2013

MOdelling Rural Economies (MORE)

Table of contents

List of figures

List of Tables

CHAPTER 1: CLASSIFICATION OF NUTS3 REGIONS

1. OBJECTIVE OF CHAPTER 1

Over the years a number of Common Agricultural Policy (CAP) reforms have led to the emergence of a CAP chapter specifically dedicated to rural development, also referred as Pillar 2, and have resulted in a progressive switch of CAP budget from Pillar 1 (i.e. direct support to farmers, including direct payments and other instruments for market regulation) to Pillar 2. Approximately 23 per cent of the CAP should be allocated to rural development measures during the period 2014-2020. The recent development of Pillar 2 calls for further research on the impact assessment of such policies. Unfortunately, the diversity of rural situations across Europe has complicated the empirical studies of the impacts of rural development and often makes any comparison between regions rather trivial.

The main objective of TASK 1 is the creation of a classification of 1303 NUTS3 regions, which reflects the heterogeneity of NUTS3 characteristics in the EU. This classification will be multidimensional. In particular, the typology will be based on the following set of four criteria: Rural Character, Accessibility, Actual economic diversification and Total Gross Domestic Product per capita. Such classification will facilitate the comparison of rural development policy impacts between regions of interest across Europe.

In the report we consider three different approaches, i.e. traditional cluster analysis, latent class models and multiple cluster structures. All approaches are presented in Section 2 with a description of their underlying assumptions and specific features. Then, Section 3 assesses the adequateness of each approach with respect to the data used for the analysis. Section 4 examines the results for each approach and some concluding remarks are provided in Section 5.

2. METHODOLOGY

2.1. "Traditional" cluster analysis

Cluster Analysis (CA) covers a rather wide collection of statistical methods that can be used to assign cases, i.e. records or units (here, NUTS3 regions), to groups (clusters) that are mutually exclusive. Group members will share some properties in common, so that the degree of associations is strong between cases of the same clusters and weak between cases of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong. The resultant classification can then provide some insights and help for the interpretation of a research topic because it may reveal associations and the structure of the data, which, in turn, may contribute to the definition of formal classification schemes or even suggest models with which to describe a population.

An effect of the classification could be the reduction of dimensionality of a database by reducing the row number (cases). The groupings produced by the analysis may (or not) prove useful for classifying cases: if the groupings discriminate between variables that are not used to implement the CA and those discriminations are useful, then results from the CA are useful.

There are many options one may select when using CA: hierarchical or non-hierarchical methods, divisive or agglomerative techniques, different distance measures (Euclidean, Manhattan ...). Generally, the classification obtained through analysis is dependent upon the particular algorithm used in the process; consequently, there is no such thing as a single correct classification, although there have been attempts to define concepts such as "optimal classification".

The first step is to decide which clustering variables will be included in the analysis; one should avoid using an abundance of clustering variables, as this increases the probability that they are no longer dissimilar. If there is a high degree of collinearity between clustering variables, they cannot have the power to discriminate between groups. For instance, if highly correlated variables (say around 0.90) are considered for CA, specific aspects covered by these variables could be overrepresented in the cluster results.

Choosing an appropriate clustering method is the second critical step in CA. Some algorithms are strictly related to the nature of variables, e.g. when a distance measure is necessary, only quantitative variables can be used for creating such measure. Alternatively, if one has a mixture of nominal and continuous variables, one must use the two-step cluster procedure because none of the distance measures used in hierarchical clustering or K-Means are suitable in that particular situation. Whichever clustering variables and algorithm are chosen, it is important to assess the validity of the analysis. The criterion validity is related to the algorithm used, nevertheless there should at least be significant differences between the clustering variables across the resultant clusters and relative small differences within clusters.

2.2. Latent class models

Latent class analysis could be thought of as an “improved” cluster analysis, which uses statistical (rather than mathematical) methodology to construct the results. In fact, traditional clustering approaches use classification algorithms that group cases together that are "near" to each other according to an *ad hoc* definition of "distance". Over the last decade, attention has shifted towards model-based approaches, especially mixture model clustering where each latent class represents a hidden cluster. Latent class models are particularly appropriate to include not only continuous variables, but also variables that are ordinal, nominal or counts, or any combination of these.

They are based on the statistical concept of likelihood, i.e. results are obtained via maximizing a log-likelihood (LL) function. The main difference is that cases are not absolutely assigned to classes; instead, using a model-based probability, they have a probability of membership for each class (see Vermunt and Magidson, 2002). There are several advantages of using a statistical model. Especially, it allows for the use of categorical variables into the analysis and the choice of the cluster criterion is less arbitrary than in traditional CA.

Moreover, the approach allows performing rigorous statistical tests in order to assess hypotheses on model parameters. For example, these tests can be exploited to detect significant differences in the distribution of variables among groups.

2.3. Multilevel latent class models

In some fields of research such as social science, medicine or economics, datasets frequently present a multilevel structure. That is, two (or more) nested levels of units are present in the data and lower level units belong to higher level units. Examples of two-level data are observations from individuals living in regions of a same country, patients hospitalized in wards of a same hospital and employees from teams of a same organisation. As the parameters of a latent class model are assumed to be the same for all level-1 units (e.g., individuals, patients, employees), when a multilevel dataset is analysed using clustering methods based on latent class models, the multilevel nature of the phenomenon will be ignored. In order to take account the multilevel structure, some of the parameters should be allowed to differ across level-2 units (regions, wards, teams).

Vermunt (2003, 2008) propose a multilevel extension of latent class models in which the dependence between observations within higher-level units is dealt with by assuming that certain model parameters differ randomly across higher-level observations. In particular, his extension is based on the introduction of a discrete random effect; this yields a latent class model in which there are not only groups of level-1 units, but also groups of level-2 units sharing the same parameter values. These multi-level models allow for the computation of cluster membership probabilities for units of both levels. Furthermore, using the approximated Bayesian rule, level-1 and level-2 units can respectively be partitioned into level-1 and level-2 clusters.

3. VARIABLES AND DATA

The classification will be based on 4 variables/indicators that are described below. All indicators are considered at NUTS3 level from the updated version of Regulation 105/2007 (2006), meaning 1303 NUTS3 regions are included in the analysis.

3.1. Eurostat classification of rural/urban typology

The variable is obtained from the revised Eurostat classification that presents 3 typologies: predominantly urban, intermediate and predominantly rural. The definition is based on per km^2 grid cells. The steps of the NUTS3 classification[1] are:

1. It creates clusters of urban grid cells with a minimum population density of 300 inhabitants per km^2 and a minimum population of 5 000. All cells outside these urban clusters are considered as rural.

2. It groups NUTS 3 regions of less than 500 km^2 with one or more of its neighbours solely for classification purposes, i.e. all the NUTS 3 regions in a grouping are classified in the same way.

3. It classifies NUTS 3 regions based on the share of population in rural grid cells. More than 50 % of the total population in rural grid cells = predominantly rural, between 20 % and 50 % in rural grid cells = intermediate and less than 20 % = predominantly urban.

Results from this classification are presented in Table 1 and Figure 1.

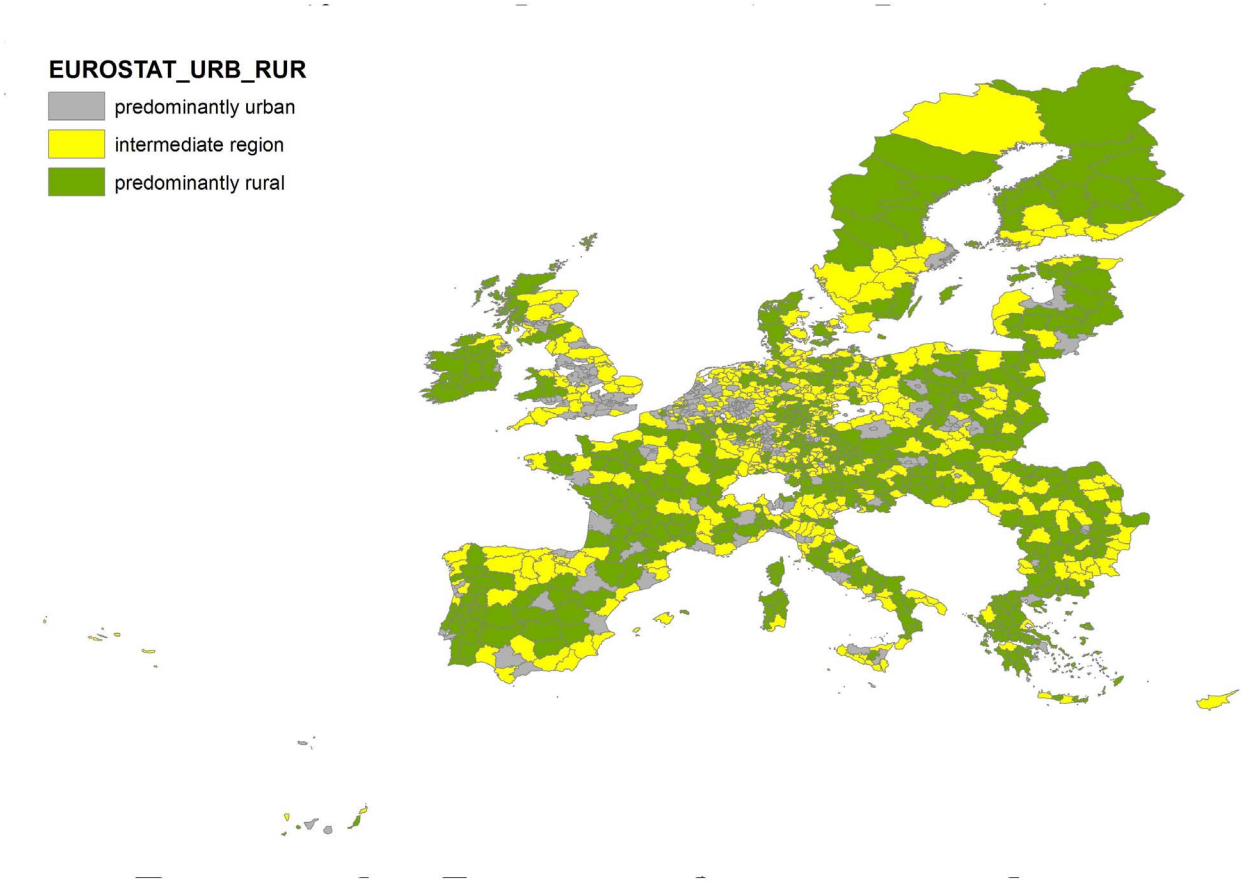
Table 1: NUTS3 classification by urban/rural predominance

Eurostat classification	**Frequency (number of NUTS3 regions)**	**Percentage (%)**
1: Predominantly urban	308	23,6
2: Intermediate region	494	37,9
3: Predominantly rural	501	38,4
Total	**1303**	**100,0**

Source: Eurostat

[1] Further details can be found at http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Urban-rural_typology.

Figure 1: Geographical distribution of NUTS3 regions by urban/rural character

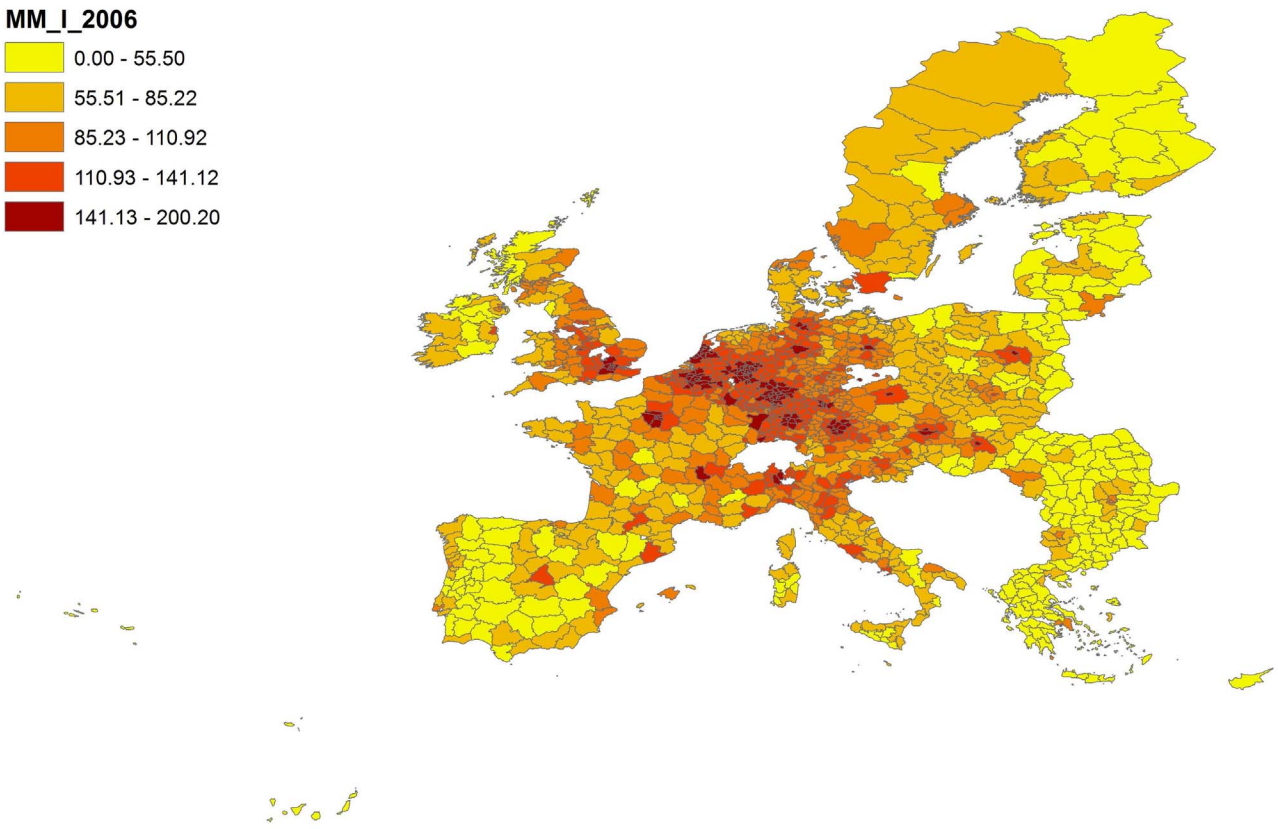


Source: Eurostat

3.2. Accessibility

The indicator used is the Multimodal potential accessibility (mm_i) standardised with the EU average, derived from the ESPON 2013 database, available at: www.espon.eu/main/Menu_ToolsandMaps/ESPON2013Database). The concept of potential accessibility enables to measure how easy (i.e. travel time) a region can be reached from other European regions by a given means of transportation. High accessibility is often considered a prerequisite for economic development, for attracting investors, increasing employment and building networks of cities (www.espon.eu). Figure 2 presents the geographical distribution of NUTS3 regions according to the index described just above.

Figure 2: Geographical distribution of NUTS3 regions by accessibility



Source: Espon

3.3. Actual economic diversification

Actual economic diversification is determined by the first digit of the diversification_economy_agriculture variable obtained in the TERA-SIAP project[2]. It describes the relative importance of agriculture in the regional economy (indicators: primary sector GVA, agricultural employment). The "actual economic diversification" includes two components: (i) overall economic diversification, measured by the relative importance of agriculture in the regional economy (using indicators: primary sector GVA, agricultural employment); and (ii) farm diversification and agricultural pluriactivity, measured by the incidence of other gainful activities (Weingarten et al.,

[2] The technical report of the TERA-SIAP project is available at: http://ftp.jrc.es/EURdoc/JRC58493.pdf.

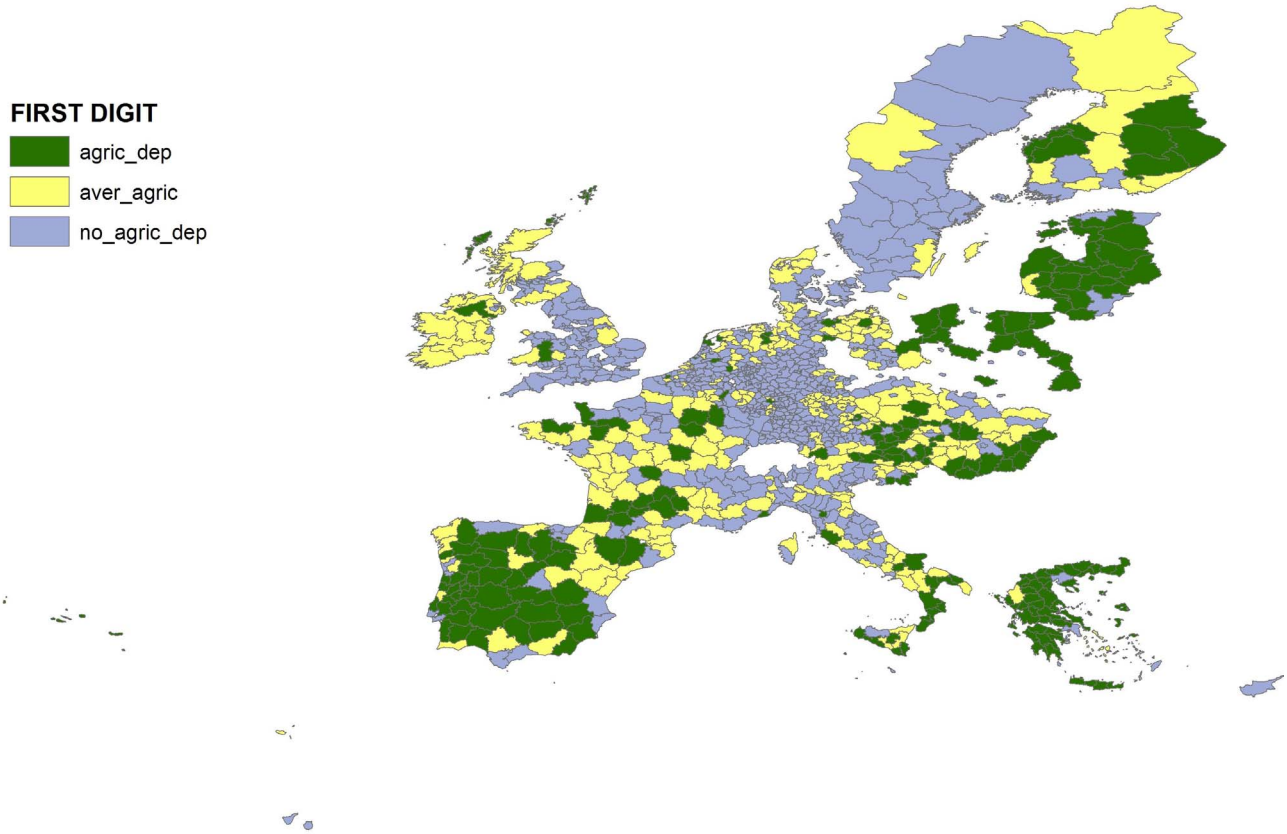
2010). Figure 3 maps the distribution of NUTS3 regions according to actual economic diversification.

Table 2: NUTS3 classification by actual economic diversification

TERA-SIAP classification	**Frequency (number of NUTS3 regions)**	**Percentage (%)**
1: Importance of agriculture above average	216	16.6
2: Average importance of agriculture	255	19.6
3: Importance of agriculture below average	670	51.4
Missing value	162	12.4
Total	**1303**	**100**

Source: TERA-SIAP

Figure 3: Geographical distribution of NUTS3 regions by actual economic diversification



Source: TERA-SIAP

3.4. Total Gross Domestic Product

The GDP per capita is at current market prices, 2009 EUR. (http://epp.eurostat.ec.europa.eu/portal/page/portal/national_accounts/data/database). Since the presence of outliers and the strong asymmetry of the distribution (Figure 4) we use the logarithmic transformation (ln_GDP) that makes patterns more visible (at least 2 patterns) in Figure 5.

Figure 4: Frequency distribution of NUTS3 regions by GDP

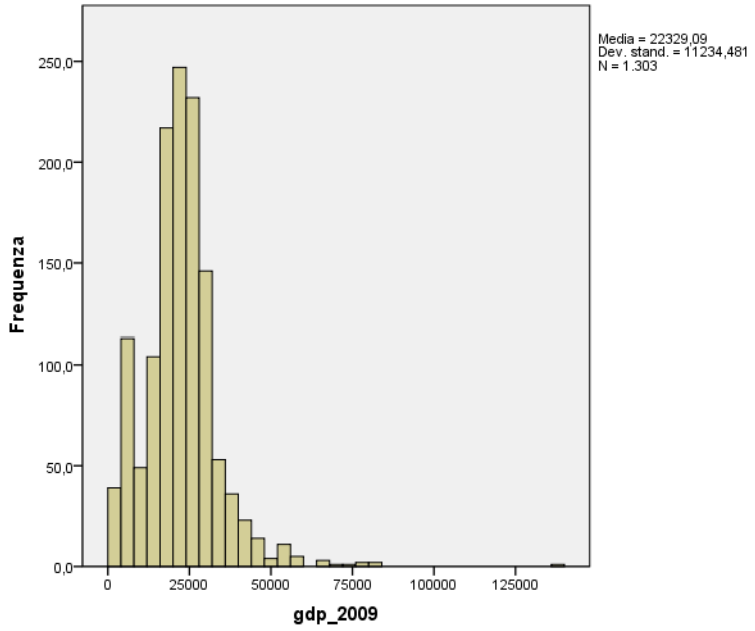


Figure 5: Frequency distribution of NUTS3 regions by logarithm of GDP (ln_GDP)

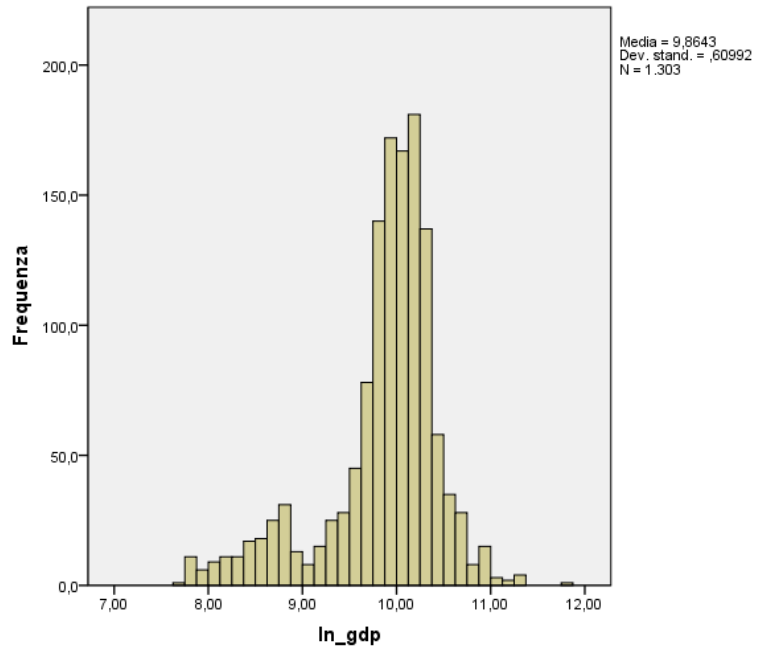
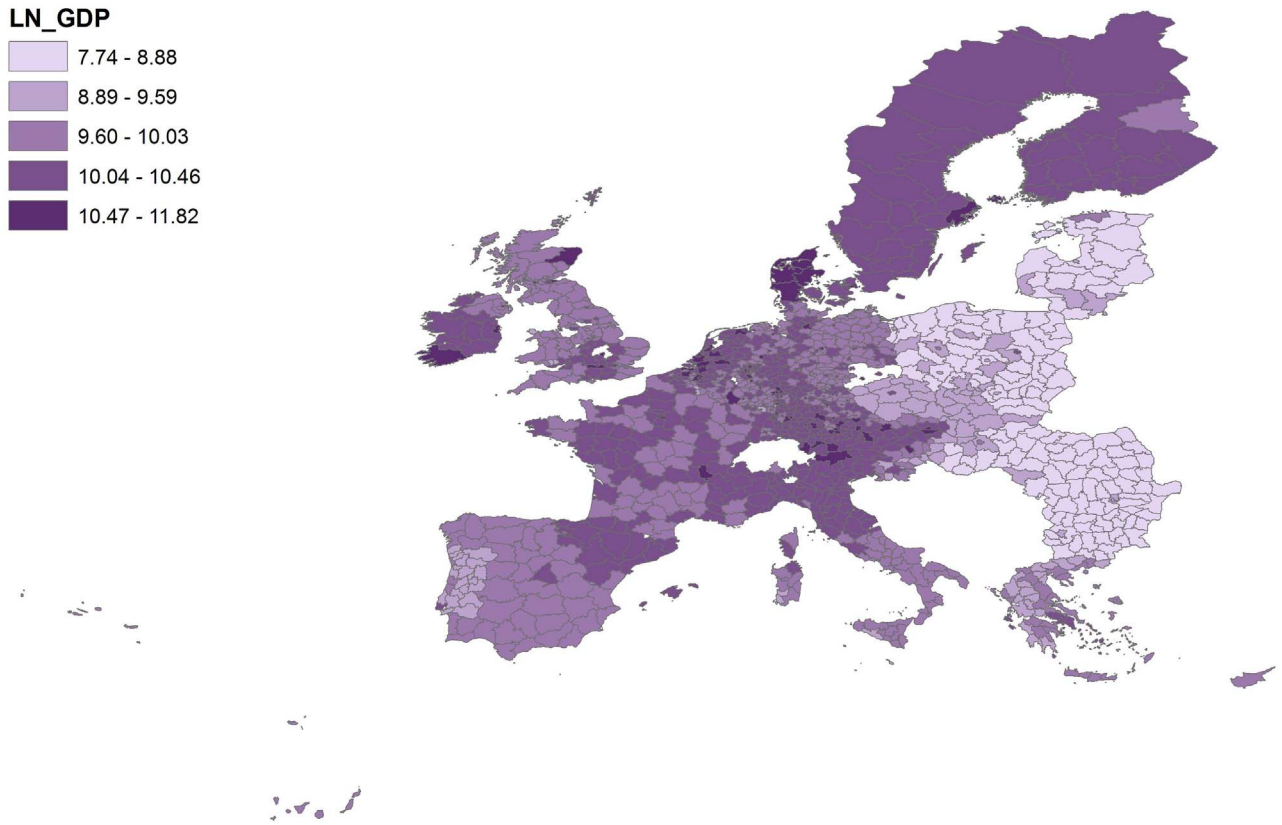


Figure 6: Geographical distribution of NUTS3 regions by ln_GDP



Source: Eurostat

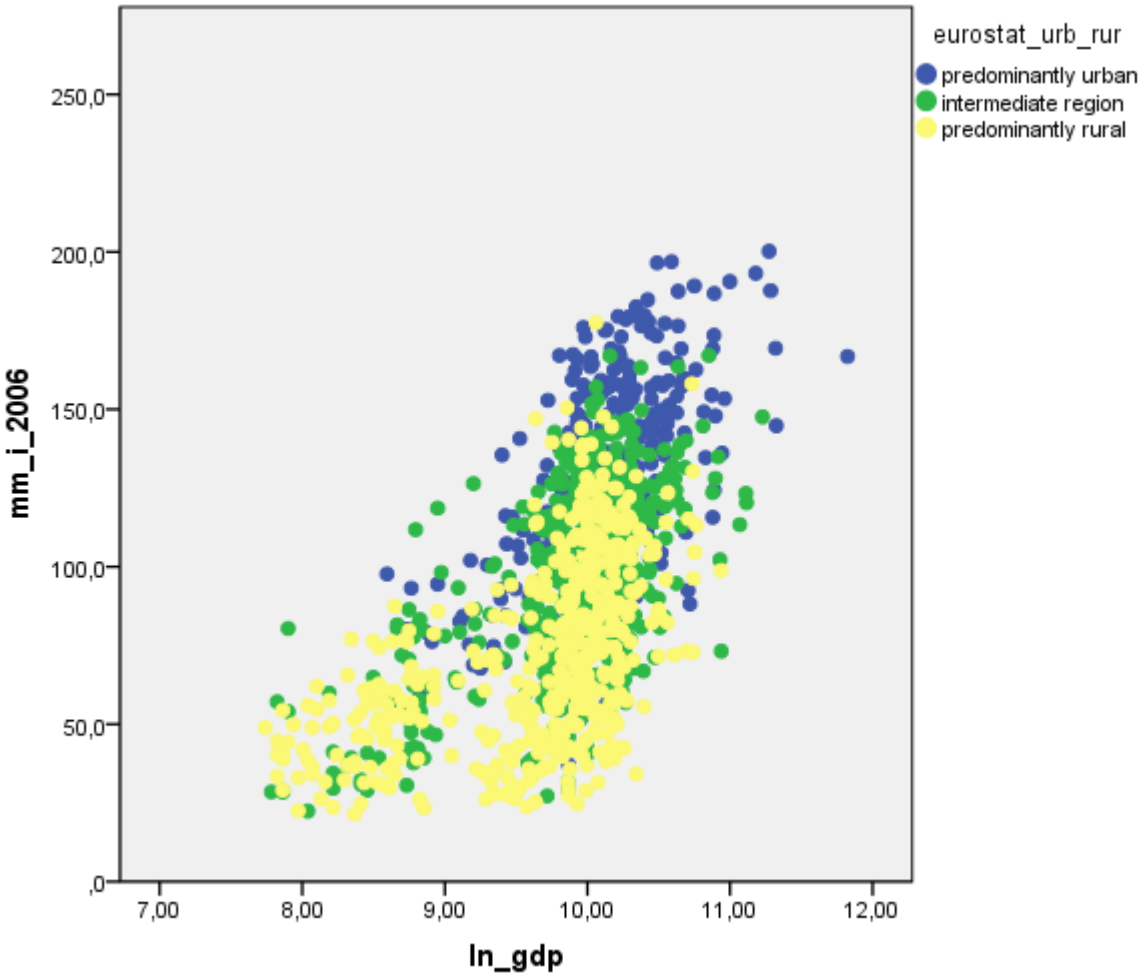
3.5. Relationship among variables

In this paragraph, a preliminary diagrammatic analysis considering the relationships among the variables is considered, with the aim of having a first glance at potential patterns in the data. The correlation between the two quantitative variables (GDP and accessibility) is 0.596. It is relatively high but not high enough to cause problems in the cluster analysis in terms of lack of significance of variables. The degree of association between the two qualitative variables is analysed in Table 3. Some cells show a strong association, but this is not unique (i.e. this frequency is different from zero only in one cell per row and per column) and all cells have frequency (empty cells indicate an impossible combination).

Table 3: Cross-tabulation frequency between urban/rural area and economic diversification

	Economic diversification (first_digit)			
Eurostat urban/rural	1: Agric_dep	2: Aver_agric	3: No_agric_dep	Missing value
1: Predominantly urban	4	12	264	28
2: Intermediate region	41	93	302	58
3: Predominantly rural	171	150	104	76

Figure 7: Scatterplot by accessibility (mm_i_2006), GDP and urban/rural (eurostat_urb_rur)



The following figures aim at highlighting the data structure in relation to potential clear patterns/groups. All graphs show the presence of clusters with similar variable features. For example, in Figure 7, NUTS3 regions classified as urban (in blue) are mostly grouped in the top-right corner, meaning they combine a high accessibility index and high GDP. Another example is displayed in Figure 8, where NUTS3 regions, the economies of which are strongly dependent from the agricultural sector (in yellow), are grouped in the bottom-left corner showing an association of low accessibility and low GDP.

Figure 8: Scatterplot by accessibility (mm_i_2006), GDP and dependence on agriculture (first digit)

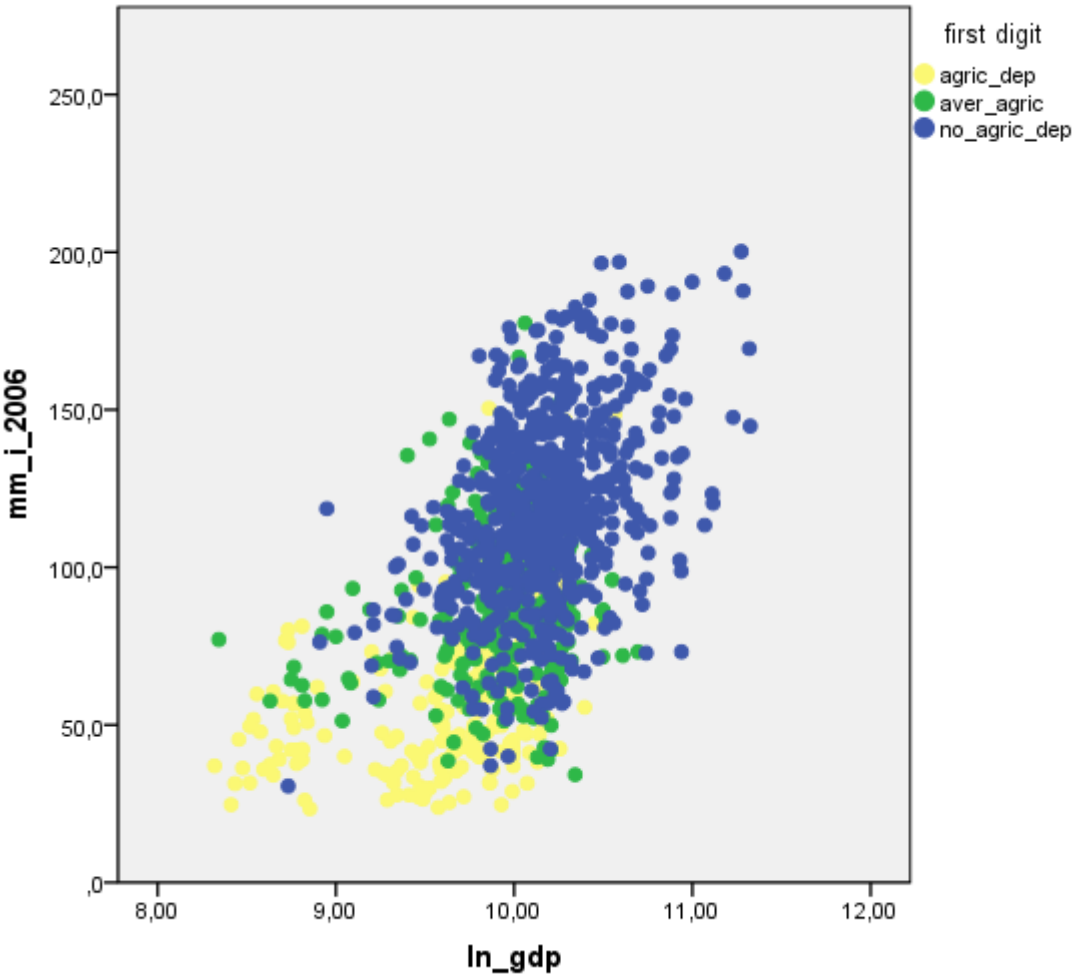


Figure 9: Scatterplots and linear trendlines of accessibility (mm_i_2006) and ln_GDP, stratified by urban/rural (eurostat_urb_rur) and Economic diversification (first_digit).

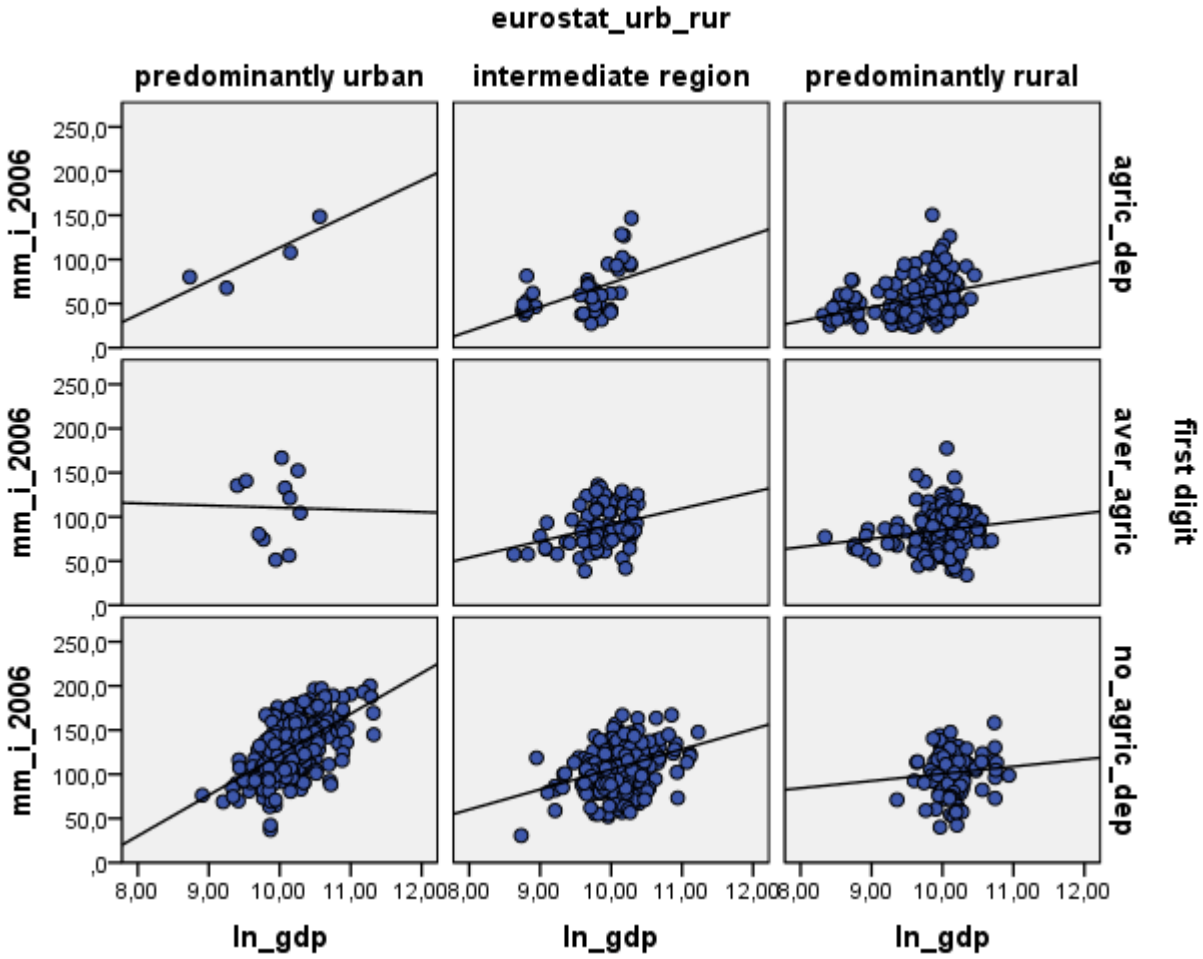


Figure 9 allows highlighting some interesting features with respect to the relationships among the four variables. Each panel in Figure 9 refers to regions characterised by a specific combination of both values of accessibility, i.e. urban/rural (eurostat_urb_rur), and economic diversification (first_digit) and contains the scatterplot for accessibility and ln_GDP, with a linear trend. First, it is worth noting that the linear trend between ln_GDP and accessibility seems to be stronger and positive when considering NUTS3 regions that are intermediate or predominantly urban (according to the EUROSTAT typology) and in which agriculture has a lower importance in terms of economic diversification (see bottom left and bottom middle panels).

Furthermore, the presence of two subgroups of NUTS3 regions that are intermediate or predominantly rural and have a strong dependence on agriculture (i.e. top right and top middle panels) is visually evident. These two subgroups are characterised by different average values for ln_GDP. This second feature is strictly related to the patterns already observed in Figure 5 (e.g. bimodal distribution). In particular, it suggests that the lowest mode (between 8 and 9) in Figure 5 corresponds to only a subset of the NUTS3 regions in which agriculture has high importance in terms of economic diversification. In other words, there are many rural regions that show a large value for ln_GDP (close to the one for urban regions).

4. RESULTS

4.1. Cluster analysis: a very rough analysis

The set of variables that we consider includes two quantitative variables (gdp and accessibility) and two qualitative variables (rural/urban character and economic diversification). The traditional cluster analysis is only feasible with quantitative variables, since they are based on the calculation of a distance matrix. A very rough solution to obtain a preliminary analysis of our variables is to recode and to interpret the two qualitative variables as ordinal ones:

- Proxy of urban index: 0=rural; 1= intermediate; 2=urban
- Proxy of economic development: 0=agriculture dependent; 1= agriculture average; 2= no dependence

Clearly, results can only give an approximation of the complexity of the situation. In the following tables, we report the output obtained with K-Means clustering, which accounts for 5 clusters. This algorithm assigns cases to clusters based on the smallest amount of distance between the cluster mean and each case. This is an iterative process that stops once the cluster means do not significantly change in successive steps (with stopping criteria, maximum center variation less than 0.05). The output of K-Means is provided in the following tables.

Table 4: Final cluster centers

	Cluster				
	1	2	3	4	5
proxi_urb	0.29	0.91	1.31	0.67	1.87
proxi_indus	0.45	1.74	1.86	1.25	1.97
mm_i_2006	47.9	108.7	134.2	81.0	167.1
ln_GDP	9.61	10.09	10.22	9.92	10.39

Table 4 shows final cluster centers: for quantitative variables, values represent the correct average in each cluster, while for ordinal variables averages are less meaningful since the variables could take only integer values. In Table 5, the number of cases in each cluster is reported. With the exception of cluster 5, the remaining clusters are more or less the same size.

Even if all variables have significant power to discriminate groups (Table 6), another problem in the analysis is shown in Table 7; that is, there is not a clear cut division between the two ordinal variables in the clusters, i.e. there is not a unique association of ordinal variable levels and clusters. In other words, this "rough" analysis demonstrates that even though all variables are significant enough to discriminate among clusters, the classification results do not illustrate a clear cut division between variables because the latter show overlapped features for some clusters.

Table 5: Number of cases in each cluster

Cluster	1	221
	2	325
	3	208
	4	295
	5	87
Valid		1136
Missing		167

Table 6: Test ANOVA for discriminant variables

		Sum of squares	df	Mean square	F	Sig.
proxi_urb	Between groups	214.434	4	53.609	129.538	0.000
	Within groups	468.058	1131	.414		
	Total	682.492	1135			
proxi_indus	Between groups	312.559	4	78.140	226.604	0.000
	Within groups	390.001	1131	.345		
	Total	702.560	1135			
mm_i_2006	Between groups	1366007.179	4	341501.795	4106.800	0.000
	Within groups	94048.539	1131	83.155		
	Total	1460055.718	1135			
ln_GDP	Between groups	61.788	4	15.447	112.359	0.000
	Within groups	155.490	1131	.137		
	Total	217.278	1135			

Table 7: Frequency in each cluster of ordinal variable codes (K-Means algorithm)

		Cluster				
		1	2	3	4	5
proxi_urb	0	162	99	26	134	2
	1	54	157	92	125	7
	2	5	69	90	36	78
proxi_indus	0	148	12	6	49	0
	1	46	62	18	123	3
	2	27	251	184	123	84

4.2. Cluster analysis: SPSS TwoStep Cluster

To deal with the problems raised in the K-Means algorithm, we carry out a SPSS TwoStep Cluster Analysis. This algorithm can produce solutions using mixtures of continuous and categorical variables. The clustering algorithm is based on a distance measure that provides the best results if:

- All variables are independent,
- All variables are continuous variables and have a normal distribution,
- Categorical variables have a multinomial distribution.

The SPSS algorithm provides an optimal automatic number of clusters, yet, since cluster analysis does not involve hypothesis testing, the task of checking whether the solution is satisfactory is left to the researcher. Various measures can be used to quantify the goodness of a cluster solution.

In Figure 10 the model summary shows that the number of automatic optimal clusters is 6 and the quality is good (see cluster quality bar). This bar represents the silhouette coefficient, which is a measure of both cohesion (i.e. how elements within a cluster are similar to one) and separation (i.e. how clusters themselves are quite different) and ranges from –1 to 1.

Figure 10: Model summary and cluster quality (TwoStep algorithm)

In our results, the silhouette measure is approximately 0.6 indicating that the cluster solution is quite good. Nevertheless, it is necessary to further examine the composition of clusters and the importance of each variable in the grouping procedure. Figure 11 illustrates the percentage frequency of each cluster, clusters 6 and 5 are the largest with 26% and 23% of NUTS3 regions, respectively, followed by cluster 1 with 15% and clusters 2 and 3 with almost the same size (13%). The smallest is cluster 4 which includes only 9% of NUTS3 regions.

Figure 11: Cluster size (TwoStep algorithm)

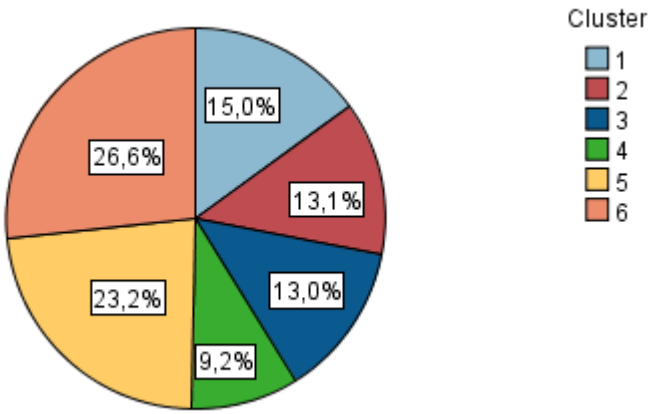
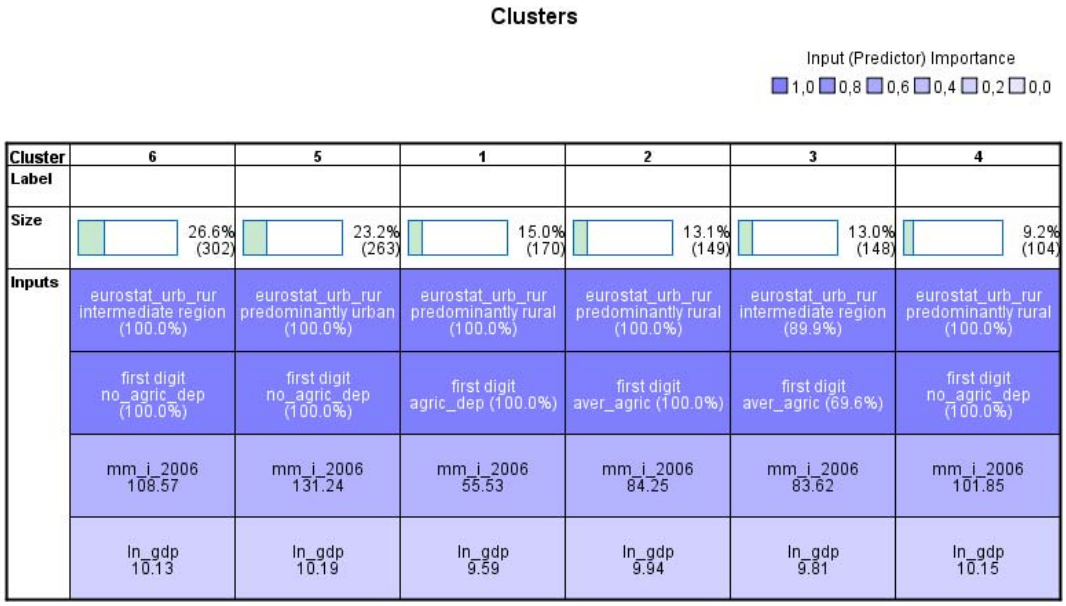


Figure 12: Feature importance (TwoStep algorithm)



The output[3] that describes the importance of each variable is shown in Figure 12. The first warning sign is that the two qualitative variables are the ones with the greatest importance (in blue) as demonstrated by the percentage in brackets. To better understand the output of this analysis and how the NUTS3 regions are classified in clusters, the next step is to analyse the composition of each cluster using cross tabulations or bar charts of the distribution of variables within each cluster.

Table 8: Frequency in each cluster related to combination of qualitative variable codes (TwoStep algorithm)

Cluster				
		1: predominantly urban	2: intermediate region	3: predominantly rural
1	1: agric_dep			170
	2: aver_agric			
	3: no_agric_dep			
2	1: agric_dep			
	2: aver_agric			149
	3: no_agric_dep			
3	1: agric_dep	4	41	
	2: aver_agric	11	92	
	3: no_agric_dep			
4	1: agric_dep			
	2: aver_agric			
	3: no_agric_dep			104
5	1: agric_dep			
	2: aver_agric			
	3: no_agric_dep	263		
6	1: agric_dep			
	2: aver_agric			
	3: no_agric_dep		302	

[3] Results are obtained via SPSS.

According to Bacher *et al.* (2004), "Summarizing the results of the simulations, SPSS TwoStep performs well if all variables are continuous. The results are less satisfactory if the variables are of mixed type. One reason for this unsatisfactory finding is the fact that differences in categorical variables are given a higher weight than differences in continuous variables. Different combinations of the categorical variables can dominate the results. In addition, SPSS TwoStep clustering is not able to correctly detect models with no cluster solutions. Latent class models show a better performance. They are able to detect models with no underlying cluster structure, they result more frequently in correct decisions and in less unbiased estimators". This statement largely applies to our analysis; given the frequencies obtained above, it is quite obvious that the grouping obtained is a simple combination of code of qualitative variables (with the exception of cluster 3). In fact, Table 8 clearly shows that some clusters are simply the combination of two levels of ordinal variables, for example cluster 1 is formed by all NUTS3 that are predominantly rural and agriculturally dependent. Finally, the TwoStep algorithm does not offer a satisfactory grouping result for our data.

4.3. Latent class models

The most important types of latent class and finite mixture models, along with their multilevel extensions[4], are particularly suited for analysing data sets containing indicators of different scale types. For each indicator, the user must specify whether it is nominal, ordinal, continuous, or a count. The parameters of the various types of models are estimated by means of Maximum Likelihood (ML) or Posterior Mode (PM) methods. The likelihood function is derived from the probability density function of the selected (multilevel) latent class model. PM methods are implemented in order to overcome possible technical problems that may arise with the ML method.

The output of our model is presented in Table 9 and Table 10. Table 9 summarizes all the models estimated on our data and helps with the model selection, i.e. the optimal number of clusters is 5 groups with the best solution being the one with the highest

[4] Latent GOLD 4.0 is a commercial software package available from Statistical Innovations Inc. that implements such models.

likelihood (which assesses how well the model fits the data expressed by the lowest Bayesian Information Criteria).

Table 9: Model summary (by latent class)

Classification		**LL**	**BIC(LL)**	**Npar**	**Class.Err.**
Model 1	1-Cluster	-9729.62	19552.48	13	0
Model 2	2-Cluster	-9206.42	18577.81	23	0.048
Model 3	3-Cluster	-9021.41	18279.51	33	0.1066
Model 4	4-Cluster	-8938.03	18184.48	43	0.1025
Model 5	**5-Cluster**	**-8890.68**	**18161.51**	**53**	**0.1364**
Model 6	6-Cluster	-8857.24	18166.35	63	0.1362
Model 7	7-Cluster	-8836.64	18196.86	73	0.1554

Table 10: Cluster profiles (by latent class)

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**
Cluster Size	0.302	0.256	0.210	0.134	0.098
Indicators					
ln_GDP (Mean)	10.008	10.181	10.123	8.617	9.748
mm_i_2006 (Mean)	87.183	111.777	133.944	54.346	46.756
First digit=1 (agric_dep)	0.125	0.002	0.007	0.646	0.995
First digit=2 (aver_agric)	0.596	0.001	0.029	0.307	0.003
First digit=3 (no_agric_dep)	0.279	0.997	0.964	0.047	0.002
eurostat_urb_rur=1 (predominantly urban)	0.039	0.116	0.909	0.019	0.0004
eurostat_urb_rur=2 (intermediate region)	0.407	0.676	0.090	0.312	0.156
eurostat_urb_rur=3 (predominantly rural)	0.554	0.209	0.001	0.67	0.843

Table 10 shows the cluster profiles. Overall, cluster 1 contains 30.2% of the cases, cluster 2 contains 25.6%, cluster 3 contains 21%, cluster 4 contains 13.4% and the remaining 9.8% are in cluster 5. Conditional probabilities show the differences in response patterns that distinguish the clusters. For example, cluster 5 is much more likely to include NUTS3 regions that are predominantly rural (0.843), in which economies are strongly dependent from agriculture, with low accessibility and low GDP.

On the contrary, Cluster 3 is characterised by NUTS3 regions that are predominantly urban (0.909), with an economy that is not agriculturally-dependent (0.964), with high GDP (10.123) and high accessibility (133.944). Cluster 2 includes NUTS3 regions classified as intermediate rural/urban, an importance of agriculture below average, high GDP and good accessibility. Cluster 4 contains NUTS3 regions with the lowest GDP, sufficient accessibility, predominantly rural area with an economy strongly or moderately dependent on agriculture. Cluster 1 is not clearly defined, as it includes mainly NUTS3 regions classified as predominantly rural but also intermediate regions, with average agricultural dependency, intermediate accessibility and high GDP. The graphical distribution of each variable in each cluster is presented in the following figures. Figure 13 presents how urban/rural areas combine with economic diversification in clusters. Figure 14 and Figure 15 display box plots for single cluster respectively for accessibility (mm_i_2006) and GDP (ln_GDP) variables. The two box plots show a clear divide among clusters since boxes and buffers are not overlapping. The cluster composition considering the two qualitative variables is not straightforward (figure 13). In fact some clusters include a mix of NUTS3 typology (for example cluster 1 include NUTS3 predominantly urban, intermediate region and predominantly rural associated with agricultural dependence, average agricultural dependence and no agricultural dependence).

Figure 13: Cluster composition by economic diversification (first_digit) and urban/rural classification (eurostat_urb_rur)

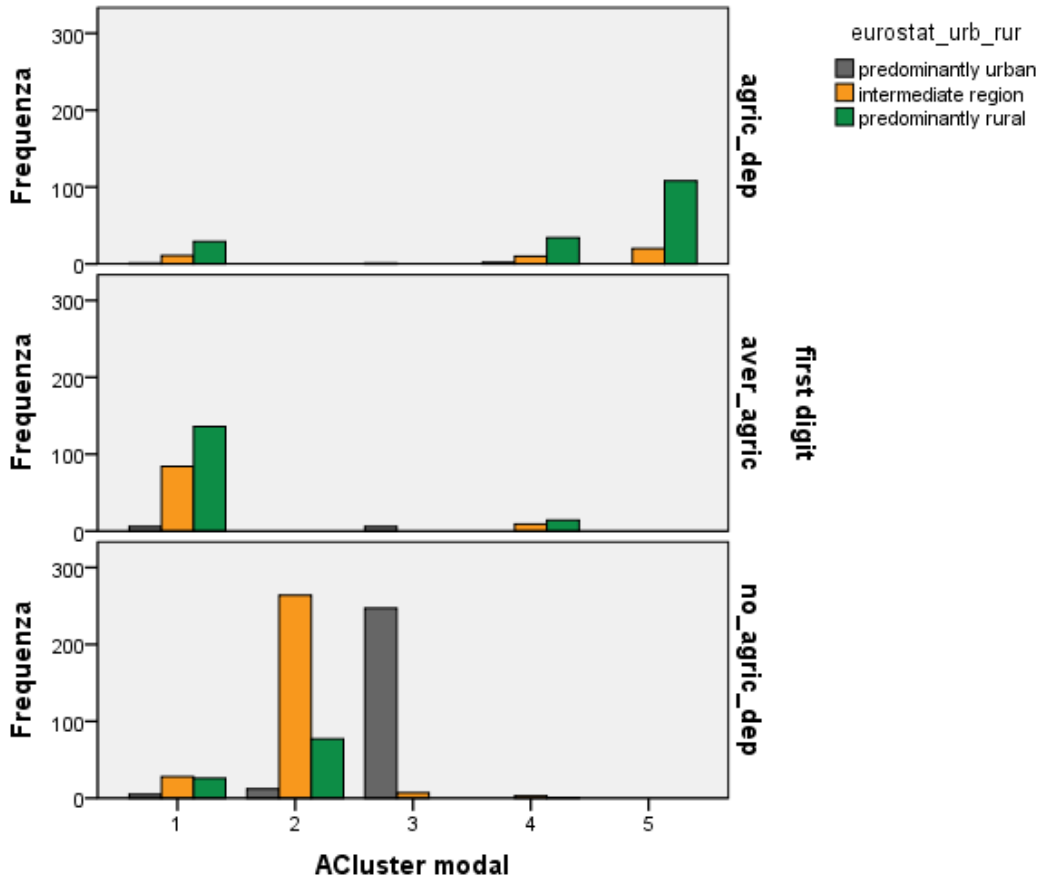


Figure 14: Accessibility variable box plot by cluster

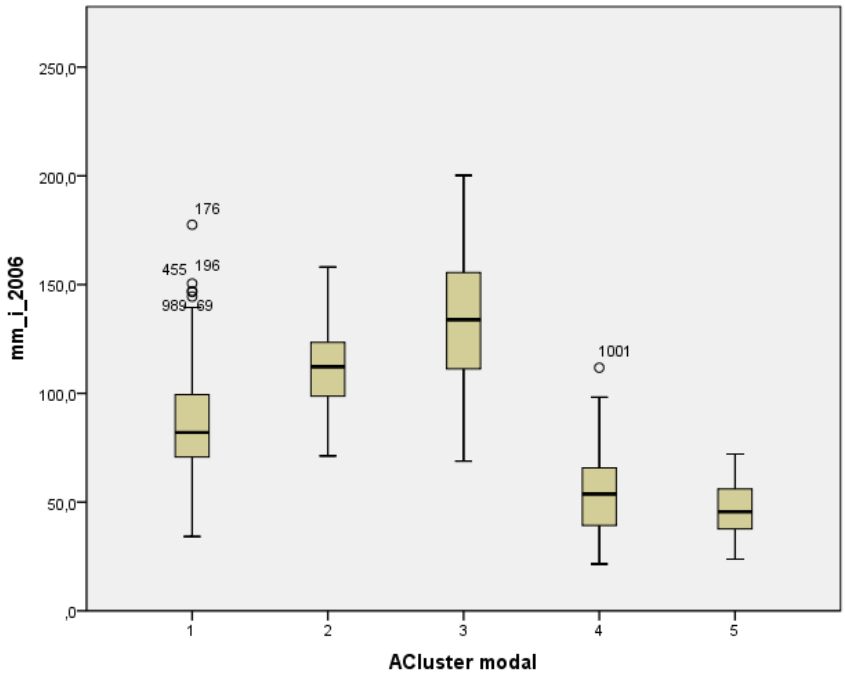
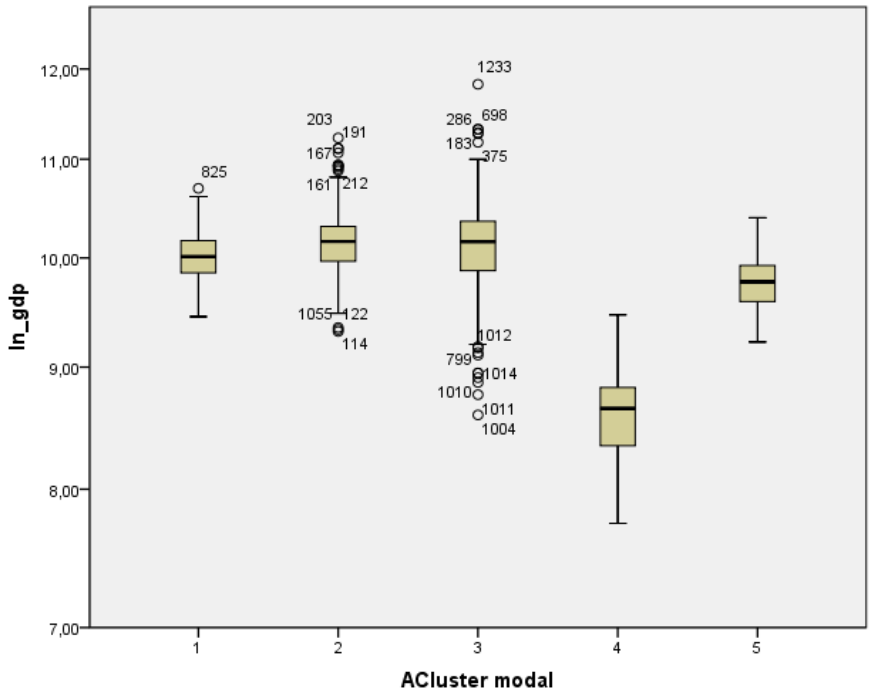


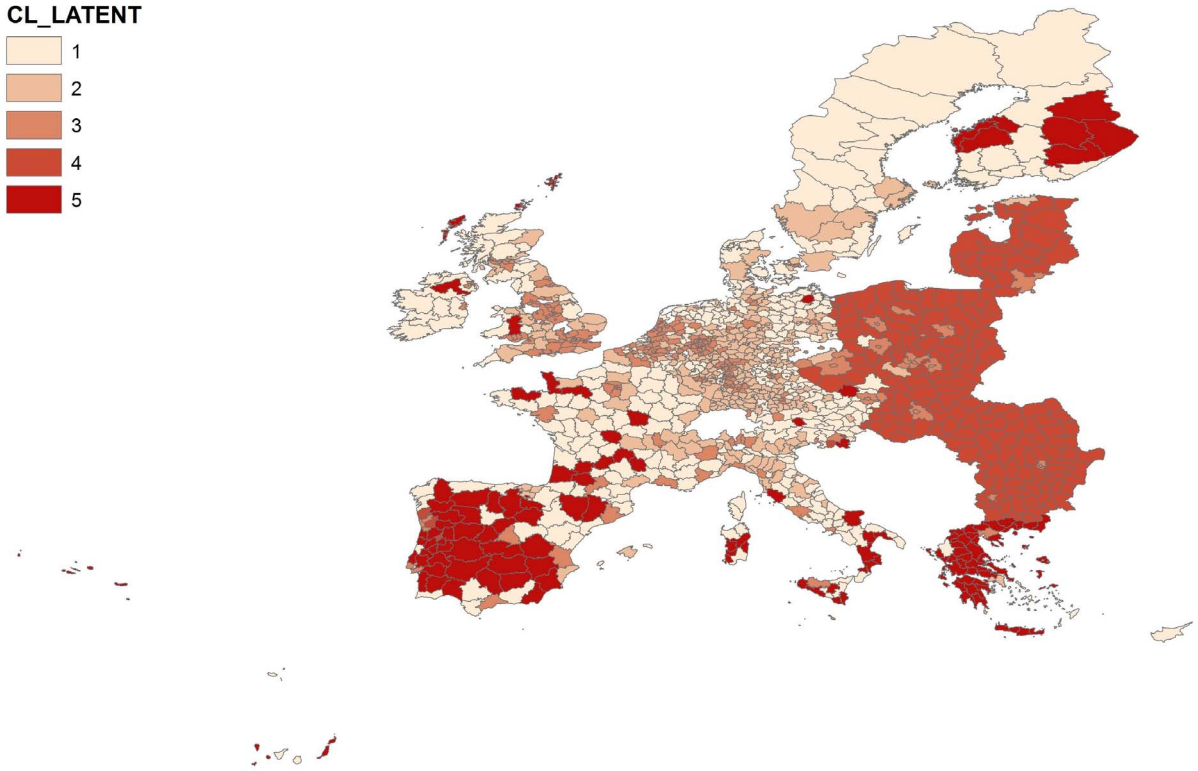
Figure 15: ln_GDP variable box plot by cluster



The map of NUTS3 classification is illustrated in Figure 16. It is useful to understand where clusters are located and to have an immediate impression of the "patchwork" of typologies. Generally, latent class models perform better than cluster analysis[5], and allow discriminating clusters in a more precise way. Clusters 2 and 3 present similar characteristics for accessibility, GDP and agriculture dependence, but cluster 2 includes mainly intermediate regions and cluster 3 predominantly urban regions.

[5] This analysis also has the advantage to cluster observations even though there are missing values in clustering variables.

Figure 16: Geographical NUTS3 classification based on the 4 variables using latent class model



4.4. Multilevel analysis results

The last analysis is performed using the best multilevel latent class model. In our case, the nested structure of the data is taken into account considering as first level the NUTS2 region and as second level the NUTS3 region. The best multilevel model is selected on the basis of the Bayesian Information Criterion (BIC). As output, the EU regions (NUTS2= first level) could be partitioned into two classes. The probability of provinces to belong to one of them is illustrated in Figure 17 (class1) and Figure 18 (class2, which is obviously complementary to class1). The distinction between the two classes is clear: the first class includes Southern and Eastern European provinces (where agriculture is important and GDP is below average); instead, the second class comprises Central and Northern Europe (characterised by high economic diversification and above average GDP).

Figure 17: Geographical distribution of probability to belong to class 1 (level-1)

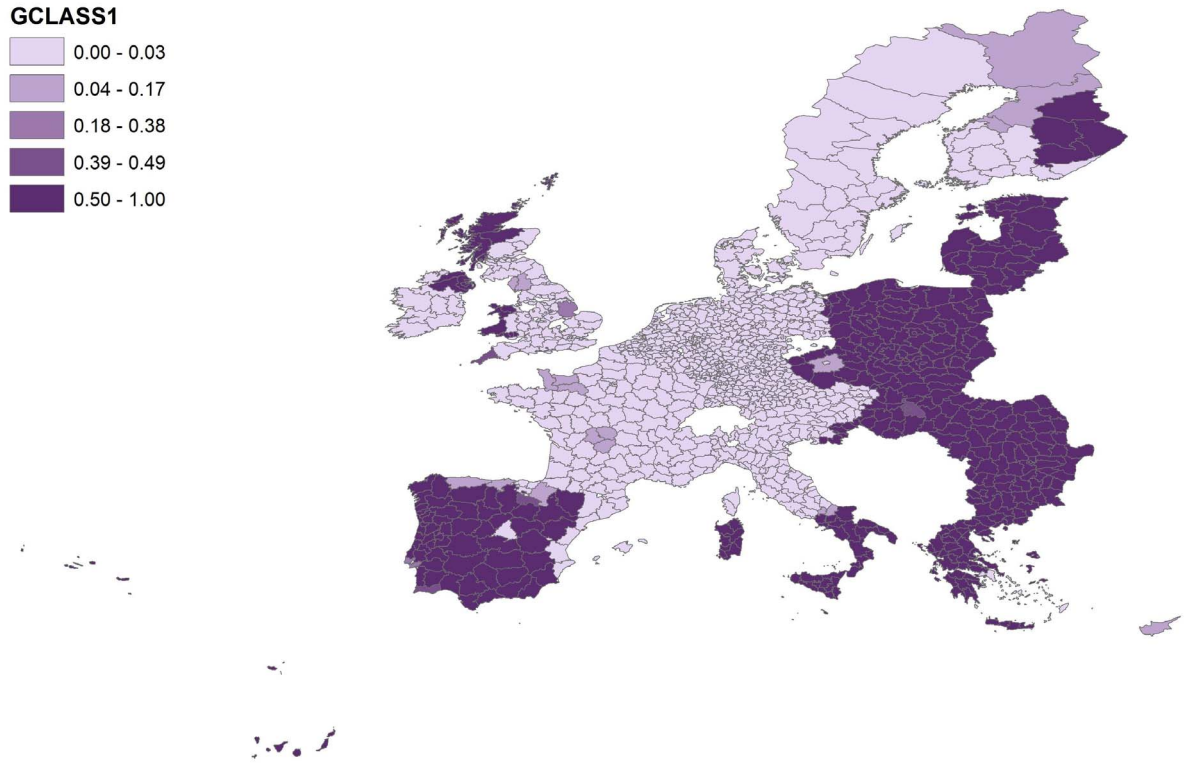


Figure 18: Geographical distribution of probability to belong to class 2 (level-1)

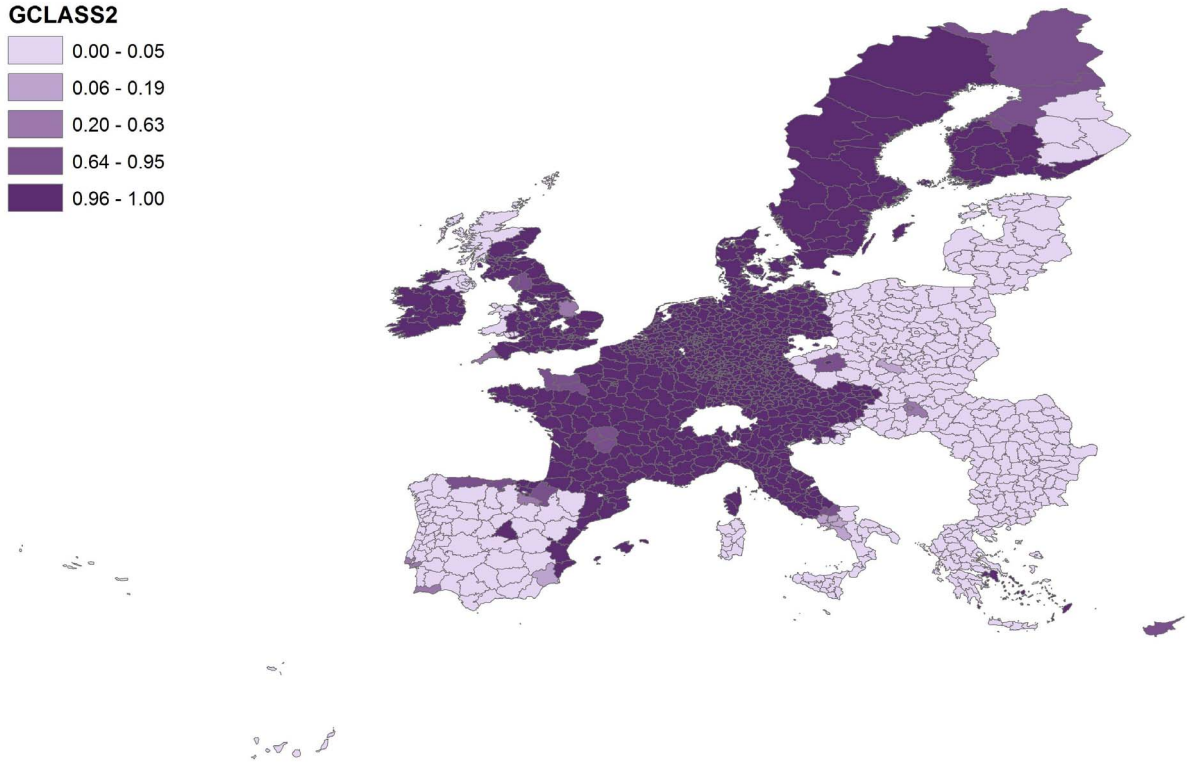


Table 11: Cluster profiles (by multiple cluster structure)

	Cluster 1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
Cluster Size	0.282	0.258	0.137	0.128	0.113	0.082
Indicators						
ln_GDP (Mean)	10.180	10.046	8.551	10.276	9.719	9.475
mm_i_2006 (Mean)	114.01	88.700	51.464	147.93	45.996	80.030
First digit=1 (agric_dep)	0.004	0.162	0.783	0.005	0.936	0.002
First digit=2 (aver_agric)	0.009	0.545	0.194	0.052	0.059	0.494
First digit=3 (no_agric_dep)	0.987	0.293	0.023	0.943	0.004	0.504
eurostat_urb_rur=1 (predominantly urban)	0.230	0.000	0.027	0.988	0.000	0.418
eurostat_urb_rur=2 (intermediate region)	0.604	0.386	0.264	0.004	0.203	0.424
eurostat_urb_rur=3 (predominantly rural)	0.165	0.614	0.709	0.009	0.796	0.158

Taking into account the multilevel structure of our data, the analysis allows for a better and deeper interpretation of the classification of NUT3 regions. In particular:

- Cluster 1 includes provinces classified as intermediate urban/rural, economically diversified, with high accessibility and high GDP.
- Cluster 2 contains rural provinces that are agriculturally dependent, with good accessibility and high GDP). Cluster 3 takes into account NUTS3 predominantly rural and agriculture dependent, with low accessibility and low GDP.

- Cluster 4 includes urban provinces with the highest GDP provinces with the highest accessibility and diversified economies.
- Cluster 5 contains rural NUTS3, strongly economically dependent from agriculture with the lowest accessibility index and low GDP.
- Finally, cluster 6 consists of urban and intermediate provinces with low GDP, intermediate accessibility and intermediate economic diversification.

Table 12: Schematic cluster features

	GDP	**Accessibility**	**Agriculture dependency**	**Rural/Intermediate/Urban**
Cluster 1: *"Rich intermediate"*	++	++	-	I
Cluster 2: *"Rich rural"*	+	+	+/-	R
Cluster 3: *"Very poor rural"*	---	-	+	R
Cluster 4: *"Rich urban"*	+++	+++	-	U
Cluster 5: *"Poor agriculture dependent"*	-	--	++	R
Cluster 6: *"Poor urban"*	--	+	-/+	U/I

The graphical distribution of each variable in each cluster is presented in the next figures (Figure 19, Figure 20 and Figure 21) that confirm the cluster interpretation described above. Comparing Figure 13 and Figure 19, it is worth noting that this results with 6 clusters that allow a better separation of NUTS3 regions leading to stronger characterisation for each cluster.

Figure 19: Cluster composition by economic diversification (first_digit) and urban/rural classification (eurostat_urb_rur) (by multiple cluster structure)

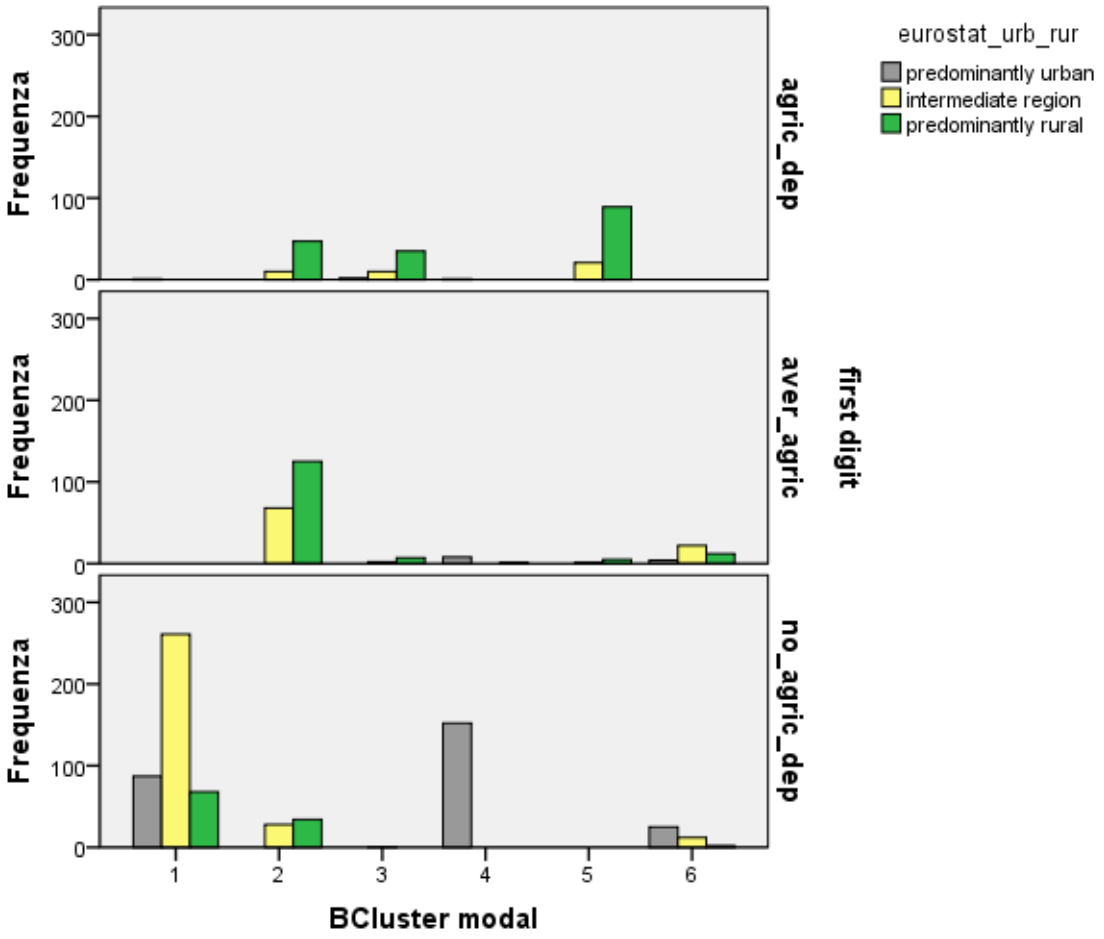


Figure 20: Accessibility variable box plot by cluster (by multiple cluster structure)

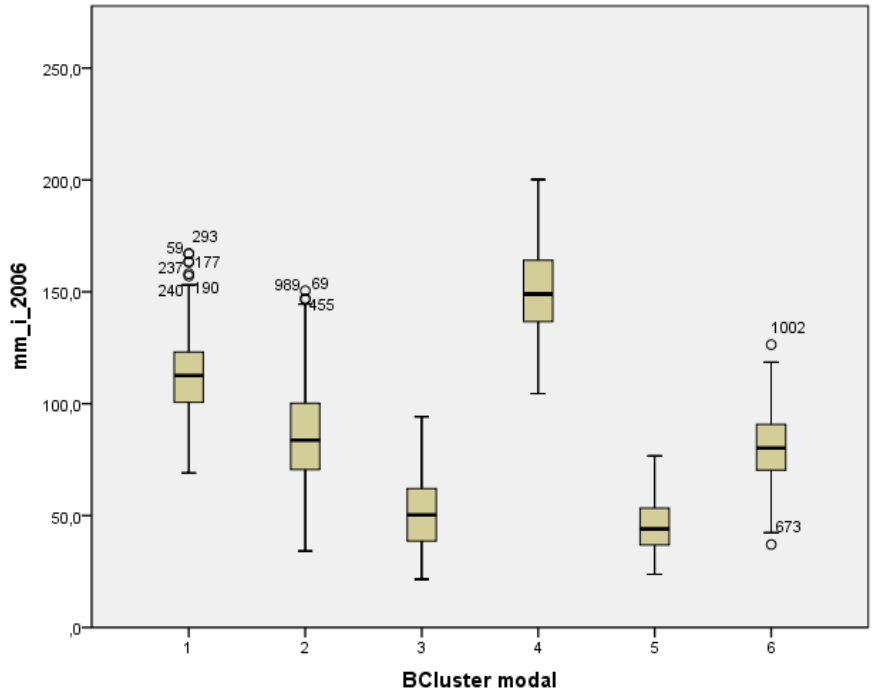


Figure 21: ln_GDP variable box plot by cluster (by multiple cluster structure)

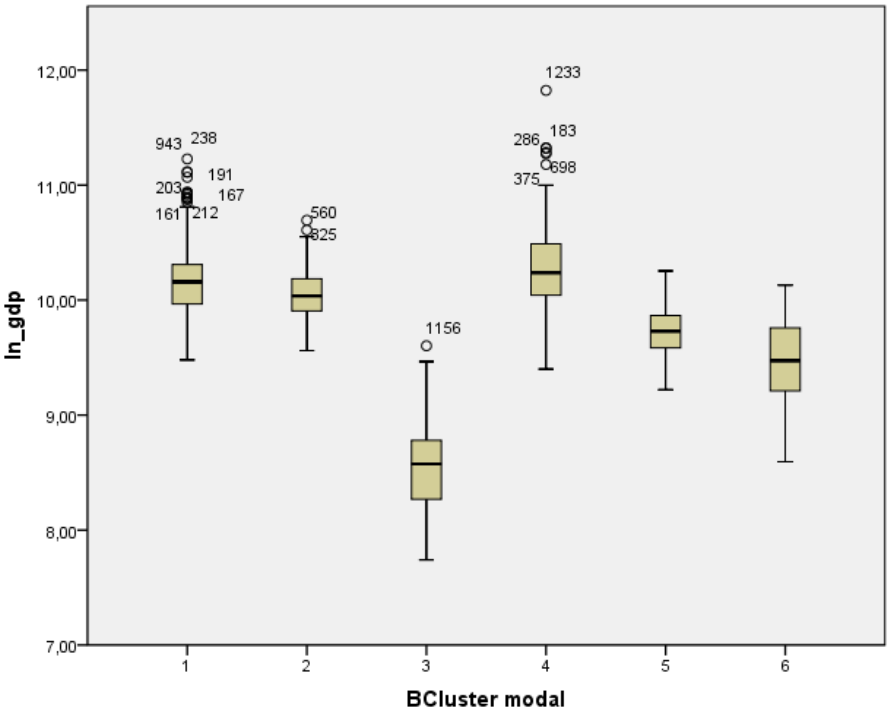


Figure 22: NUTS3 distribution among the 6 clusters and the 4 variables

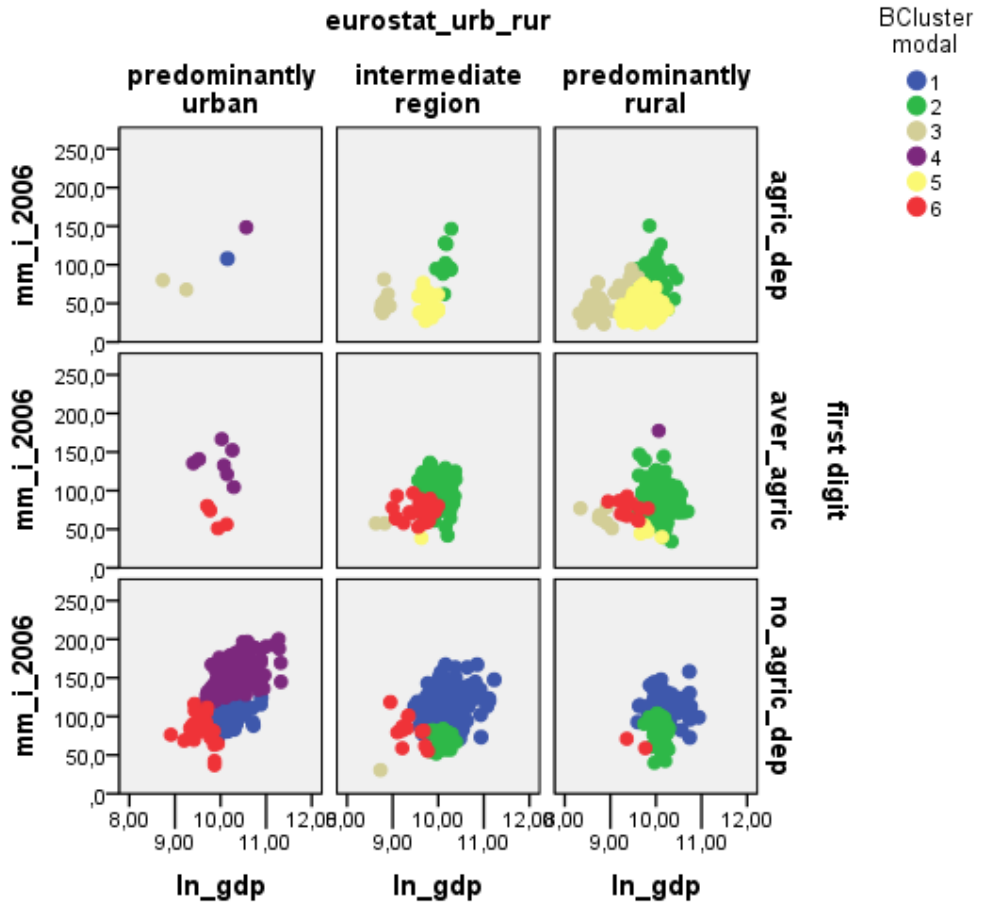
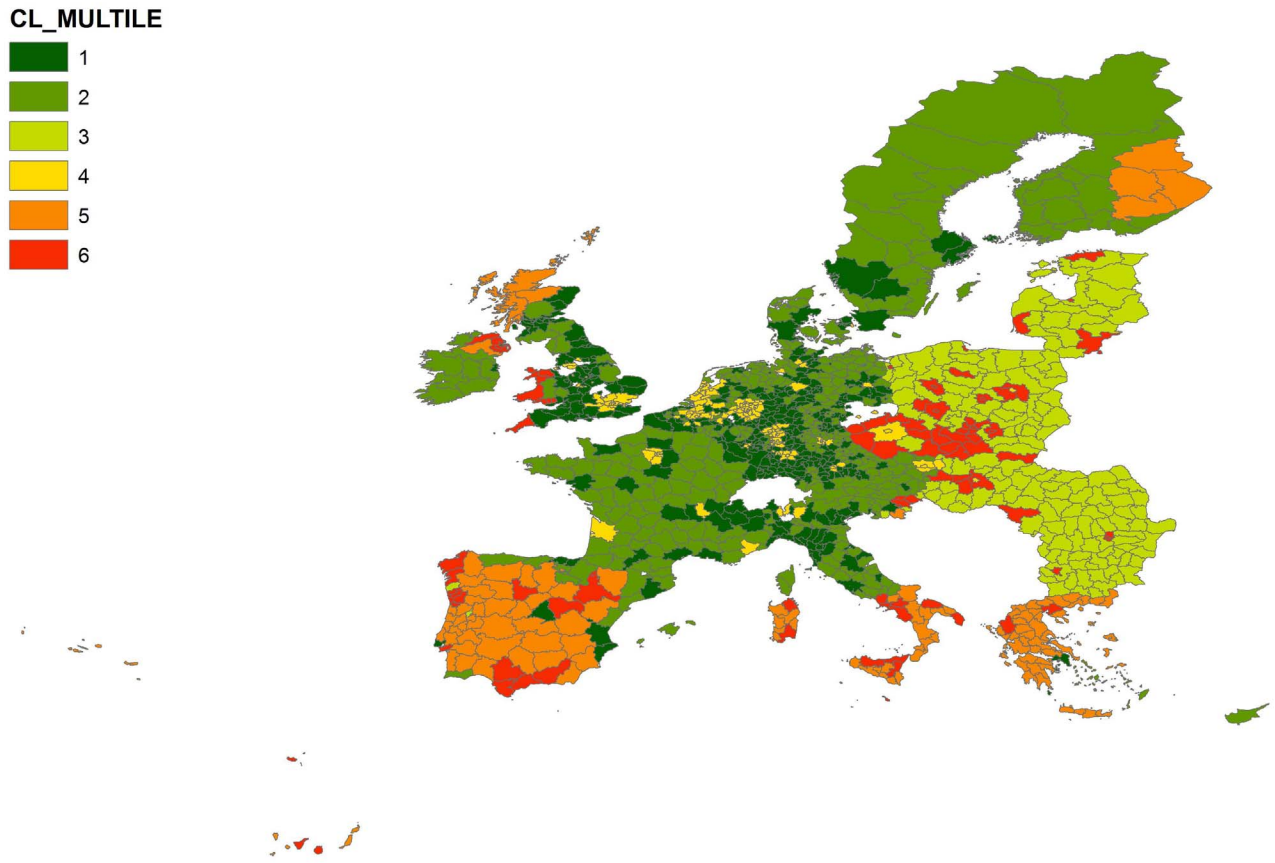


Figure 22 presents the same scatterplot than in Figure 9, but now, we include different colours to represent different clusters. For each dot cloud colours are clearly separated, meaning that NUTS3 regions belonging to a cluster are well separate to others. The map of NUTS3 classification taking into account the region to which they belong is illustrated in Figure 23. The map shows that cluster 1 ("rich intermediate") includes mainly regions in the UK, Germany, France, Southern Sweden and Northern Italy; cluster 2 ("rich rural") consists of almost all remaining Sweden, parts of Finland, all occidental EU (France, Nederland, Germany, Belgium, Denmark, UK); in cluster 3 ("very poor rural") mainly Eastern European NUTS3 regions can be found; cluster 4 ("rich urban") represents regions with capitals or large cities (e.g. Paris and London); cluster 5 ("poor agriculture dependent") includes mainly Northern UK, Southern Italy, Greece and Spain; cluster 6 ("poor urban") collect NUTS3 regions in Eastern Europe, Southern Spain and Southern Italy.

Figure 23: Geographical NUTS3 classification based on the 4 variables and taking into account the region (NUTS2) to which they belong



5. DISCUSSION OF RESULTS

Four different methods have been applied in order to classify NUTS3 regions, considering four characteristics: rurality, economic diversification, accessibility and GDP (ln_GDP). The first attempt is a very rough analysis using proxies for ordinal variables instead of qualitative ones. All the variables have significant power to discriminate, but resultant clusters show a composition that is not clearly identifiable, since there are overlapped features among them.The SPSS TwoStep algorithm is applied obtaining 6 clusters. Using this method it is possible to include qualitative and quantitative variables, but results from our data show a strong relevance of qualitative variables.. In fact, the clusters obtained are merely a combination of qualitative attributes. The latent class model gives 5 clusters. Results are relatively good, but two

clusters are difficult to interpret (i.e. clusters 2 and 3) and include a mix of features for the same variables. Taking into account the multilevel structure of the data, i.e. each NUTS3 region belongs to a unique NUTS2 region, we obtain a classification in two level (NUTS 2 regions) classes and 6 clusters (NUTS3 regions). The interpretation of these clusters is straightforward and seems to catch better different groups of NUTS3 regions.

CHAPTER 2: SAMPLE SIZE AND SAMPLING PROCEDURE

1. OBJECTIVE OF CHAPTER 2

The second task is strictly related to the NUTS3 classification obtained in task 1. Based on results obtained in task 1, an analysis of different alternative sampling schemes is carried on. The aim is to answer the following question: if a statistically significant sample was to be drawn from the total sample of 1303 regions, what would the sample size considering the diversity observed in the set of 4 variables used for Task 1? This report provides some answers in terms of combinations sample size/sampling procedure and considers simple random sampling and stratification sampling with different approaches for allocating strata sample size. The report is organised as follows: Section 7 introduces the different sampling procedures. Section 8 reports results provided by the application of different sample procedures and section 9 briefly discusses the results.

2. METHODOLOGY: HOW TO DETERMINE THE SAMPLE SIZE

2.1. Random sample

A simple random sample is obtained by a method that gives the same chance to be selected to every population unit; consequently all possible samples are equally likely to be selected. It can be drawn, for example, using random number tables, lottery numbers, or with a systematic procedure. An important issue is the determination of the number of units to be included in the sample. If the sample is too large, then effort, time and money are somewhat wasted. Conversely, if the number of sampled units is too small, inadequate information may be collected, which diminishes the precision of the results and therefore their utility.

Theoretically, the size of the sample is linked to a specified level of precision. The maximum difference between estimate and the parameter value that can be tolerated is called *permissible error*. Once the permissible error has been specified, the sample size,

which meets those requirements, is determined. Since the amount of error differs from sample to sample, the error is specified by this probability:

$$Pr\left(\left|\hat{\theta} - \theta\right| < E\right) = 1 - \alpha$$

Where θ is the population parameter, $\hat{\theta}$ its estimate, E the permissible error and $1 - \alpha$ the level of confidence (typically 90%, 95% or 99%). The required sample size is determined by equating half width of the confidence interval to the permissible error E and solving the resulting equation for the sample size n (included in the estimator variance of θ). Thus, it is understandable that the sample size determination is dependent on the variance of the estimator that is dependent on the typology of the parameter in which we are interested in. If θ is a proportion P of some population characteristics to be estimated, then[6]:

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 P(1-P)}{E^2} \qquad (1)$$

Where $z_{\alpha/2}$=1.96 with a 95% confidence level, 2.58 with a 99% or 1.56 with 90%.

The determination of the sample size requires knowledge of the variance of the estimator that is connected to the population variance of the characteristic surveyed (here the variability of ln_GDP – GDP, mm_i_2006 – accessibility, first digit TERA-SIAP – agricultural dependence and eurosta_urb_rur – rural/urban character). The required sample size can be determined by using prior information on the variance of the population, but it is usually difficult to obtain reliable information on these parameters. One possibility is to recover this information from a previous study or to use a small preliminary sample to estimate the population parameter values, which in turn are used to determine the final sample size. A different option is to calculate the "worst" situation possible, the maximum variability of P, i.e. P=0.5:

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 0.25}{E^2}$$

[6] A specific case is when the parameter is the mean, then the formula becomes: $n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$

In a finite population, i.e. when the total number of observations N is not very large, a correction is needed for Equation (1) (e.g. Krejcie & Morgan, 1970). Using this correction, Equation (1) becomes:

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 P(1-P)N}{(N-1)E^2 + \left(z_{\frac{\alpha}{2}}\right)^2 P(1-P)} \qquad (2)$$

2.2. Stratified sample

The precision of an estimate depends on the sample size and on the variability among population units. Therefore, apart from increasing the sample size, another means of increasing estimate precision could be to use a stratified random sample. The first step is to divide the population units into groups (strata) such that the variability *within the groups* is the smallest while it is the largest *among groups*. Obviously, strata are also chosen to divide a population into categories relevant to the research question. Then, smaller samples can be drawn randomly and independently from each group. Probabilistic methods to select units allow making statistical valid conclusions from the collected data.

The advantages of stratified random sampling are multiple:

- Since the population is split in strata and samples are drawn from each stratum, it is unlikely that any essential population group will be completely excluded.
- It is possible to use different sampling designs in different strata.
- When there are population extreme values, they can be grouped into a separate stratum, thereby reducing the variability within other strata.
- When strata are formed using administrative boundaries, for the sake of convenience, for example considering geographical localisation as a stratifying variable, the cost of the survey is expected to be less for a stratified sample.
- The stratified random sample can improve the representation of particular strata, as well as ensuring that these strata are not over-represented. Together, this helps to compare strata, as well as make more valid inferences from the sample to the population.
- Since the variability within strata is reduced, the stratification normally provides more efficient estimates than random (unstratified) sampling. In fact, a stratified procedure

improves the potential for the units to be more spread out over the population. Furthermore, where samples are of the same size, a stratified random sample can provide greater precision than a simple random sample. Because of the greater precision of a stratified sample, it may be possible to use a smaller sample.

Stratified sampling suffers from one main limitation in that it needs the availability of a complete list of the population, requiring that each unit from the population must (only) belong to a stratum. To create a stratified random sample, the steps are:

(a) defining the population;
(b) choosing the relevant stratification;
(c) listing the population according to the chosen stratification;
(d) choosing the sample size and the sample allocation to different strata;
(e) using a simple random or systematic sample to drawn sample units.

Besides the sample size and the variability among the population unit, the precision of the estimate based on the stratified sample also depends on the sample allocation to different strata. Three main methods of sample allocation are proposed in literature: equal allocation, proportional allocation and optimum allocation.

Equal allocation

In this case, the number of sampling units selected for each stratum is the same. Then,

$$n_h = n/H$$

Where n_h is the sample size for stratum *h*, *H* the number of strata and *n* total sample size. This method is used when stratum sizes do not differ much from each other and the information about the variation within strata is lacking.

Proportional allocation

With proportional stratification, the sample size of each stratum is proportional to the population size of the stratum. Strata sample sizes are determined by the following equation:

$$n_h = (N_h/N) * n \qquad (3)$$

Where n_h is the sample size for stratum *h*, N_h the population size for stratum *h*, *N* total population size, and *n* total sample size. Because of its simplicity, this method is often used; it is likely to be near the optimum allocation (see below) for a fixed sample size, when strata variances are almost same. A major disadvantage of proportional allocation is that sample size in a stratum may be low hence providing unreliable stratum-specific results.

Optimum allocation

Using optimum allocation, the sample size for each stratum is directly proportional to the stratum variance and inversely proportional to the average unit cost of data collection in the stratum. This optimum stratum allocation yields estimates with the smallest possible variance for a fixed total budget (i.e. the more precise estimates with a fixed budget). When the data collection cost is equal for each unit in stratum, the optimum allocation is:

$$n_h = n\frac{N_h\sigma_h}{\sum_{h=1}^{H} N_h\sigma_h} \qquad (4)$$

Where n_h is the sample size for stratum *h*, N_h the population size for stratum *h*, σ_h the standard error for the stratum *h* and *n* total sample size.

Obviously, σ_h is unknown and the problem is to substitute it with a "good" estimate that could be obtained from previous related surveys, expert opinion, data from a pre-test or a pilot study, or from knowledge of the statistical range of these values in the population. Reasonably good approximations for σ_h are likely to yield estimates, whose variances are very close to the minimum possible variance.

When the parameter of interest is a proportion *P* of the population of one characteristic, then the optimum allocation in a precautionary approach (i.e. the worst situation with the maximum variance and then the precautionary size) is:

$$n_h = n\frac{N_h\sqrt{0.25}}{\sum_{h=1}^{H} N_h\sqrt{0.25}}$$

Some final remarks:

- The optimum allocation technique allocates a larger sample size to the larger and more variable stratum.
- The estimate precision and allocation in the stratum are strictly connected to the variance and to its estimate/proxy used in the allocation formula.
- Unless unit costs differ widely among strata, proportional stratified sampling is almost always preferred when estimating proportions mainly because it is more practical and the precision is almost the same as with the optimum allocation.

3. RESULTS

3.1. Definition of a random sample

In a simple random sample, subjects in the population are sampled by a random process, using either a random number generator or a random number table, so that each unit has the same probability of being selected for the sample. In our case the population size is N=1303 NUTS3 regions and we consider the case where the interested parameter is the proportion P of one characteristic in the population. In Table 13 the sample size *n* needed for specific confidence intervals and errors is calculated based on Equation (2), considering different values of P.

Note that from Equation (1), there is an inverse relationship between sample size and the margin of error, i.e. smaller sample sizes will yield larger margins of error. Furthermore, there is also a direct relationship between sample size and the confidence level, that is, smaller sample sizes will yield smaller confidence of error. Last, there is a direct relationship between the variability of population and sample size.

Table 13: Determination of sample size for simple random sample for different confidence intervals, errors and proportions

Proportion expected in population *P*	**Confidence level (1-α)**	**Error admitted *E***	**Sample size *n***
0.2	90%	± 5%	154
0.2	95%	± 5%	207
0.2	90%	± 10%	43
0.2	95%	± 10%	59
0.5	90%	± 5%	226
0.5	**95%**	**± 5%**	**297**
0.5	90%	± 10%	65
0.5	95%	± 10%	90
0.7	90%	± 5%	195
0.7	95%	± 5%	259
0.7	90%	± 10%	55
0.7	95%	± 10%	77

3.2. Stratification: basic aspects of our study

In some cases, the researcher has access to an "auxiliary variable" believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve the accuracy of the sample design. One option is to use the auxiliary variable as a basis for stratification. This method is useful, as it increases the estimate precision (or decreases the sample size needed for a fixed precision), but only if the stratification variable is correlated to the variable of interest, and only if it is possible to hypothesize that strata are created as homogeneously within units and heterogeneously among strata.

Therefore, in our study the stratification with one or more of the four variables (i.e. GDP, rurality, accessibility, and agricultural dependence) only makes sense for the analysis of some other variables correlated to them. The natural stratification is

obtained by cluster analysis illustrated in Chapter 1 of this report, where all four variables have been used to classify the NUTS3 regions using a multilevel latent class model. Following the results achieved in Chapter 1 of this report, we consider 6 strata and in the following paragraphs we present four possibilities to allocate sample units in strata:

(a) Proportional allocation;

(b) Optimal allocation using accessibility variable;

(c) Optimal allocation using GDP variable;

(d) Optimal allocation using a hybrid variable, which is a combination of accessibility and GDP variables.

In order to obtain the optimal allocation, the variance of auxiliary variables needs be calculated, so only quantitative variables could be used and therefore (i.e. it is not possible to use rurality and agriculture dependency).

3.3. Stratification: proportional allocation

The sample size in each stratum must be proportional to the dimension of the strata. The fundamental hypothesis is that larger strata have larger variability and need be sampled more. The proportional allocation is determined in Table 14, using Equation (3), considering the worst situation in Table 13, i.e. n=297 (greatest variability, confidence level= 95%, margin of error= ±5%).

Table 14: Determination of sample size in each stratum with proportional allocation

Stratum (Cluster)	**Dimension of stratum (N_h)**	**N_h/N**	**Sample size in stratum n_h**
1	425	0.326	97
2	324	0.248	74
3	153	0.117	35
4	176	0.135	40
5	128	0.098	29
6	97	0.074	22

3.4. Stratification: optimal allocation with accessibility as auxiliary variable

In the case of optimal allocation, the sample size in each stratum is correlated to the variance in each stratum, true optimal allocation assumes knowledge of the variances σ_h^2. In practice, of course, these quantities will not be known. However, their estimates can often be obtained either from a preliminary pilot study of the population. Alternatively, we can assume that these are unchanged from past studies of the same population, or are the same as the relative sizes of the stratum variances of another variable whose values are known for all population elements[7]. In our study, we assume the accessibility variable to be highly correlated with the variable of interest (e.g. use of agri-environmental funds at NUTS3 level) and use the accessibility variance in each stratum to determine the size. Obviously, the greater the correlation, the more similar the result will be to the "true" optimal allocation.

Table 15 presents the determination of optimal allocation that is obtained using Equation (4), considering the worst situation in Table 13, i.e. n=297 (greatest variability, confidence level= 95%, margin of error= ±5%).

Table 15: Determination of sample size in each stratum with optimal allocation

Stratum (Cluster)	**Dimension of stratum (N_h)**	**Standard deviation of accessibility (s_h)**	**Optimal sample size in stratum (n_h)**
1	425	17.345	93
2	324	21.851	89
3	153	16.887	33
4	176	19.038	42
5	128	12.279	20
6	97	16.682	20

It should be stressed that differences in stratum sample size between Table 14 and Table 15 are linked to the variability observed in each stratum, i.e. a higher standard deviation leads to an increase of sample size in the stratum. Since total sample size is

[7] This is because we are only interested in the relative sizes of these variances between the strata.

the same (n=297), it is possible to calculate the design effect, i.e. the ratio of the variance of a statistic with a complex sample design to the variance of that statistic with a simple random sample. This is a valuable tool for sample design as it defines the increase of the estimator precision in a complex design. In our case:

$$D_{eff} = V(\bar{X})_{STRAT} / V(\bar{X})_{SRS}$$

Where D_{eff} is the design effect, $V(\bar{X})_{SRS}$ the variance of the mean estimator in simple random sample and $V(\bar{X})_{STRAT}$ the same in a stratified sample. Assuming that the sample strata averages and the variances of the variable accessibility are highly correlated to the same characteristics of the variable of interest, we can use the accessibility features:

$$V(\bar{X})_{STRAT} = \sum_{h=1}^{H} W_h \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2 = \quad 0.845$$

Table 16 reports the intermediate calculus. For the simple random sampling, we obtain:

$$V(\bar{X})_{SRS} = \frac{N-n}{N}\frac{s^2}{n} = \frac{1303-297}{1303}\frac{1368.112}{297} = 3.5564$$

The efficiency of this stratified sampling scheme is $D_{eff} = V(\bar{X})_{STRAT} / V(\bar{X})_{SRS}$= 0.2377, meaning that this stratified sampling is about 4.21 times more efficient than a simple random sampling; in practice, this stratified sampling is as about as efficient as a simple random sampling of 1250 regions (1250=297/0.2377). This concept that is directly related to the design effect is called effective sample size, and is usually denoted as

$$n_{eff} = n / D_{eff}$$

The effective sample size is the size of a simple random sample that would yield the same level of precision for the survey estimate as that attained by the complex design.

Table 16: Determination of variance of mean estimator in optimal allocation with accessibility as auxiliary variable

Stratum (Cluster)	**Dimension of stratum (N_h)**	**Optimal sample size in stratum (n_h)**	**Variance of accessibility (s^2_h) (A)**	**$W_h = n_h/N_h$ (B)**	**$1/n_h$ (C)**	**$1 - n_h/N_h$ (D)**	**A*B*C*D**
1	425	93	300.9	0.326	0.011	0.781	0.269
2	324	89	477.5	0.248	0.011	0.725	0.239
3	153	33	285.2	0.117	0.031	0.787	0.095
4	176	42	362.5	0.135	0.024	0.760	0.119
5	128	20	150.8	0.098	0.051	0.845	0.062
6	97	20	278.3	0.074	0.049	0.789	0.059
							0.845

3.5. Stratification: optimal allocation with GDP as auxiliary variable

Considering now the logarithm of GDP as an auxiliary variable, the optimal allocation is shown in Table 17. The differences in optimal allocation in Table 15 and Table 17 (last columns) are due to the differences in variability of the two variables (accessibility and lnGDP) within strata. Table 18 illustrates the intermediate calculus to define $V(\bar{X})_{STRAT}$ = 0.000208.

For the simple random sampling, we obtain:

$$V(\bar{X})_{SRS} = \frac{N-n}{N}\frac{s^2}{n} = \frac{1303-297}{1303}\frac{0.372}{297} = 0.000967$$

The efficiency of this stratified sampling scheme, based on GDP as auxiliary variable, is $D_{eff} = V(X)_{STRAT}/V(X)_{SRS}$=0.2153, meaning that this stratified sampling is about 4.64 times more efficient than a simple random sampling. The effective sample size in this case is n_{eff}= 1380 (1380=297/0.2153).

Table 17: Determination of sample size in each stratum with optimal allocation with ln_GDP as auxiliary variable

Stratum (Cluster)	**Dimension of stratum (N_h)**	**Standard deviation of lnGDP (s_h)**	**Optimal sample size in stratum (n_h)**
1	425	0.2919	99
2	324	0.1964	51
3	153	0.3815	47
4	176	0.3738	53
5	128	0.2167	22
6	97	0.3338	26

Table 18: Determination of variance of mean estimator in optimal allocation with ln_GDP as auxiliary variable

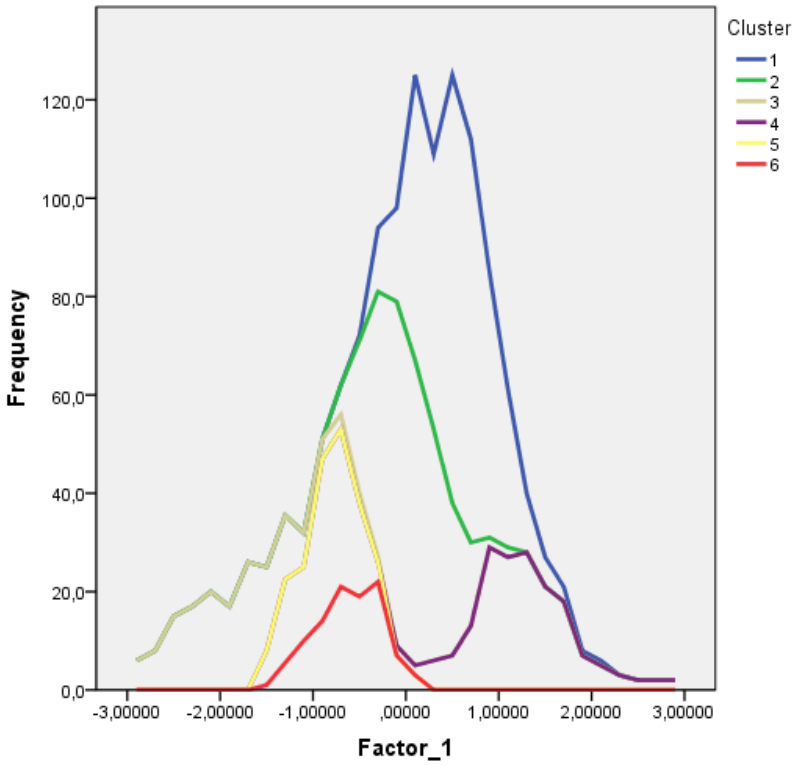
Stratum (Cluster)	**Dimension of stratum (N_h)**	**Optimal sample size in stratum (n_h)**	**Variance of lnGDP (s^2_h) (A)**	**$W_h = n_h/N_h$ (B)**	**$1/n_h$ (C)**	**$1 - n_h/N_h$ (D)**	**A*B*C*D**
1	425	99	0.085	0.326	0.010	0.766	0.000070
2	324	51	0.039	0.249	0.019	0.843	0.000040
3	153	47	0.146	0.117	0.021	0.695	0.000030
4	176	53	0.139	0.135	0.019	0.702	0.000034
5	128	22	0.047	0.098	0.045	0.827	0.000017
6	97	26	0.111	0.074	0.039	0.734	0.000018
							0.000208

3.6. Stratification: optimal allocation with a "hybrid" auxiliary variable

The choice of which variable is better to use to allocate the sample is linked to the variable of interest. The best auxiliary variable is the one that is most correlated to the variable of interest, then the ones that have the most similar variability within strata.

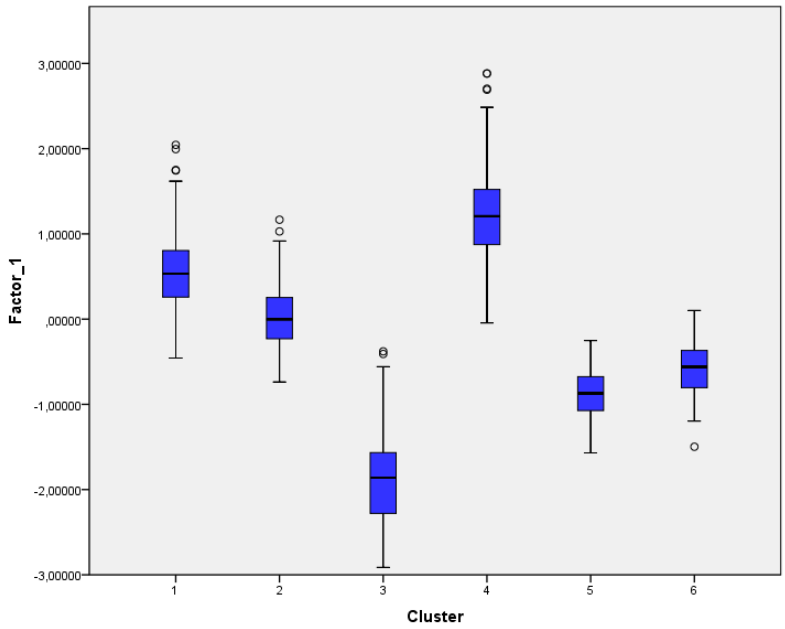
An alternative option is to try to combine both variables in a new hybrid variable and then to use it for optimal allocation of the sample. We create a new hybrid variable utilising the principal component analysis (PCA). PCA is a means of discerning simple structure from the interrelationship of variables; it is a way of identifying patterns in data and expressing the data in such a way that it highlights their similarities and differences. PCA is considered useful to find patterns in high-dimensional data; in our case, we use it as a method of data reduction, in particular as a possibility to combine two correlated variables (accessibility and ln_GDP) towards creating a new variable. We are interested in obtaining the component score (the values of the new variable) with the purpose of using it as an auxiliary variable in stratified optimal allocation.

Figure 24: Frequency distribution of hybrid variable (Factor_1) by cluster



The PCA analysis leads to one factor (called Factor_1 in the database) that accounts for 79.81% of the variance (about 80% is an acceptable result).This factor performs well in the 6 clusters, as it has different distributions (mean, variability and trend), in Figure 24. Figure 25 shows how Factor_1 has different means and levels of variability in the clusters. In fact the boxes are not overlapped and different value of quartiles meaning that it could be used as a stratification variable.

Figure 25: Box plot of Factor_1 variable by cluster



The optimal allocation in each stratum is defined in Table 19. Differences among the two previous methods are relevant. The sample size obtained by Factor_1 is positioned in the middle between ln_GDP allocation and the accessibility allocation for clusters 1, 2, 3 and 5. Therefore it is the lowest for cluster 6 and equal to ln_GDP allocation for cluster 4.

Table 19: Determination of sample size in each stratum with optimal allocation with Factor_1 as auxiliary variable

Stratum (Cluster)	Dimension of stratum (N_h)	Standard deviation of Factor_1 (s_h)	Optimal sample size in stratum (n_h)
1	425	0.412	97
2	324	0.354	64
3	153	0.522	45
4	176	0.537	53
5	128	0.297	21
6	97	0.322	17

Table 20 illustrates the intermediate calculus to define $V(\bar{X})_{STRAT}$ = 0.00043. For the simple random sampling, we obtain:

$$V(\bar{X})_{SRS} = \frac{N-n}{N}\frac{s^2}{n} = \frac{1303-297}{1303}\frac{1}{297} = 0.0026$$

The efficiency of this stratified sampling scheme, based on the hybrid variable, is $D_{eff} = V(\bar{X})_{STRAT}/V(\bar{X})_{SRS}$=0.1655, meaning that stratification with optimal allocation with Factor_1 is about 6 times more efficient than a simple random sampling. The effective sample size in this case is n_{eff}= 1795 (1795=297/0.1655).

The better performance of Factor_1 as an auxiliary variable is the result of a good combination of the variability of both accessibility and ln_GDP variables. Even if we consider only one factor of PCA and it captures only 80% of the total variability, the results show a significant increase in estimate precision. For example, using a sample size of 50 units with a stratified optimal allocation, we can obtain the same precision as a simple random sample of 297 units (P=0.5 greatest variability, confidence level= 95%, margin of error= ±5% in Table 13).

Table 20: Determination of variance of mean estimator in optimal allocation with Factor_1 as auxiliary variable

Stratum (Cluster)	**Dimension of stratum (N_h)**	**Optimal sample size in stratum (n_h)**	**Variance of Factor_1 (s^2_h) (A)**	**$W_h = n_h/N_h$ (B)**	**$1/n_h$ (C)**	**$1 - n_h/N_h$ (D)**	**A*B*C*D**
1	425	97	0.169	0.326	0.010	0.771	0.000143
2	324	64	0.125	0.249	0.016	0.803	0.000097
3	153	45	0.273	0.117	0.023	0.70	0.000060
4	176	53	0.288	0.135	0.019	0.701	0.000070
5	128	21	0.088	0.098	0.047	0.834	0.000034
6	97	17	0.103	0.074	0.058	0.821	0.000027
							0.00043

4. DISCUSSION OF RESULTS

Some concluding remarks are needed to correctly use the results of this report. First, the report highlights the need for a link/correlation between stratification variable(s) and the variable of interest. This correlation means that the average and variance of the variable of interest change with the values of these auxiliary variables, and so it makes sense to partition the population by defining strata on the basis of these auxiliary variables in order to control the variation in our sample by sampling independently from the different strata.

The researcher may at times be interested in estimating population characteristics for several variables. If all variables of interest are closely related to a single auxiliary variable (say X) and information for all population units on X is available, then stratification and allocation is a good way to proceed. If all the interest variables are not related to a single auxiliary variable but are related to more than one auxiliary variable, the procedure of stratification and allocation needs to be modified. One possibility is to

undertake a multiple stratification. In this method, population units are first stratified using the most important auxiliary variable and the strata formed are called primary strata; then each primary stratum is further stratified using another auxiliary variable (secondary strata). Instead, another possibility is the creation of one (or more) variables using reduction data analysis such as Principal Component Analysis or Factor Analysis. This is the approach followed in this report. In this case, the analysis is useful to identify new meaningful variables without much loss of information starting from correlated variables.

Moreover, the stratification procedure and allocation sample in strata are possible only if researchers have access to one or more auxiliary variables, the values of which are known for the entire population. Furthermore, another consideration is required with regard to sample size and the numbers needed for the data analysis; if descriptive statistics are to be used, (mean, frequencies, variability, etc), then nearly any sample size will suffice. On the other hand, a good size sample, e.g. 200-500, is needed for multiple regressions, analysis of covariance, or log-linear analysis, which could be performed for a more rigorous state impact evaluation. The sample size should be appropriate for the analysis that is planned. In addition, an adjustment to the sample size may be needed to accommodate a comparative analysis of subgroups (e.g., such as an evaluation of program participants with nonparticipants).

Finally, the sample size formulas provide the number of responses required. Many researchers commonly add 10% to the sample size to compensate for units that the researcher is unable to contact. The sample size is also often increased by 20-30% to compensate for non-responses. Thus, the number of mailed surveys or planned interviews can be substantially larger than the number required for a desired level of confidence and precision.

REFERENCES

Bacher J., Wenzig K., Vogler M. (2004). SPSS TwoStep Clustering – A First Evaluation. In Dijkum C., Blasius J., Durand C.(eds.), Recent Developments and Applications in Social Research Methodology. Proceedings of the RC33 Sixth International Conference on Social Science Methodology, Amsterdam 2004.

Krejcie R.V., Morgan D.W. (1970). Determining Sample Size for Research Activities. Educational and Psychological Measurement, 30, 607-610.

Vermunt, J.K. (2003). Multilevel latent class models. Sociological Methodology, 33, 213-239.

Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. Statistical Methods in Medical Research, 17, 33-51.

Vermunt J.K., Magidson J. (2002). Latent class cluster analysis. In J.A. Hagenaars, A.L. McCutcheon (eds.), Advances in Latent Class Analysis, Cambridge University Press.

Weingarten P., Neumeier S., Copus A., Psaltopoulos D., Skuras D., Balamou E. (2010). Building a Typology of European Rural Areas for the Spatial Impact Assessment of Policies (TERA-SIAP), Luxembourg: Publications Office of the European Union, EUR 24398 EN.

ACKNOWLEDGEMENTS

The authors would like to thank Jacques Delincé for comments on an earlier draft of the report.

European Commission
EUR 26263 – Joint Research Centre – Institute for Prospective Technological Studies

Title: A classification of European NUTS3 regions

Authors: Meri Raggi, Sébastien Mary, Fabien Santini, Sergio Gomez y Paloma

Luxembourg: Publications Office of the European Union

2013- 56 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424 (online)

ISBN 978-92-79-34483-1 (pdf)

doi:10.2791/35200

Abstract
Over the years a number of Common Agricultural Policy (CAP) reforms have led to the emergence of a CAP chapter specifically dedicated to rural development, also referred as Pillar 2, and have resulted in a progressive switch of CAP budget from Pillar 1 (i.e. direct support to farmers, including direct payments and other instruments for market regulation) to Pillar 2. Approximately 23 per cent of the CAP should be allocated to rural development measures during the period 2014-2020. The recent development of Pillar 2 calls for further research on the impact assessment of such policies. Unfortunately, the diversity of rural situations across Europe has complicated the empirical studies of the impacts of rural development and often makes any comparison between regions rather trivial. The main objective of this report is the creation of a classification of 1303 NUTS3 regions, which reflects the heterogeneity of NUTS3 characteristics in the EU. This classification is multidimensional. In particular, the typology is based on the following set of four criteria: Rural Character, Accessibility, Actual economic diversification and Total Gross Domestic Product per capita. Such classification will facilitate the comparison of rural development policy impacts between regions of interest across Europe.

LF-NA-26263-EN-N

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new standards, methods and tools, and sharing and transferring its know-how to the Member States and international community.

Key policy areas include: environment and climate change; energy and transport; agriculture and food security; health and consumer protection; information society and digital agenda; safety and security including nuclear; all supported through a cross-cutting and multi-disciplinary approach.

Publications Office

ISBN 978-92-79-34483-1

9 789279 344831