Version 30/10/2009

The ESPON 2013 Programme

ReRisk Regions at Risk of Energy Poverty

Applied Research Project 2013/1/5

Updated Interim Report: Area Typologies - Clustering

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

This report presents the interim results of an Applied Research Project conducted within the framework of the ESPON 2013 Programme, partly financed by the European Regional Development Fund.

The partnership behind the ESPON Programme consists of the EU Commission and the Member States of the EU27, plus Iceland, Liechtenstein, Norway and Switzerland. Each partner is represented in the ESPON Monitoring Committee.

This report does not necessarily reflect the opinion of the members of the Monitoring Committee.

Information on the ESPON Programme and projects can be found on www.espon.eu

The web site provides the possibility to download and examine the most recent documents produced by finalised and ongoing ESPON projects.

This basic report exists only in an electronic version.

List of authors

Dr. Stamatis Kalogirou, NTUA

Table of contents

Figures

Maps

Tables

1. Executive summary (max. 5 pages)

The classification of the EU Regions into groups of regions with similar characteristics is a helpful tool in the study of the risk of energy poverty in Europe. This classification may result in regional typologies that will assist policy makers in to understand the picture of Europe in various aspects and form their policy agenda accordingly.

In the previous phases of this project the typologies suggested were meant to be such that allow the assessment of risk for of energy poverty as well as the spatial differentials of this risk. This would help policy makers to make informed decisions. In order to assess potential energy poverty it was apparent that it is necessary to account for economic development and prosperity of the region, infrastructure and access to the supply of energy, demographic structure of the population and of course weather conditions.

An indicative set of types of regions we originally proposed follows:

1. Metropolitan Regions (high accessibility, well developed, above national average household incomes, completion holds energy prices low, economically active households)

2. Evolving Regions (smaller towns with high growth, improvement in income and quality of life)

3. Hidden Risk Regions (areas that seem to do well, but due to the population structure, e.g. high proportion of older people, access to energy, or energy demanding climate conditions may face problems)

4. Lagging Regions (regions with current or eminent energy poverty)

Based on the recent literature, we rejected the use of Principal component analysis and fuzzy classification although they are two well established classification methodologies for this kind of analysis. Instead, we proposed to employ techniques that allow for a straight forward classification of regions. This in combination with an effective visualization of the results would allow a good communication of these to the policy makers. We also proposed that it would be interesting to also adopt an area classification method that is based on the theory and applications of geodemographics.

Although there have been limitations in the data availability that would help addressing the above typology, this report presents an extensive analysis of a set of indicators and a set of alternative classifications of the regions based on these indicators. It has been decided to allow the reader to link the classification we present here with the above typologies or other ESPON typologies based on the characteristics of the clusters. An extensive use of tables, graphs, histograms, boxplots and maps has been made at all stages because we strongly believe in the power of visualisation in the communication of statistics.

The production of an area typology is a long process that consists of several steps. We present these steps here but a more technical discussion is presented in the next section.

Step 1. Data Input: assessing row data for the regional typology

When data from different countries are put together in a single geographical layer it is very common that they do not to have been harmonised in order to account for the different approaches the data sources use for these data collection. At this stage data source(s) for each variable / indicator are checked to ensure that the data fit in a single distribution. This is important to ensure as high as possible quality data for the regions. The Inasmet-Tecnalia working group has done this assessment and provided a solid set of indicators.

We identified potential gaps in the dataset and the need for new variables/ indicators that can be derived from existing data, such as accessibility measures and urban sprawl measures, but it has not be possible to produce these data to date.

Step 2. Preparing data for the classification

It is always necessary to check the nature of the data in terms of their measure and prepare them at a form that can be inserted to the classification algorithm. One example has been the wind power potential that was originally provided at a form o wind power density. We also needed to check if some variables would make a real contribution in the analysis.

Step 3. Evaluation of input variables in terms of statistical inference

In this step of the process the appropriateness of each variable / indicator in terms of statistical inference is assessed. Some variables may be replaced or excluded from the classification because they will a) have small size, b) be correlated with other variables, c) have extreme values, or d) be much-skewed.

The classification methodologies assume that variables/indicators follow a Normal/Gaussian distribution. In many cases in the real world this is not the case. Therefore it is necessary to perform descriptive statistical analysis for each variable in order to identify those with extreme values or those exhibiting high skewness. Extreme values are usually excluded or smoothed.

However skewed variables require more attention. If the problem of skewness exists the criterion of normal distribution is violated, it is thus necessary to perform a normalisation / standardisation process to the variable in order to ensure the classification method will not suffer from misspecification bias. Another way to address this problem without replacing the variable with its normalised version is by applying a low weight.

The last part of this step consists of performing pairwise correlations in order to identify highly correlated pairs or groups of variables. It is likely that two or more variables will have similar values or distribution and therefore a high degree of correlation. This is also inappropriate for the classification algorithm because one variable may be replacing the affect of another during the classification process and thus, wrong conclusions for the type of regions may be derived. An obvious way to address this issue is by performing factor analysis and replacing the group of correlated variables with a product variable. We rejected this approach because it will then be hard to describe a region and even harder to perform scenario based analysis. Another way of reducing the potential bias is by applying lower weights to the correlated variables. Choosing the best representative variable is also common practice for experienced researchers. For the latter, opinions of energy experts were taken into account.

Step 4. Weight selection

It is apparent from the theory and empirical work that not each variable should be assumed to have an equal influence in the classification of a region. Furthermore, variables with high magnitude valued could dominate the clustering results. Based on theory and depending on the research questions the region typology is trying to answer, some indicators should receive higher weights than others.

Applying a weight to each variable is a way of addressing data issues as well as ensuring the proper influence of each variable to the regional typology. However we address the issues of data by removing problematic indicators. The influence of each variable is not only assessed on the basis of the research question and previous empirical findings for the appropriateness of each variable to the required typology but also on the evaluation of the results that may force the choice of different weights than those applied in the first run of the classification.

The final set of indicators was assessed by the ReRisk research team and the energy experts participating in the ReRisks Workshop III held in Bilbao. Based on the expert’s opinion about the appropriateness of each indicator and their ranking of the indicator’s importance in terms of the policy implications of the results of this project a single set of weights has been produced and use here.

Step 5. Clustering

It is apparent from the geography literature that among others, there are two main categories of clustering algorithms: stepwise, top-down methods and iterative location–reallocation methods. In the former the algorithm examines all possible combinations of areas into classes and converges to a fixed number of classes which are then interpreted. In the latter category of algorithms the number of classes are predefined and the algorithm allocates all areas to classes ensuring that the within a class variation is minimised. An example algorithm of the latter is the K-means clustering.

If a K-means algorithm with four classes is applied it is likely that the results will much the suggested typology. However, the lack of key variables cannot ensure that this will be the case. To ensure best performance it will be necessary to properly adjust the weights applied to each variable / indicator. This has also not been addressed.

However the geodemographic-like approach, that could be a superior one, uses an alternative way of region typology. One important aspect is that a high number of classes is produced or selected. An optimisation process follows in which the classes are assessed and may be merged manually or by repeating the clustering steps with different configuration. Hierarchical clustering is more appropriate here as it provides a clustering tree that shows the groupings in terms of the statistical output. The resulted classes are then numbered, interpreted and labelled accordingly. If there are too many classes and communicating the results is inefficient, these are grouped to categories resulting in a cluster hierarchy. It is possible that the classes are grouped to match the four types of regions specified above; however this is not known a priory. Although an attempt has been made to employ this technique, this is far from being a proper geodemographic-like approach mainly due to data availability. This should be a research question for future analysis especially if the analysis is applied at a finer geographical scale.

Finally, visualisation tools have been applied and a short description of the profile of each cluster is provided in order to ensure good communication of the results to the policy makers and the reader in general.

2. Methodology

In the Interim Report of this project, a thorough presentation of the available data and their sources has been presented. From these data some 20 variables have been identified and are available for the area typology exercise. However, in order to ensure quality input and to comply with the criteria of the clustering algorithms in terms of statistical inference, it is necessary to statistically assess the characteristics of each of these variables. This is a standard procedure in exploratory analysis in quantitative geography to ensure data quality (input) and robustness of the results (output).

The area typologies of the 287 EU Regions will be based on clustering these regions based on the available variables. The clustering procedure has about seven main steps referring to the data quality checking, clustering and presenting the results. The main steps of this procedure are shown in Figure 1 bellow.

The main data issue in concern are that some of the values for the EU Regions have missing values. Other issues include the distributing of the values that ideally should be normal and the correlation between the indicators that should be independent to ensure high quality analysis. One should also ensure that the values are also spatial independent, however the clustering algorithms employed here do not account for the spatial dependence in the data. This issue has not been discussed properly in the clustering literature, thus we ignore this fact of spatial data.

The decision of which indicators to include in the clustering exercise was based both in the preliminary descriptive analysis as well as on the expertise of the research team and the experts who participated in workshop III opinions in the context of the scenario building process. Of course one could not completely address Steps 1-4 of the flowchart bellow (Figure 1) as we do not have an a priori knowledge of what each variables distribution and spatial structure should be.

In the following section we look at the data quality in terms of statistical inference. This is that any statistical algorithm, such as k-means clustering that works on the basis of minimising the sum of squares of some sort (here the distance from cluster centres) it is a requirement that the input variables have a normal distribution and are independent from each other. It is also necessary that outliers due to error or miscalculation are removed in order to avoid biased results.

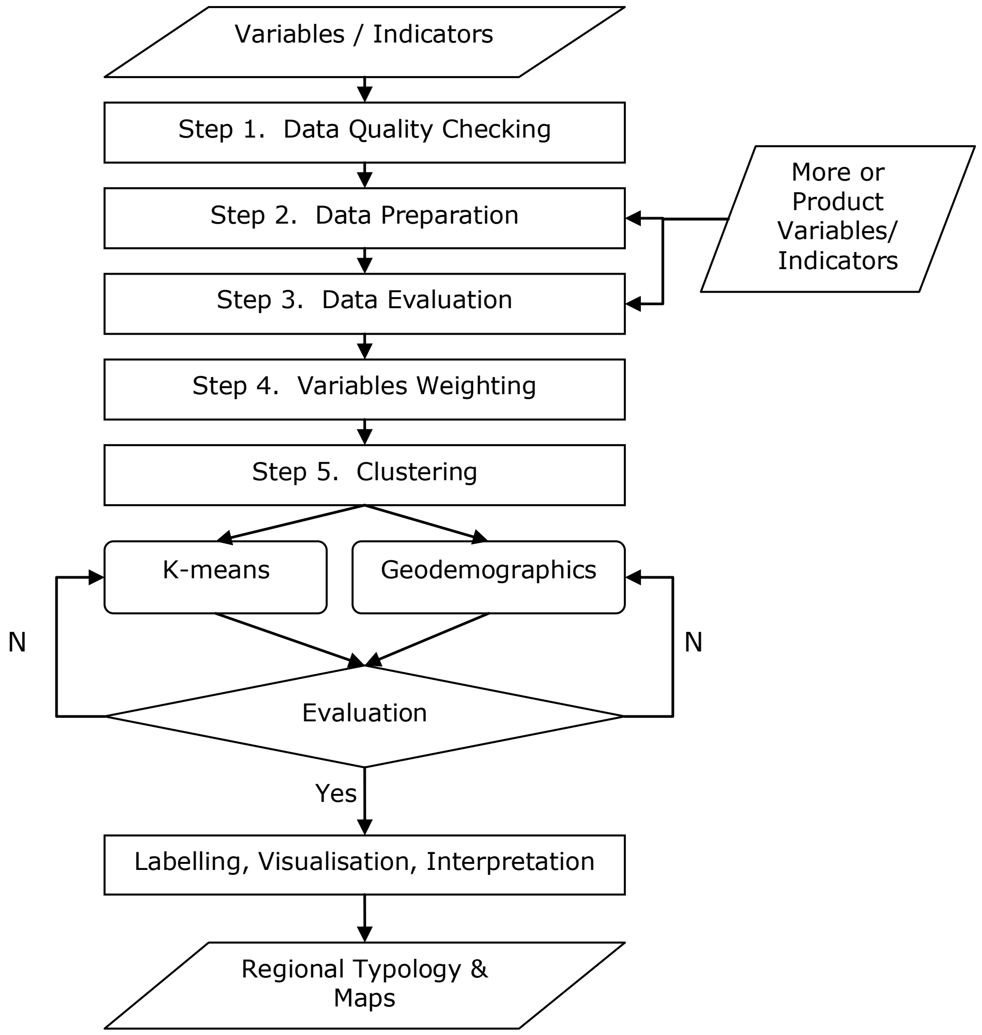


Figure 2.1 Steps for clustering procedure

It is necessary also to pay special attention on the issue of missing values. Clustering algorithms allow two options: one is to classify all regions with as many variables as possible and the other is to leave a region with at least one missing value out of the clustering exercise. The former option ensures as many classified regions as possible, but the variable set for classification changes in each region introducing bias in the results. The latter option apparently will leave several regions out of the classification exercise, but is the standard in clustering software.

Several algorithms and software has been used for the clustering. K-means and hierarchical clustering are the main function available both in commercial (SPSS) and open source statistical software (R).

The k-means procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that can handle large numbers of cases. The procedure tries to form groups that do differ. The reason for choosing to apply a k-means cluster analysis is that it allows for the grouping of regions into categories of similar rates for a set of variables. It is a quick algorithm the results of which can be easily mapped (Kalogirou, 2003).

The k-means clustering algorithm is described in detail by Hartigan (1975). The k-means used here is an efficient version of the algorithm presented in Hartigan and Wong (1979). The aim of the K-means algorithm is to divide M points in N dimensions into K clusters so that the within-cluster sum of squares is minimized. It is not practical to require that the solution has minimal sum of squares against all partitions, except when M, N are small and K=2. We seek instead "local" optima, solutions such that no movement of a point from one cluster to another will reduce the within-cluster sum of squares (Hartigan and Wong, 1979, p. 100).

Geodemographics have originally designed to classify areas based on the socio-economic profile and demographic structure of the people living in these areas. They allow for a straightforward understanding of the average person living in an area. The inclusion of energy and climate related variables in the analysis can produce an extended geodemographics-like system. Such systems use a two level hierarchical classification that can be top-down or bottom up. The top classes are groupings of the bottom classes that are more detailed. Thus, one can have an initial reading of the regions and their classification. However the system allows for a more detailed classification of the regions that may be useful for specific policy making actions. Unfortunately, the lack of appropriate data and the fact that NUTS II regional are very large in size and cover diverse population groups limit the advantages of this method and its proper application to the dataset.

The Geodemographics approach can be also implemented using k-means clustering. Based on the literature (Harris et al., 2005) this method is appropriate for high geographic details, such as the UK output areas (Singleton and Longley, 2008; Vickers and Rees, 2007). Thus, although originally proposed as an alternative to the standard clustering it has been decided that would not contribute to the conclusions of this research and should form a future agenda where the same research questions are asked for more detailed geographies, such as Local Authorities (Webber and Craig, 1978).

3. Analysis and Results

This section contains a discussion of the clustering exercise. Initially a descriptive analysis of the data is presented. This assists in the selection of the appropriate indicators for the clustering based on statistical inference theory. The descriptive analysis looks at the structure of the data and the correlation between the variables. Out of the 20 indicators available at this stage only 11 were appropriate for the clustering in statistical terms.

Furthermore, these 11 variables were at the scrutiny of the ReRisk research team and the energy experts participating in the ReRisks Workshop III held in Bilbao. Based on the expert's opinion about the appropriateness of each indicator and their ranking of the indicator's importance in terms of the policy implications of the results of this project, only 9 indicators finally selected for the clustering.

3.1. Data Assessment

This section discusses issues in relation to the data appropriateness for input into the clustering exercise and does not replace any previous discussion on data issues.

3.1.1. List of Indicators

For the clustering analysis 20 indicators for 287 Regions in Europe have been available. The names of these indicators are presented in Table 1 grouped in five categories. The indicators in bold fonts pass both the statistical and policy relevance tests, whereas those in italic fonts failed the latter tests. Indicators in standard fronts found inappropriate for analysis.

Table 1 Final set of indicators

Category	Indicators
Climate conditions	**Mean maximum temperature July (Max T July)**
	Mean minimum temperature January (Min T Jan)
	Mean annual temperature (Mean T)
	Mean maximum annual temperature (Max T)
	Mean minimum annual temperature (Min T)
Economic structure	**% employment in industries with high energy purchases**
	% of GVA in industries with high energy purchases
	Private energy use
Transport dependency	**Spending on transport fuel for freight as % of GDP**
	Population commuting to other regions / population working in the same region
	Employment in the transport sector as % of total employment
	Age of car park (Average age of cars)
	Number of passengers travelling by air / total population
Social dimension	**Long-term unemployment rate**
	Disposable income in households
	Age dependency ratio
	Economic activity rate
Production potential of renewables	**Wind Power Energy Potential 2005**
	PV potential
Other	Region Area Size

3.1.2. Exploratory Data Analysis

In this section an attempt to analyse the data and provide information on their distribution and spatial structure is being made. The Interim Report discusses data issues in relation to their sources, relevance to the analysis and some indication of areas with high or low values. In this section a thorough investigation of the data on 20 variables is presented. The investigation answers questions about the quality of the variables in terms of statistical inference (i.e. normal distribution, non correlation) in order to assess their appropriateness for analysis using clustering algorithms.

Descriptive Statistics of all indicators

In order to understand the distribution of each variable a descriptive analysis of the data is necessary. Table 2 shows the basic descriptive statistics of all 20 variables listed above.

Table 2 Descriptive Statistics

Table 2a. Climate Conditions indicators

Statistics		Mean T	Max T	Min T	Max T July	Min T Jan
N	Valid	264	264	264	264	264
	Missing	23	23	23	23	23
Mean		11.00641	22.02166	0.27459	31.62409	-8.2607
Median		10.50567	21.62667	-0.06917	31.70333	-8.13333
Mode		2.070714	13.31071	-12.7894	31.24667	-12.5333
Std. Deviation		2.949196	3.261765	3.608596	3.565457	5.75646
Variance		8.70	10.64	13.02	12.71	33.14
Skewness		0.50	0.07	0.37	-0.20	-0.24
Kurtosis		1.07	-0.09	2.49	-0.17	0.76
Range		17.58	17.07	25.72	18.05	37.93
Sum		2905.69	5813.72	72.49	8348.76	-2180.83

Table 2b. Economic structure indicators

		% of employment in industries with high energy purchases	% of GVA in industries with high energy purchases	Private energy use
N	Valid	268	217	262
	Missing	19	70	25
Mean		4.42	8.29	1096.64
Median		3.91	7.83	1140.77
Mode		0.41	1.15	1102.44
Std. Deviation		2.61	4.27	351.81
Variance		6.84	18.20	123772.99
Skewness		1.27	0.82	1.39
Kurtosis		1.93	0.95	11.99
Range		13.81	23.99	3378.54
Sum		1184.89	1798.20	287319.07

Table 2c. Transport dependency indicators

		Air passengers/ population	% Workers commuting	% Employment in transport	Average age of cars	% Fuel costs of freight transport
N	Valid	197	259	269	40	262
	Missing	90	28	18	247	25
Mean		2.91	9.46	10.10	10.46	2.53
Median		1.08	5.40	8.83	9.18	2.11
Mode		0.00	0.00	0.00	7.92	0.50
Std. Deviation		5.05	12.60	6.10	1.73	3.00
Variance		18.79	158.88	37.21	9.03	3.01
Skewness		0.05	3.39	3.70	1.86	0.03
Kurtosis		8.24	15.64	22.28	-1.22	7.61
Range		28.60	98.22	60.26	11.47	14.18
Sum		574.23	2449.96	2716.83	418.31	663.96

Table 2d. Social dimension indicators

		Long term unemployment rate	Economic activity rate	Disposable income in households	Age dependency ratio
N	Valid	271	269	228	277
	Missing	16	18	59	10
Mean		39.22	57.53	13316.31	24.59
Median		40.19	57.50	14294.40	24.58
Mode		15.17	53.90	3146	6.16
Std. Deviation		6.26	16.15	4186.52	21358.09
Variance		260.77	39.18	17526950.55	25.54
Skewness		0.25	0.06	-0.64	3.27
Kurtosis		-0.68	0.83	-0.4	0.48
Range		85.41	39.70	18956.70	36.22
Sum		10628.35	15474.60	3036118.3	6810.84

Table 2e. Production potential of renewable and other indicators

		Wind Power Energy Potential	PV potential	Region Area
N	Valid	279	256	287
	Missing	8	31	0
Mean		142525.10	979.24	16942.37
Median		73938.80	892.59	9970.33
Mode		0	839.51	8.30
Std. Deviation		207470.20	190.71	4.33
Variance		43043876014.07	36368.63	456167986.37
Skewness		3.70	0.97	2.58
Kurtosis		19.16	-0.14	14.38
Range		1795408.00	830.09	164653.77
Sum		39764494.28	250686.11	4862460.95

Descriptive statistics analysis provides a first indication on the data distribution, outliers and missing values. It provides the big picture of the dataset and it allows for a first evaluation of the dataset’s ability to assist in classifying EU Regions into meaningful clusters. It is apparent that several indicators have a significant number of missing values. For example the variable "average age of cars" should not be included in the analysis because less than one in seven values is available.

Since the data are spatial, one should also look at their spatial structure. In order to do this it is necessary to produce maps and apply spatial autocorrelation diagnostics to each variable. In order to assess spatial autocorrelation, the global and local Moran's I statistics are calculated (Anselin, 2003, 2004; Cliff and Ord, 1973, 1981; Moran, 1948).

In the following section the more detailed exploratory analysis for each variable is presented for each indicator. Due to the limitation of this document the Figures and Maps are presented in Annexes I and II, respectively.

3.1.2.1 Climate Conditions indicators

All five Climate Conditions indicators have a rather good shape normal distribution but there are some outliers due to the diversity of climate in Europe (Figure 1 in Annex I). This data originally available in monthly averages for the period 1994 -2008 for each NUTS II EU Regions formed these variables. Correlation analysis indicates that all but two are highly correlated with each other.

Table 3 Climate variables correlation

Correlations

		Mean T	Max T	Min T	Max T July	Min T Jan
Mean T	Pearson Correlation	1	.874**	.928**	.670**	.804**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	264	264	264	264	264
Max T	Pearson Correlation	.874**	1	.660**	.913**	.481**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	264	264	264	264	264
Min T	Pearson Correlation	.928**	.660**	1	.399**	.933**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	264	264	264	264	264
Max T July	Pearson Correlation	.670**	.913**	.399**	1	.147*
	Sig. (2-tailed)	.000	.000	.000		.017
	N	264	264	264	264	264
Min T Jan	Pearson Correlation	.804**	.481**	.933**	.147*	1
	Sig. (2-tailed)	.000	.000	.000	.017	
	N	264	264	264	264	264

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The indicators that are not correlated and thus suitable for clustering analysis is *Mean maximum temperature July* and *Mean minimum temperature January* (Maps 1 & 2 in Annex II). These indicators are also important for the analysis in terms of energy since very hot areas have high energy demands for cooling and very cold areas for heating.

More specifically, Mean maximum July temperature is relevant for identifying the regions with high cooling demand in the summer time and will become more important as temperatures rise as a consequence of climate change.

Mean minimum January temperature is equivalent to regional demand for heating in the winter. All temperature-related data was facilitated by JRC Ispra - IPSC - MARS Unit.

3.1.2.2 Economic structure indicators

% of employment in industries with high energy purchases

Values above 10% proportion of employment in industries with high energy purchases come out as outliers. These are located in most of North Italy and the Czech Republic (Figure 2 and Map 3).

% of GVA in industries with high energy purchases

There are 70 values missing in this variable and the spatial patterns presented in Map 4 do not show some clear cut spatial trends. The highest values appear to be in regions in the Czech Republic and the Netherlands. These observations suggest an expected poor performance in clustering analysis if this variable is included.

Private energy use

The distribution is normal with an outlier located in Luxemburg. The high values are located in central European Regions, South West England, Wales and South Finland and Sweden Regions. Lower values are found in the EU Regions located in East Europe, the Balkans and Cyprus (Figure 4 and Maps 5 & 6).

3.1.2.3 Transport dependency indicators

Spending on transport fuel for freight as % of GDP

Regions in Bulgaria and Romania and generally region in East Europe and Spain appear to exhibit significantly higher values in the proportion of fuel costs of freight transport than the 2.53% of the EU Regions average. The former regions have thus a higher vulnerability in fuel prices (Figure 5 and Map 7).

Population commuting to other regions / population working in the same region

This variable has also high positive kurtosis result in a number of outliers (Figure 6). These are regions in Belgium and the UK where many people live in a different region than they commute. This observation may help to pick up some effects but in terms of statistical analysis is prone to bias because of the shape of the distribution.

The spatial patterns also show high proportions of communing in central Europe and less in the peripheries (Map 8). This is rather expected because of the high development of transport networks in central Europe and mainly in industrial regions as opposed to the regions in the Balkan Peninsula.

Employment in the transport sector as % of total employment

The proportion of people working in transport related jobs shows some interesting spatial patterns (Map 9). Metropolitan areas such as Bonn, Paris, Rome, Madrid, Prague, and Bratislava have high values. This is probably due to the fact that they are airport hubs and logistics centres service the high populations with consumer goods. There also some other regions such as Corsica perhaps related to the tourism and shipping industry. Some of the above areas are outliers but these values should not be excluded as they represent a reality.

Age of car park (Average age of cars)

This is a variable with only a few values available and is definitely poor for the clustering analysis. However, the few available values of the variable exhibit a good shape normal distribution (Figure 8).

Number of passengers travelling by air / total population

This indicator shows a distribution with a high positive kurtosis (a long tail in the normal distribution curve in Figure 9). The majority of the values are between 0-11.57% (this is mean +/- 2 std. Dev.). Map 11 shows the spatial distribution of the values of the variable for the 197 regions data are available for suggesting a clear EU metropolis and touristic island – rural and less populated EU regions divide. The boxplot in Figure 9 allows for the identification of outliers. This variable may be problematic in terms of statistical inference although it has been chosen as a significant indicator for this project.

3.1.2.4 Social dimension indicators

Long term unemployment rate

There is an apparent strong spatial inequality in the long term unemployment rate. Map 12 shows the spatial distribution of the values of the variable for the 271 regions data is available for. The values range from 0 to 85.50% with a mean value of 39.00% and a rather good shape normal distribution (Figure 10a). There are no outliers (Figure 10b).

The Moran’s I global index is 0.7029 showing a strong spatial autocorrelation. This suggests that there are spatial clusters of similarly high or low values formed by regions in certain parts of Europe. Map 13 shows the Local Indicators of Spatial Association (LISA) helping the researcher to identify spatial clusters of high values (Germany, Eastern and South-eastern Europe) and low values (Nordic countries, Spain, North Italy and the UK).

Disposable income in households

The variable exhibits a good shape normal distribution (Figure 11). The minimum value of 3146.00 Euros appears as a low outlier and is a region in Romania (Figure 11). There are several missing values, however this is an important value for the analysis and could not be omitted.

The spatial distribution of the data does not show any unknown patterns. There is an obvious East – West Europe divide and a Great Metropolitan Areas – countryside regions divide (Map 14).

Age dependency ratio

In terms of statistical inference, this is a generally well performing variable (Figure 12). The maps suggest a South Europe Rest-of-Europe divide. Most regions in South European countries including France and some regions in Central-East Europe are regions where a high ratio of age dependency is observed (Maps 15 & 16).

Economic activity rate

In terms of statistical inference, this is a generally well performing variable (Figure 13). The maps suggest a North-South Europe divide. Most regions in North European countries including Iceland and some regions in Northeast Europe are regions exhibiting high economic activity rate (Maps 17 & 18).

Social Indicators correlation analysis

The analysis bellow tries to identify any correlation between the social indicators available to this project. These four indicators are: long term unemployment rate; disposable income in households; economic activity rate; and age dependency ratio. Table 4 shows the Pearson's correlation coefficients for all possible pairs of these indicators. It is apparent that there is a significant strong correlation between the Economic activity rate and the Long term unemployment rate. This is also confirmed in Figure 14 which shows a scatter plot of the two variables. In terms of statistical inference this means that only one of the two variables should take part in clustering.

Table 4 Social Indicators correlation

Correlations

		Long term unemployment rate	Disposable income in households	Economic activity rate	Age dependency ratio
Long term unemployment rate	Pearson Correlation	1	-,296**	-,463**	-,126*
	Sig. (2-tailed)		,000	,000	,040
	N	271	228	268	266
Disposable income in households	Pearson Correlation	-.296**	1	,373**	,345**
	Sig. (2-tailed)	.000		,000	,000
	N	228	228	227	225
Economic activity rate	Pearson Correlation	-.463**	,373**	1	-,260**
	Sig. (2-tailed)	.000	,000		,000
	N	268	227	269	266
Age dependency ratio	Pearson Correlation	-.126*	,345**	-,260**	1
	Sig. (2-tailed)	.040	,000	,000	
	N	266	225	266	277

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

3.1.2.5 Production potential of renewable and other indicators

Wind Power Energy Potential 2005

This variable represents the potential energy from wind power stations and is a multiplication of the area size and the wind potential density. The values are very high and the distribution very skewed. Several high values appear as outliers (Figure 15) mainly referring to regions in the Nordic countries (Map 19). This variable if not normalised in prone to dominate the cluster analysis.

This original data on wind intensity in the regions was prepared in GIS format by the European Topic Centre on Air and Climate change (ETC/ACC), led by PBL the Netherlands, on request of the EEA (EEA, 2009). It has been converted to NUTS 2 level by the NTUA researchers, who collaborate in the ReRisk project and the help of the ESPON database project (ECT-LUSI from UAB). It identifies those regions in Europe, which have the highest potential for producing electricity from wind power. However, the EEA has introduced some restrictions when calculating the maximum potential, mainly due to environmental reasons. ReRisk has followed these recommendations, using the "restrained" wind potential for the regional analysis.

PV potential

The regional potential for produce electricity from PV panels has been calculated and supplied by the Joint Research Centre's Sunbird data base, which forms part of the SOLAREC action at the JRC Renewable Energies Unit. The data refers to the yearly total of estimated solar electricity generation (for horizontal, vertical, optimally-inclined planes) [kWh] within the built environment.

This is a generally well performing variable (Figure 16). Spatial patterns suggest a North – South Europe divide (Map 20).

Region Area Size

This variable is the result of own calculations of the region's area using Geographical Information Systems (GIS) analysis. GIS extracts the geographical area of the polygon that represents each region. This resulted in a variable with values in all regions that replaced the original variable. However the region areas size is expected to be highly correlated with the wind energy potential.

3.2. Data normalisation and weighting

It is common place in the geography literature about clustering that data in their original form may not result in efficient clusters if the variables that are included in the analysis have data with very different means and variances and their distributions are skewed (e.g. Batagelj et al., 2006; Harris et al., 2005; Milligan and Cooper, 1988; Su et al., 2009).

Harris et al. (2005, p. 152) suggest that "In an ideal world we would include as clustering variables only those which have a bell curved, normal (or Gaussian) distribution. In practice many important dimensions that need to be included in a classification are not normally distributed."

Indeed in our analysis many variables have skewness and kurtosis statistics that indicate a problematic normal distribution. In a normal distribution the skewness and kurtosis should be 0 (Kalogirou, 2003). A near zero values, such as for climate and socioeconomic indicators is acceptable. However there are indicators such as wind energy potential and % commuters that exhibit high values of skewness and kurtosis.

In order to address this issue it is necessary to standardise or normalise the data. Milligan and Cooper (1988) suggest seven methods of standardisation. Here we employ two methods of standardisation, the z-score (Formula 1) and the normalisation to the sum of the values (Formula 2).

The z-score is very common in the literature. "The z-score method addresses the differential scale of the original variables by transforming the variables to have unit variance; however, the z-score method places no specific restrictions on the ranges of the transformed variables." (Su et al., 2009, p. 281). The formula for calculating the z_i score for the value x_i of an indicator X is

$$z_i = \frac{x_i - \bar{x}}{\sigma} \tag{1}$$

where $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ is the mean and $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$ is the variance of X. The results of the application of this method in our dataset are presented in Section 3.3.4.

Milligan and Cooper (1988, p. 185) recognise that "standardization based on normalizing to the sum of the observations has been suggested: $Z_6 = X/\Sigma X$. Formula Z_6 will normalize the sum of the transformed values to 1.00 and the transformed mean will equal 1 / n. As such, the mean will be constant across variables, but the variances will differ." The formula for calculating the q_i score (the Z_6 proposed by Milligan and Cooper) for the value x_i of an indicator *X is*

$$q_i = \frac{x_i}{\sum_{i=1}^{N} x_i} \tag{2}$$

The q-score defined here ensures that the indicators will have an equal influence in the clustering. This allows the easy use of the weights for the indicators. By just multiplying each value of the indicator with the indicator's weight, we ensure that the indicator with the highest weight will have more influence in the convergence of the k-means algorithm. The results of the application of this method in our dataset are presented in Section 3.3.4.

According to Harris et al. (2005, p 162) "The next stage in the clustering process used by Experian involves the calculation of the means and standard deviations of the input variables, and the standardization of the data. An important feature of this process is that these, and all subsequent, computations are population weighted. That is to say that when calculating the means and standard deviations the algorithm gives correspondingly more attention to the values of zones with high populations than to those with low." Thus, the new z-scored note pz_i bellow for the value x_i of an indicator X is

$$pz_i = \frac{w_i x_i - \bar{x}}{\sigma} \tag{3}$$

where $\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$ is the mean and $\sigma^2 = \frac{\sum_{i=1}^{N} \left(w_i x_i - \bar{x}\right)^2}{\sum_{i=1}^{N} w_i}$ is the variance of *X* and w_i is the weight assigned to the ith EU Region and is proportional to that Region's population count.

3.3. Clustering Results

The first attempt for clustering assumed that variables should be independent, should have less than 20% missing values and should have a good shape normal distribution. However no outliers where removed from the variables. The introduction of weighting of the indicators appears in just one clustering exercise.

In this section the results of four independent clustering applications are presented:

1. A k-means with 4 clusters on 9 original indicators
2. A k-means with 20 clusters based on 9 original indicators
3. A k-means with 4 clusters based on the z values of the original indicators
4. A k-means with 4 clusters based on the weighted normalized values of the original indicators

The variables included in the clustering are:

- **Climate conditions**
  - Mean maximum temperature July
  - Mean minimum temperature January
- **Economic structure**
  - % employment in industries with high energy purchases
- **Transport dependency**
  - Fuel costs of freight transport
  - % workers commuting
- **Social dimension**
  - Long-term unemployment rate
  - Disposable income in households
- **Production potential of renewables**
  - Wind power potential
  - PV potential

Economic activity rate and private energy used were those variable excluded due to high correlation with some of the indicators we included. Age dependency ration and % employment in transport excluded as a result of low score from energy experts.

3.3.1. A 4-clusters k-means

The k-means clustering algorithm available in commercial statistics software was applied (SPSS) to the dataset consisting of the above 9 indicators. For this k-means clustering exercise the following assumptions were made:

- Number of clusters: Four clusters
- For the algorithm convergence: 100 maximum iterations and conversion at 0 using running means
- For the missing values: exclude cases pairwise

The cluster centres are presented in Table 5 along with the number of regions that were assigned to each cluster. Figures 3.1 – 3.7 present the data of Table 5 for each group of variables by mean of spider graphs. The latter make it easy to compare the cluster centre with the mean value for each indicator. A brief description of the characteristics of each cluster as a complement to the figures follows. Finally, Map 3.1 shows the membership of each EU Region in one of the four clusters in difference colour.

Table 5 Final cluster centres

Indicator	***Mean***	***Cluster Centres***			
		1	**2**	**3**	**4**
Maximum temperature July	31.62	31.48	26.72	32.09	26.91
Minimum temperature January	-8.26	-8.08	-16.47	-7.55	-28.65
% employment in industries with high energy purchases	4.42	3.332	6.094	4.649	4.106
Fuel costs of freight transport	2.53	3.270	2.415	2.328	1.765
% workers commuting	9.46	7.516	3.281	10.648	1.678
Long-term unemployment rate	39.22	41.88	22.87	39.88	31.38
Disposable income in households	13316.31	12433.91	11290.35	13842.29	11045.80
Wind power potential	142525.07	244988.29	747448.75	54103.16	1795408.00
PV potential	979.24	981.427	816.930	989.996	793.711
Number of Cases		67	17	200	2

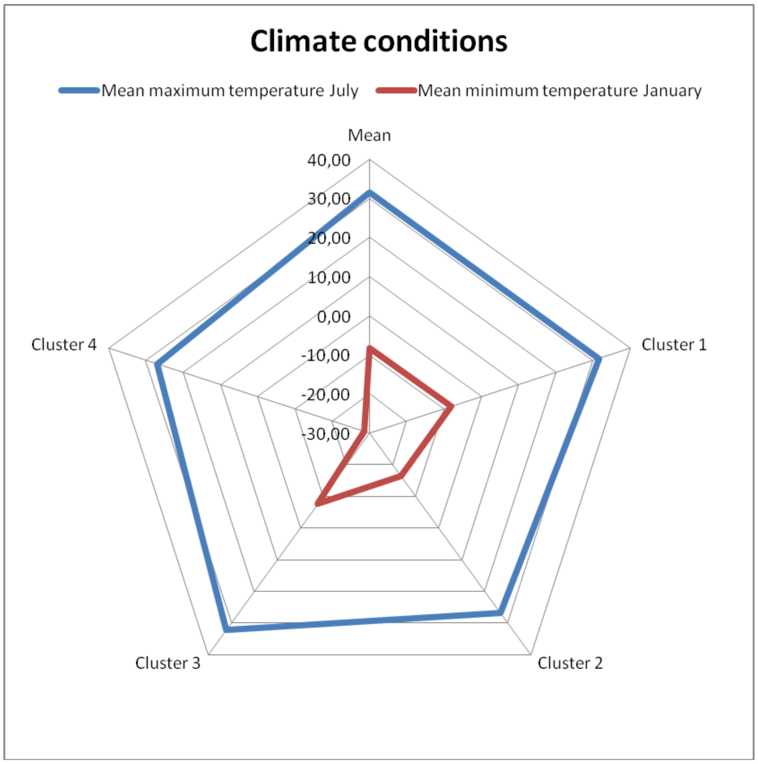


Figure 3.1. Cluster centres spider graph: Climate Conditions

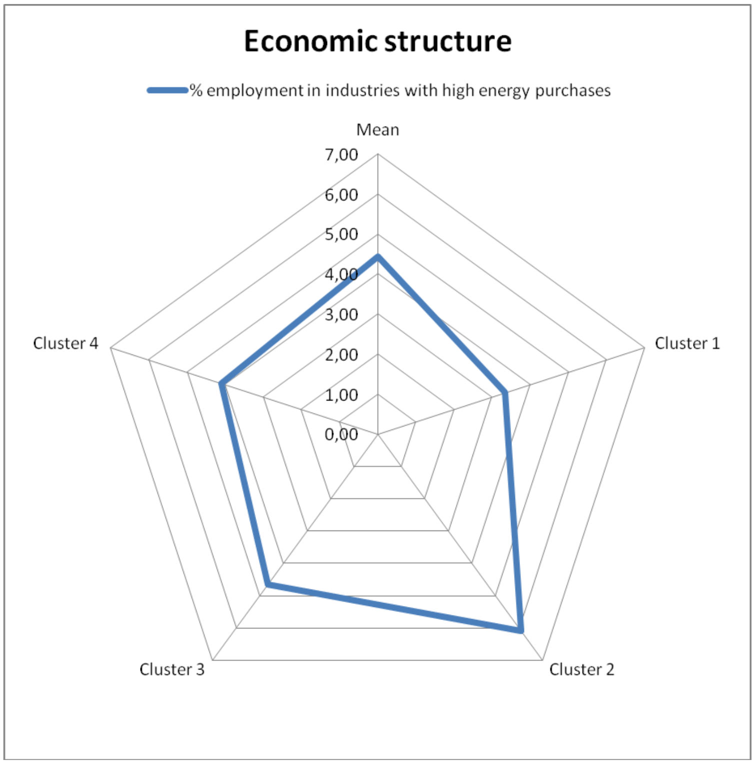


Figure 3.2 Cluster centres spider graph: Economic Structure

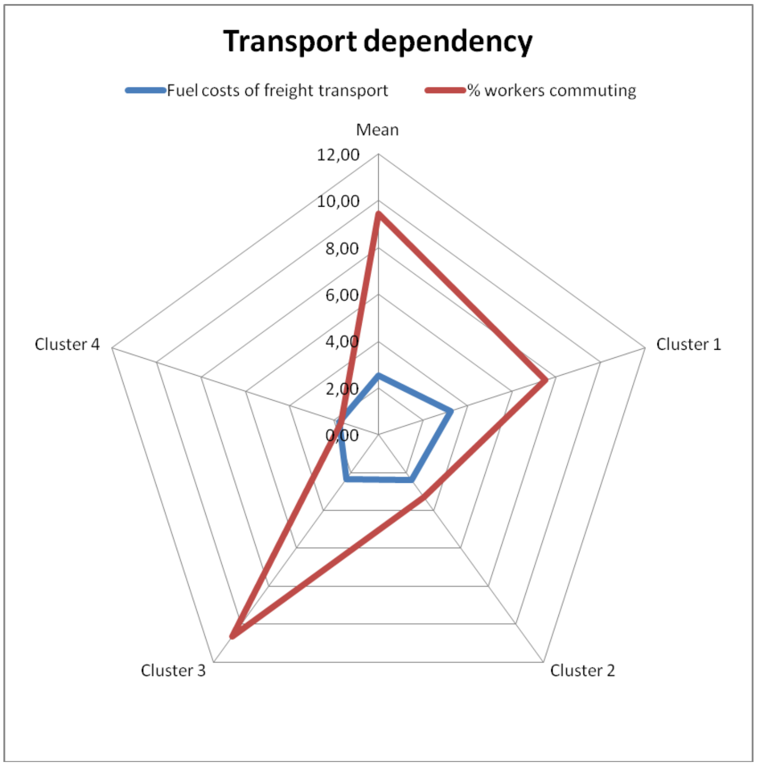


Figure 3.3 Cluster centres spider graph: Transport dependency

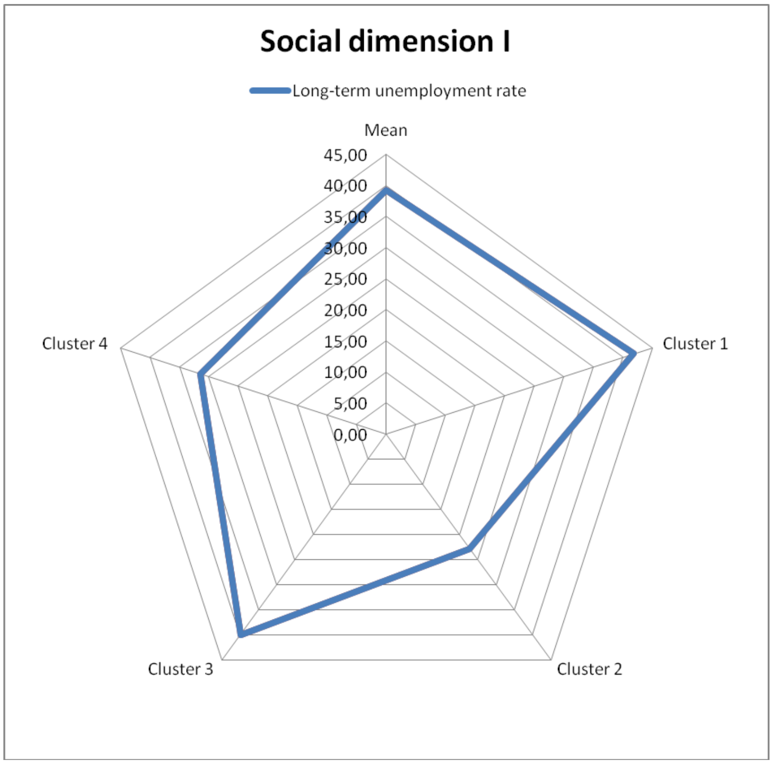


Figure 3.4 Cluster centres spider graph: Social Dimension

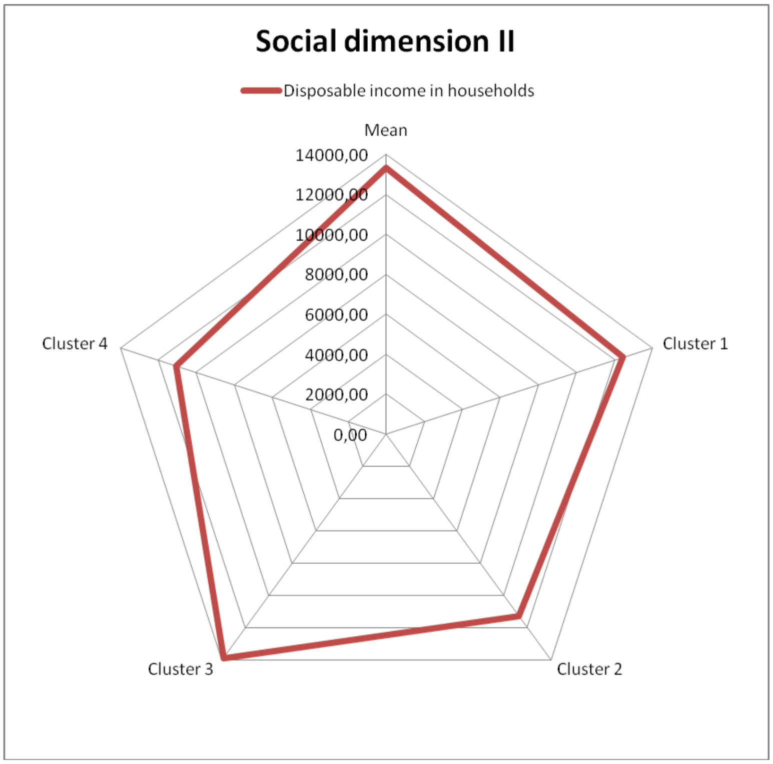


Figure 3.5 Cluster centres spider graph: Social Dimension

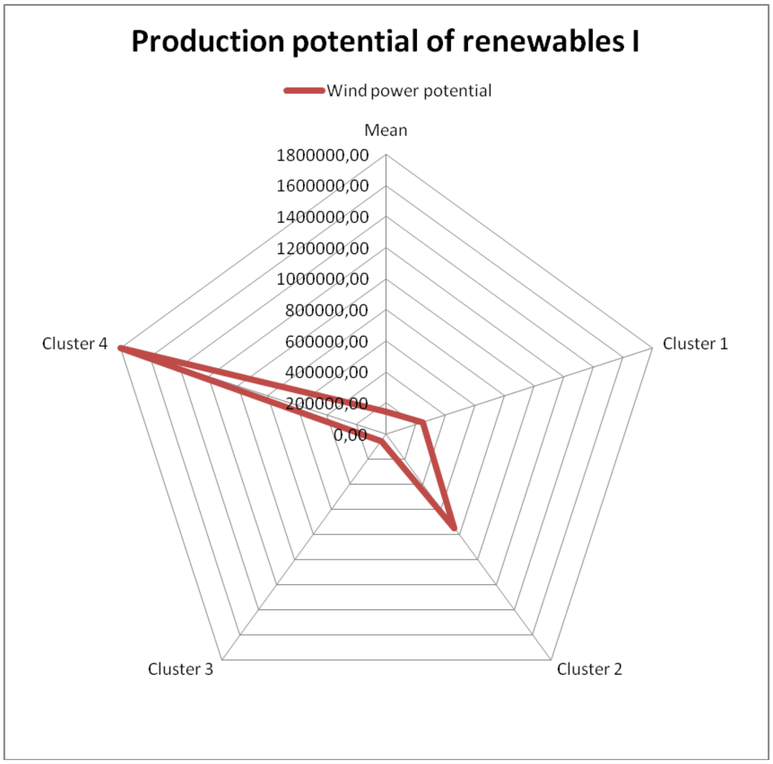


Figure 3.6 Cluster centres spider graph: Production potential or renewables

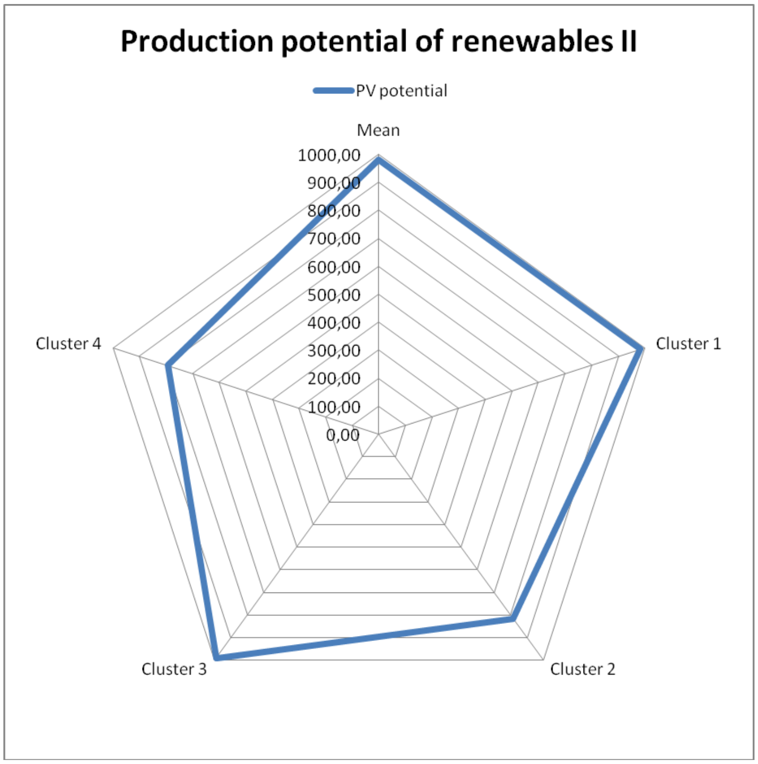
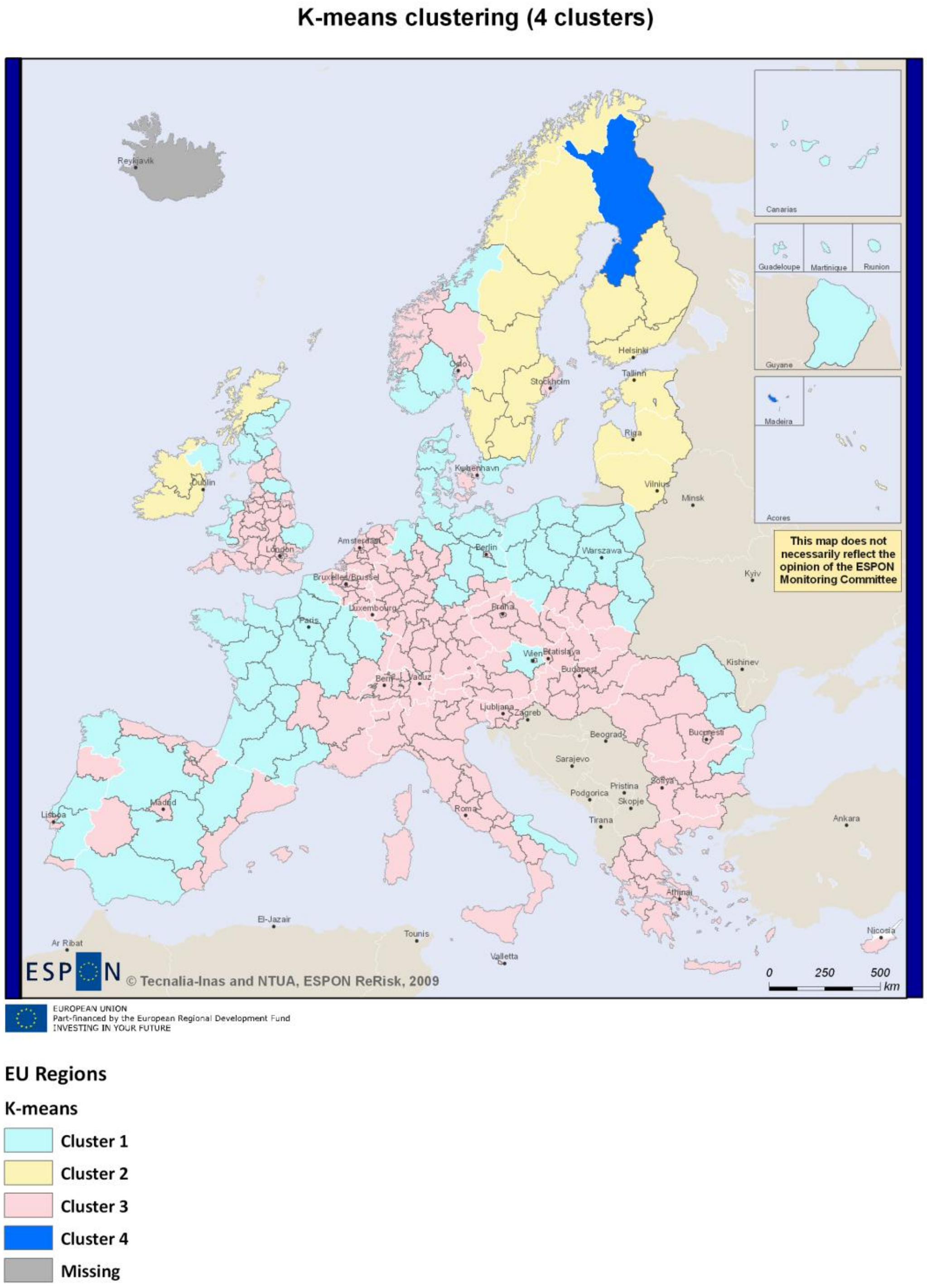


Figure 3.7 Cluster centres spider graph: Production potential or renewables

Description of the characteristics of each cluster

This is a basic clustering without any data processing of the indicators and prone to their differences in terms of the magnitude of the values and distribution of each variable. As a result, variables such as wind energy capacity are expected to dominate. Cluster 1 appears to represent the average EU regions in terms of most of the 9 indicators except the economic structure and transport dependency variables. Most of the regions are rural. Cluster 2 differentiates significantly from Cluster 1 in most indicators, especial climate (lower January temperatures), economic structure (higher), transport dependency (lower) and social dimension (lower). The regions classified in Cluster 2 are mainly located in North Europe and have higher wind power potential than Europe’s average region. What differentiates Cluster 3 regions from Europe’s average region is the very low (almost 0) wind power potential and higher commuting. Cluster 4 gets two regions with extreme values in mean January temperature and wind power potential. The assumption that the latter variable dominates the cluster membership is confirmed.



Source: Own elaboration based on Eurostat data

Map 3.1 K-means clustering membership of EU regions (NUTS II): 4 Clusters

3.3.2. A 20-clusters k-means

The same k-means clustering algorithm was applied as in the previous section using the same options. This time the number of clusters is 20.

The cluster centres are presented in Table 6 along with the number of regions that were assigned to each cluster. Figures 3.8 – 3.14 present the data of Table 5 for each group of variables by mean of spider graphs. Map 3.2 shows the membership of each EU Region in one of the twenty clusters in difference colour.

Table 6 Final cluster centres

Cluster	Max July	Min Jan	Empl HEI	Fuel Costs	% Comtg	LT Un	HHold Income	Wind P Energy	PV Output	N
1	26.91	-28.65	4.19	1.76	1.68	27.41	10522.65	1795408.00	793.71	2
2	31.59	-15.75	1.45	2.79	0.73	55.38	8519.60	395949.20	865.77	2
3	27.81	-17.20	7.60	3.00	2.44	29.14	10533.60	621730.67	844.88	3
4	30.79	-8.88	3.19	3.20	3.91	36.47	11499.02	282204.62	939.67	14
5	31.63	-8.76	4.33	2.73	9.27	45.00	13256.02	91363.32	948.14	44
6	27.49	-14.93	7.48	1.53	7.06	16.11	13255.25	545744.00	897.21	2
7	30.94	-5.48	2.13	2.96	3.68	39.49	11912.19	337562.40	951.64	7
8	31.26	-8.21	3.52	3.14	11.24	40.91	12607.99	230451.27	984.81	17
9	32.14	-8.54	4.09	3.52	11.42	40.26	12345.83	175541.00	1011.43	20
10	32.98	-7.18	4.52	1.74	9.00	38.40	14345.59	7658.25	1054.87	48
11	32.38	-6.30	5.08	1.98	14.06	37.41	14428.37	32856.49	1010.36	43
12	31.93	-4.62	2.90	2.35	4.45	38.32	15087.30	424388.00	1060.95	1
13	24.92	-6.99	3.86	2.19	0.81	26.84	11726.37	718091.00	790.53	4
14	31.20	-8.72	4.34	3.35	4.90	39.89	12550.42	128366.90	976.41	21
15	28.09	-20.09	9.17	2.66	4.85	20.46	9871.63	846190.67	838.61	3
16	31.08	-4.80	3.46	2.25	6.80	36.63	15574.00	474980.00	1006.45	2
17	27.44	-25.77	2.93	3.48	3.01	22.77	10884.00	1031076.00	786.75	1
18	25.37	-28.34	7.13	2.79	3.15	14.26	12180.80	1245316.00	815.03	1
19	25.61	-23.24	4.34	3.09	3.05	12.51	13094.60	657586.00	753.50	2
20	31.87	-7.52	4.58	2.44	11.85	41.49	14224.81	59082.80	957.99	49
Mean	31.6	-8.26	4.42	2.53	9.46	39.22	13316.3	142525.07	979.24	

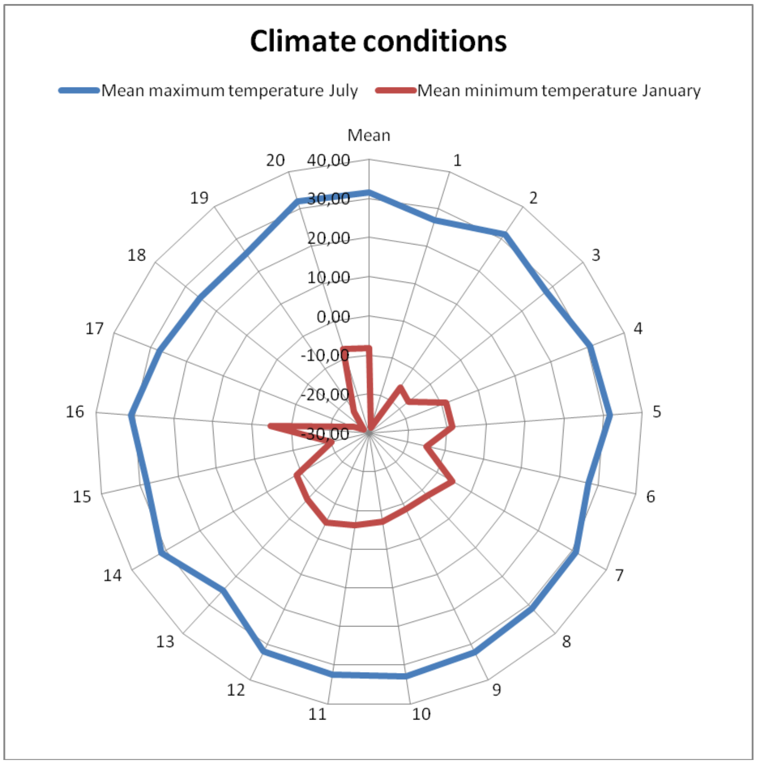


Figure 3.8 Cluster centres spider graph: Climate Conditions

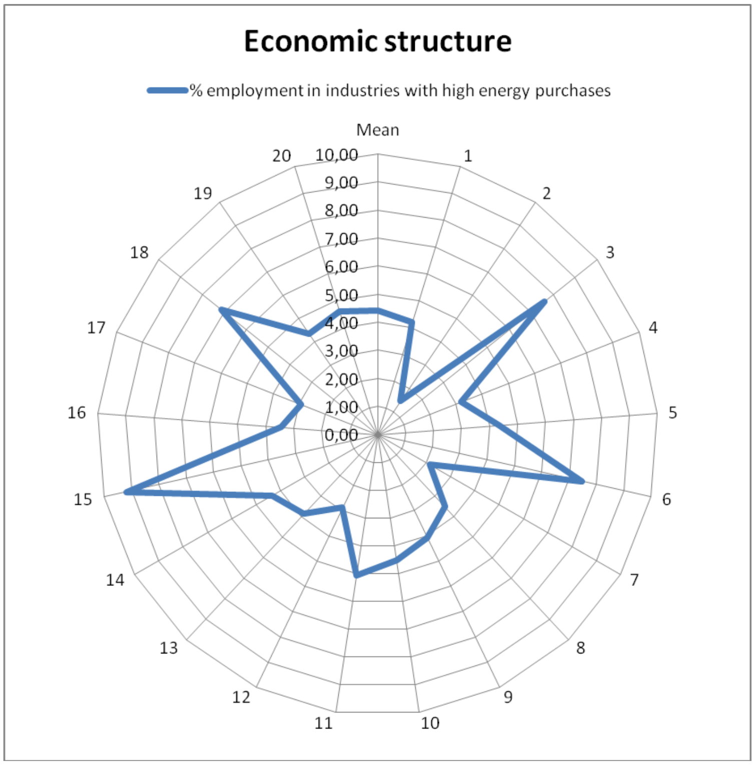


Figure 3.9 Cluster centres spider graph: Economic Structure

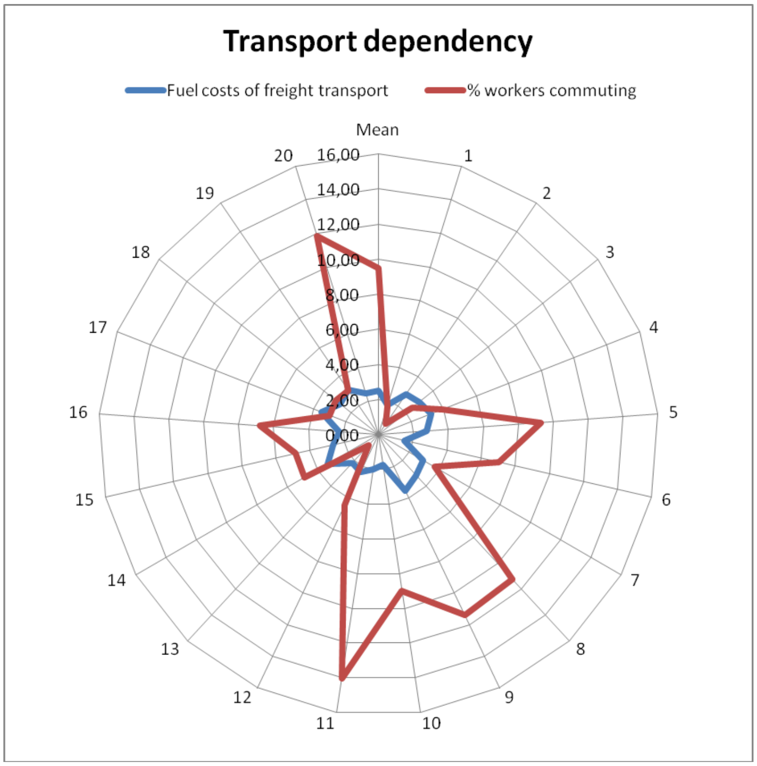


Figure 3.10 Cluster centres spider graph: Transport dependency

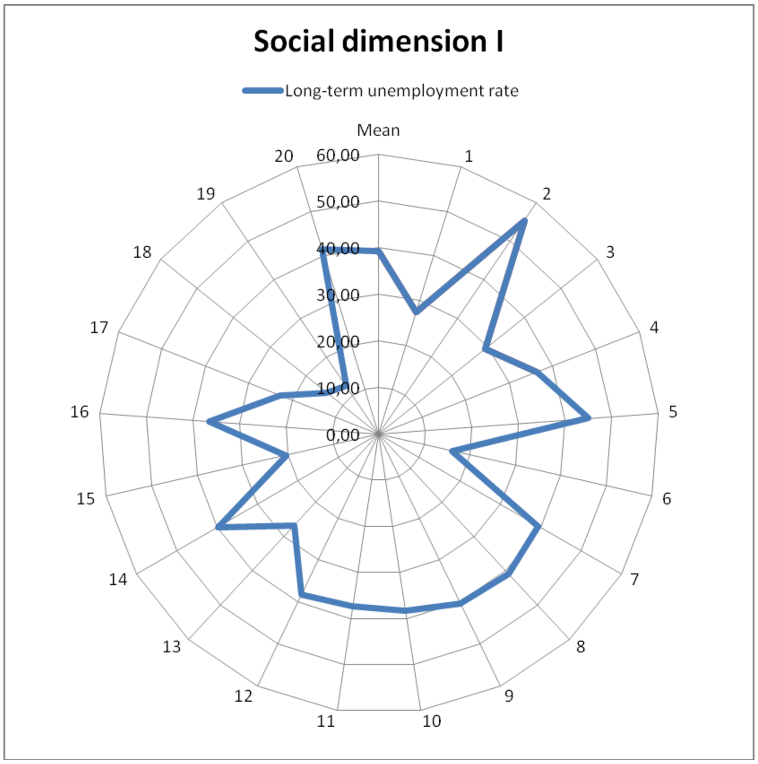


Figure 3.11 Cluster centres spider graph: Social Dimension

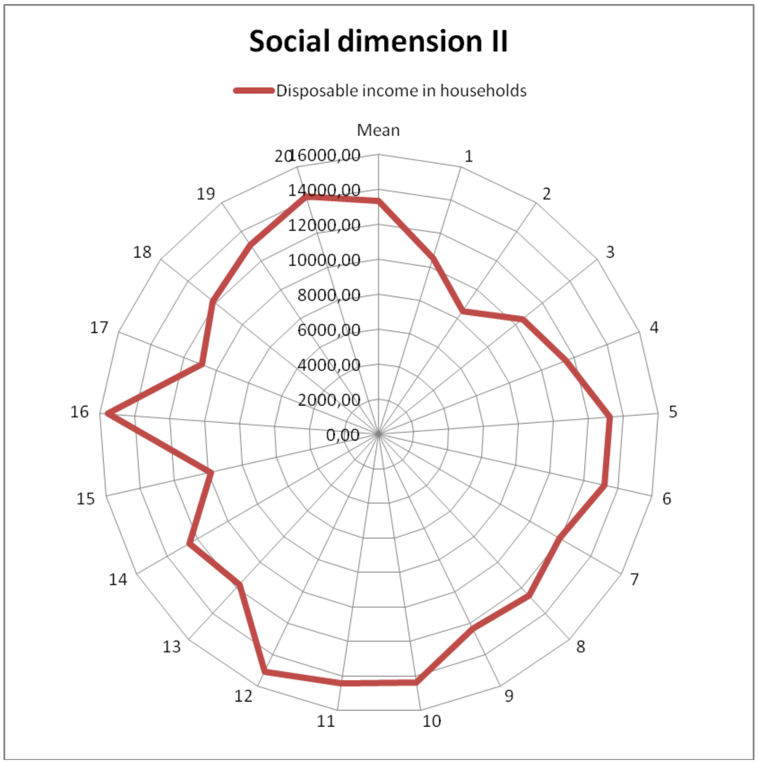


Figure 3.12 Cluster centres spider graph: Social Dimension

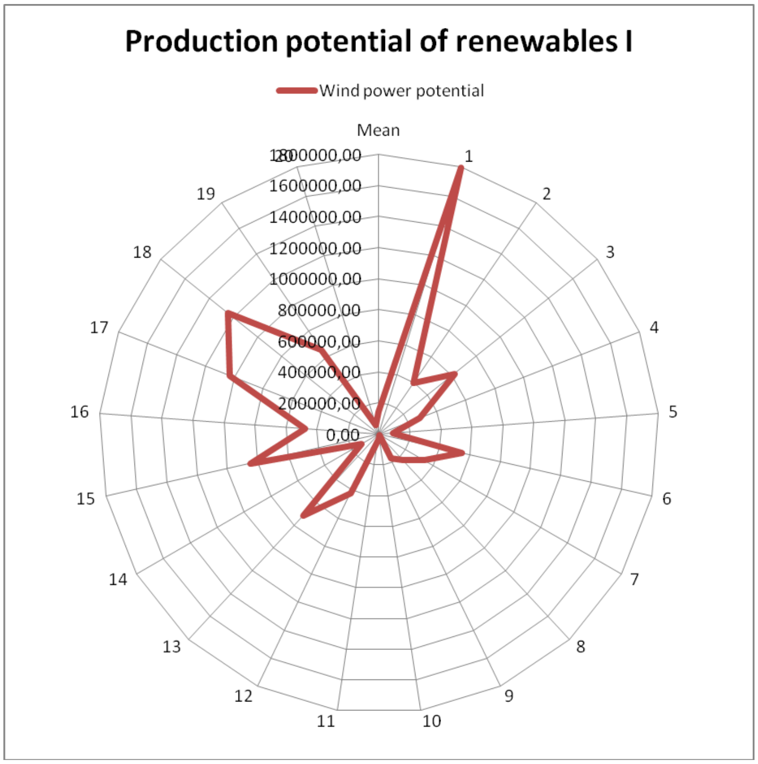


Figure 3.13 Cluster centres spider graph: Production potential or renewables

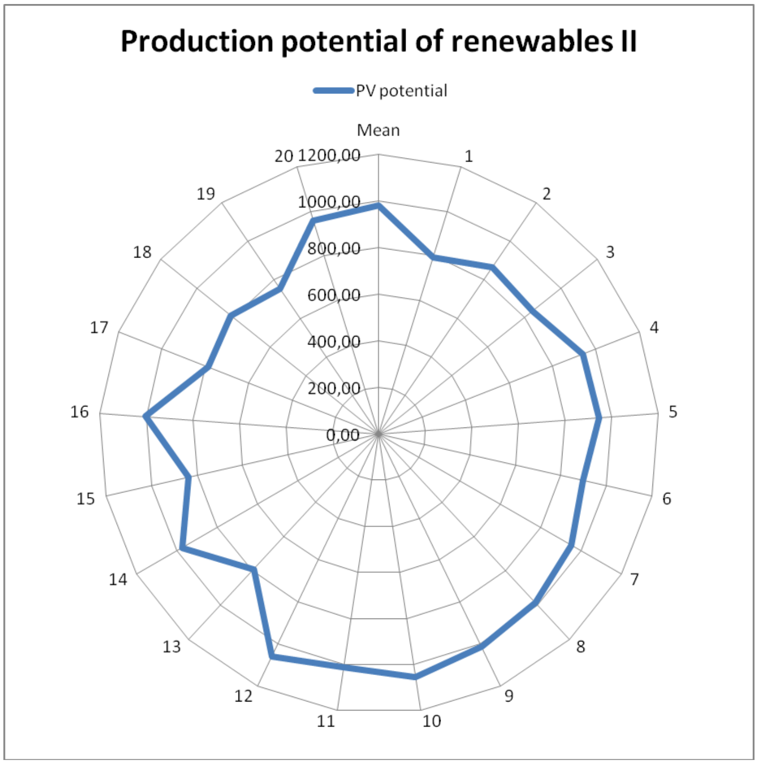
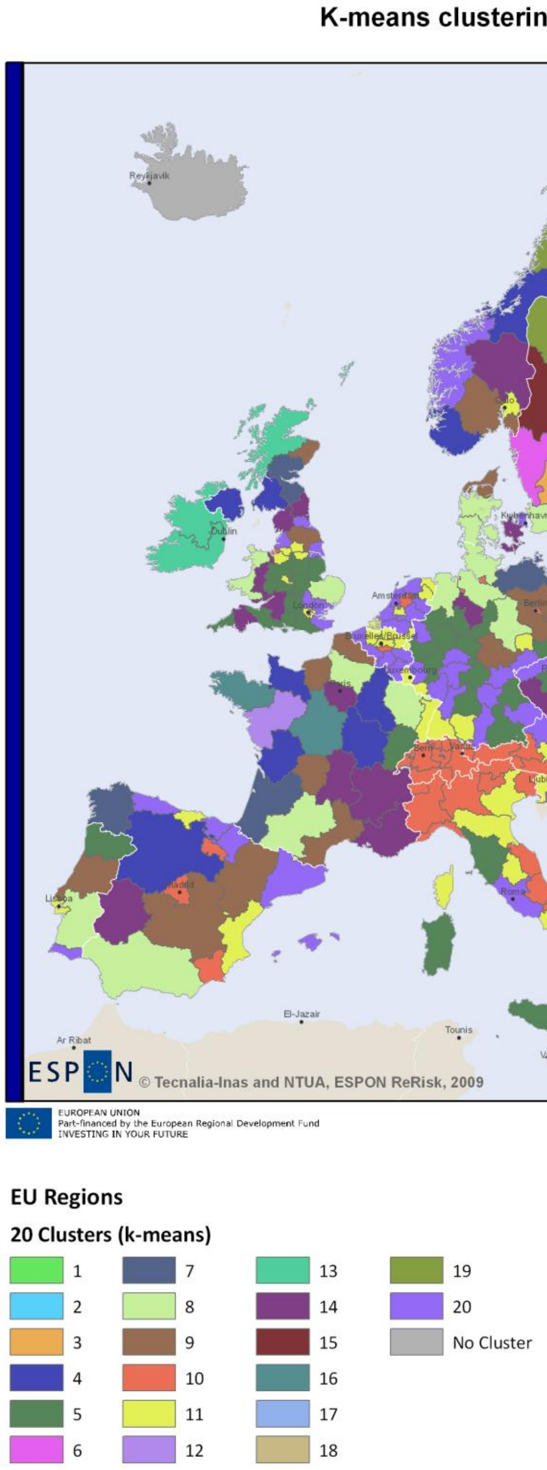


Figure 3.14 Cluster centres spider graph: Production potential or renewables

Description of the characteristics of each cluster

This is a basic clustering without any data processing of the indicators and prone to their differences in terms of the magnitude of the values and distribution of each variable. This time 20 clusters have been chosen, allowing for diversity in the formation of groups of Regions with similar characteristics.

It is necessary to identify single-member clusters such as Cluster 18 and two-members clusters such as Cluster 1 that appear to have high differences in some variables with the mean values for all European Regions. On the contrary, clusters such as Cluster 10 and 20 with several members represent the average EU regions in terms of most of the 9 indicators except the economic structure and transport dependency variables.



Source: Own elaboration based on Eurostat data

Map 3.2 K-means clustering membership of EU regions (NUTS II): 20 Clusters

3.3.3 A 4 clusters k-means of the z values of the original indicators

The same k-means clustering algorithm was applied as in the previous two sections using the same options. The number of clusters is 4 but the instead of the original variables, their z-scores (Section 3.2) where included in the analysis.

The cluster centres of the original variables based on the resulted membership are presented in Table 7 along with the number of regions that were assigned to each cluster. Figures 3.15 – 3.21 present the data of Table 5 for each group of variables by mean of spider graphs. Map 3.3 shows the membership of each EU Region in one of the four clusters in difference colour.

Table 7 Final cluster centres

	Mean	*Cluster Centres*			
		1	**2**	**3**	**4**
Maximum temperature July	31.62	26.57	27.11	30.56	34.52
Minimum temperature January	-8.26	-27.58	-14.66	-6.77	-9.21
% employment in industries with high energy purchases	4.42	5.50	6.11	3.72	5.31
Fuel costs of freight transport	2.53	3.10	2.45	1.74	3.94
% workers commuting	9.46	3.14	3.50	13.79	3.04
Long-term unemployment rate	39.22	17.81	22.54	36.30	48.18
Disposable income in households	13316.31	11273.17	11271.97	15770.37	8451.33
Wind power potential	142525.07	1357266.67	625305.44	93376.80	103240.96
PV potential	979.24	798.50	816.06	920.24	1118.18
Number of Cases		3	17	173	93

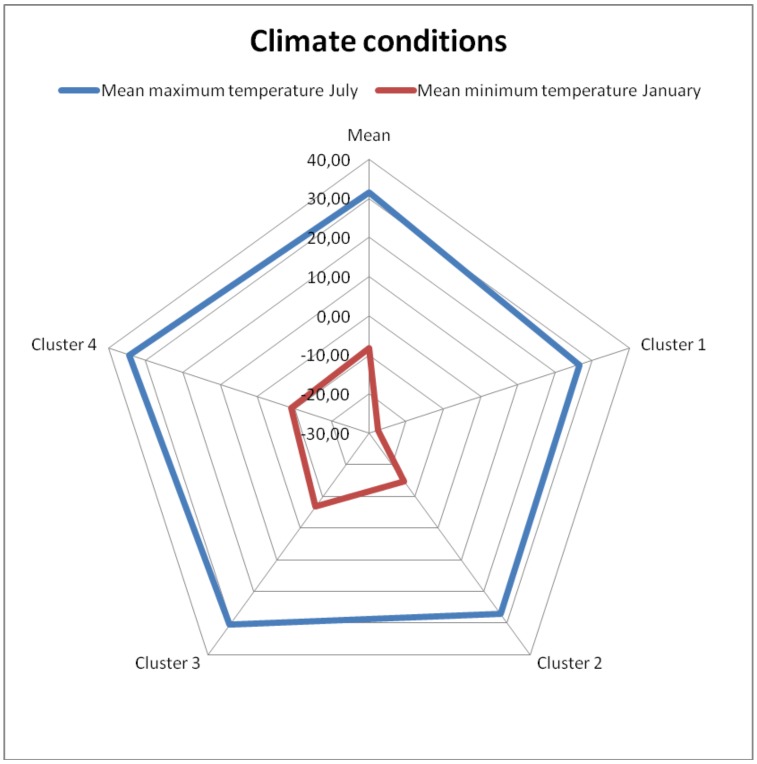


Figure 3.15 Cluster centres spider graph: Climate Conditions

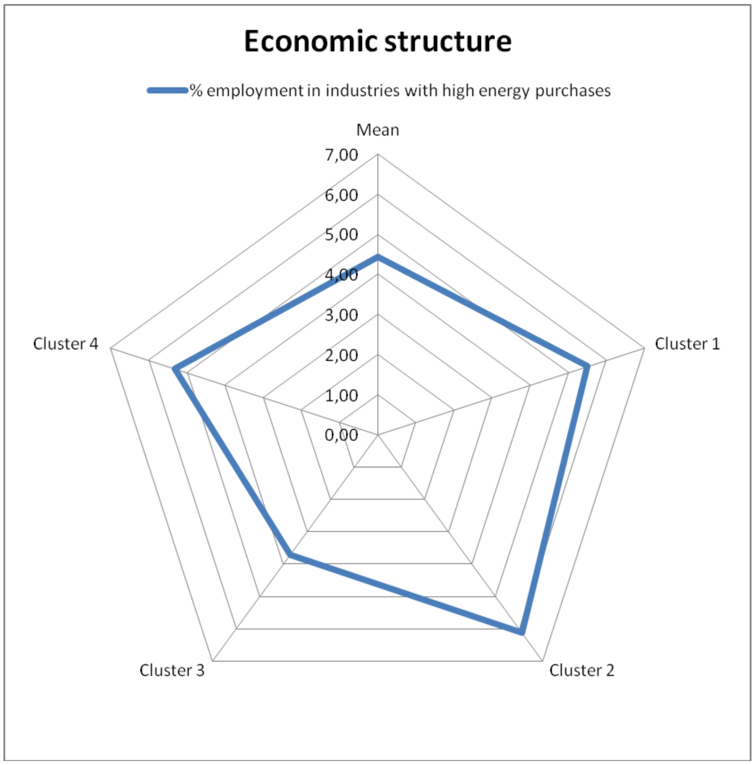


Figure 3.16 Cluster centres spider graph: Economic Structure

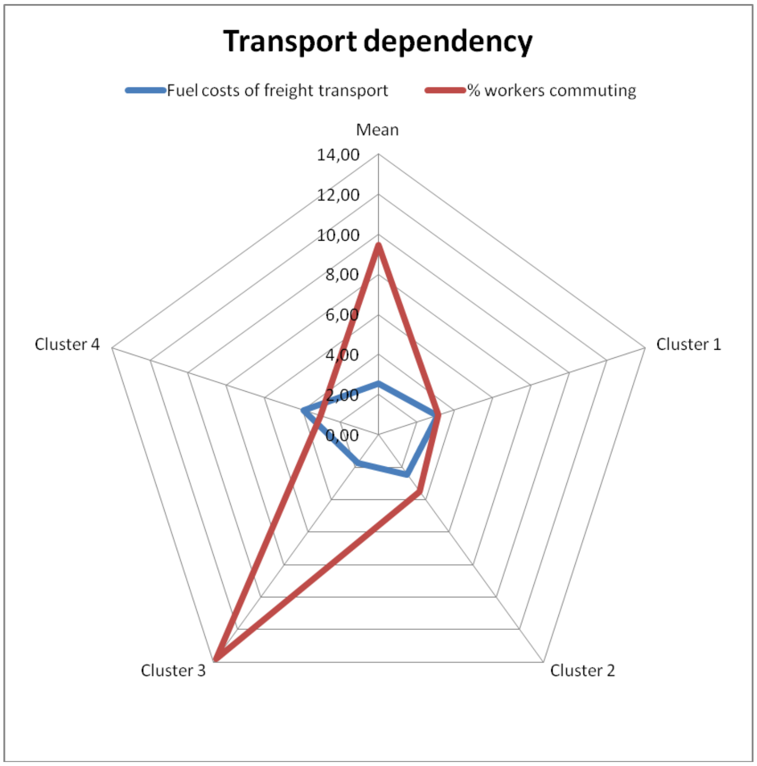


Figure 3.17 Cluster centres spider graph: Transport dependency

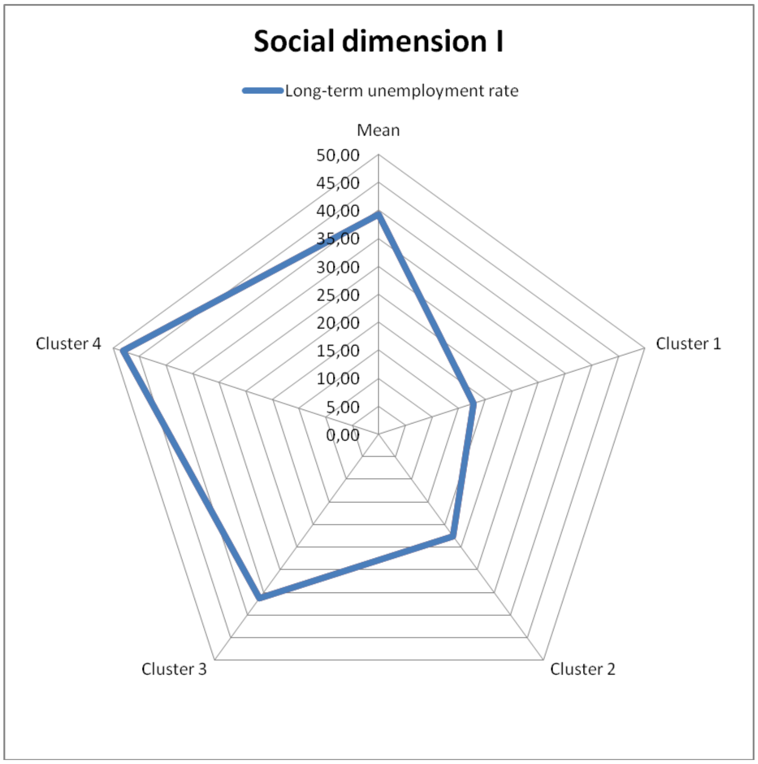


Figure 3.18 Cluster centres spider graph: Social Dimension

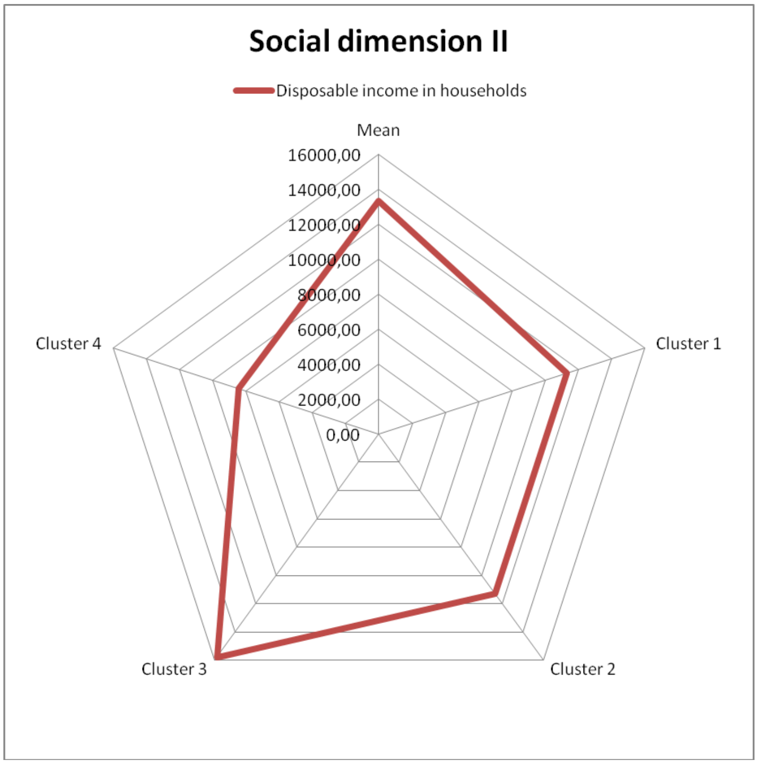


Figure 3.19 Cluster centres spider graph: Social Dimension

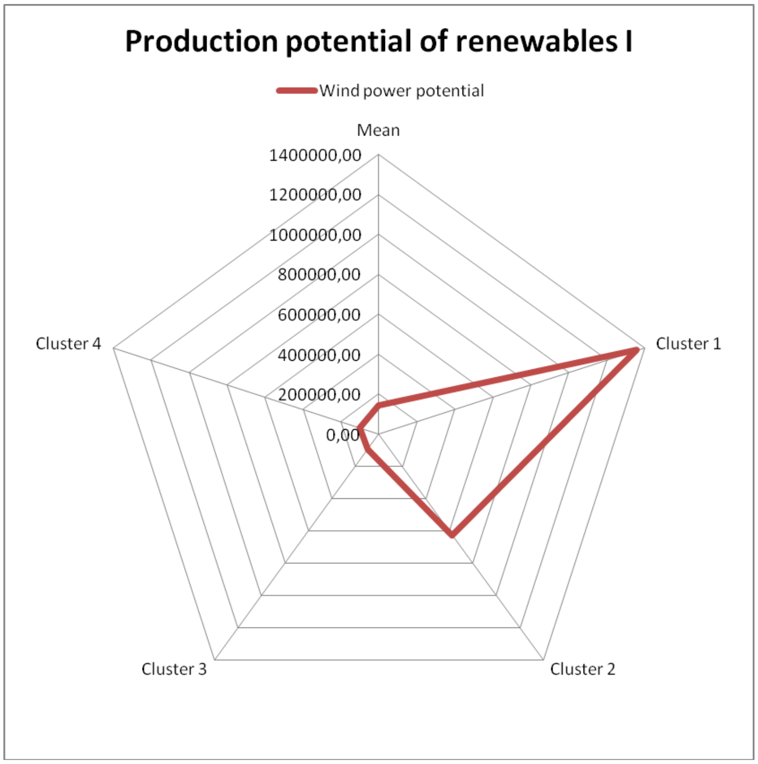


Figure 3.20 Cluster centres spider graph: Production potential or renewables

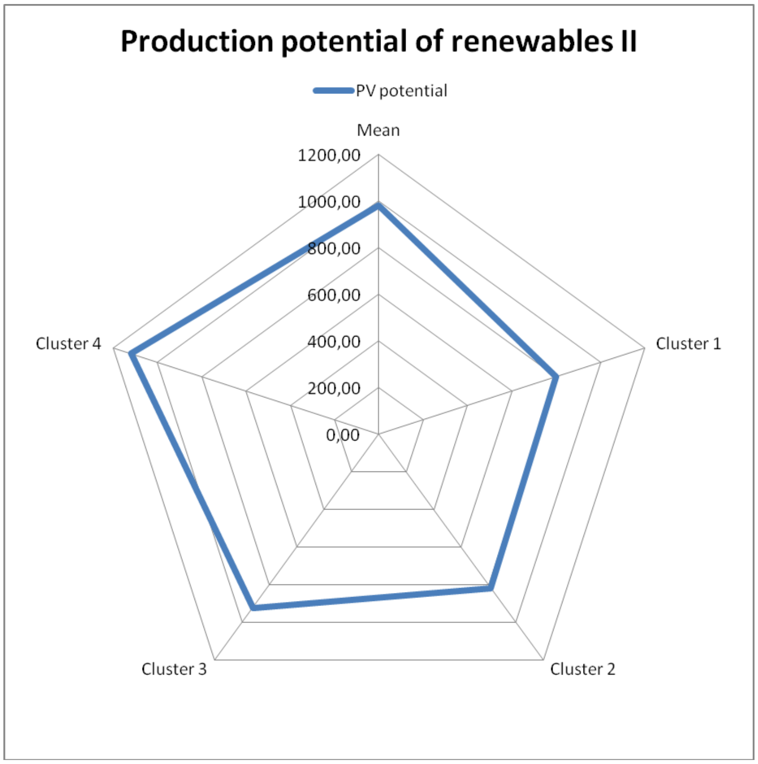


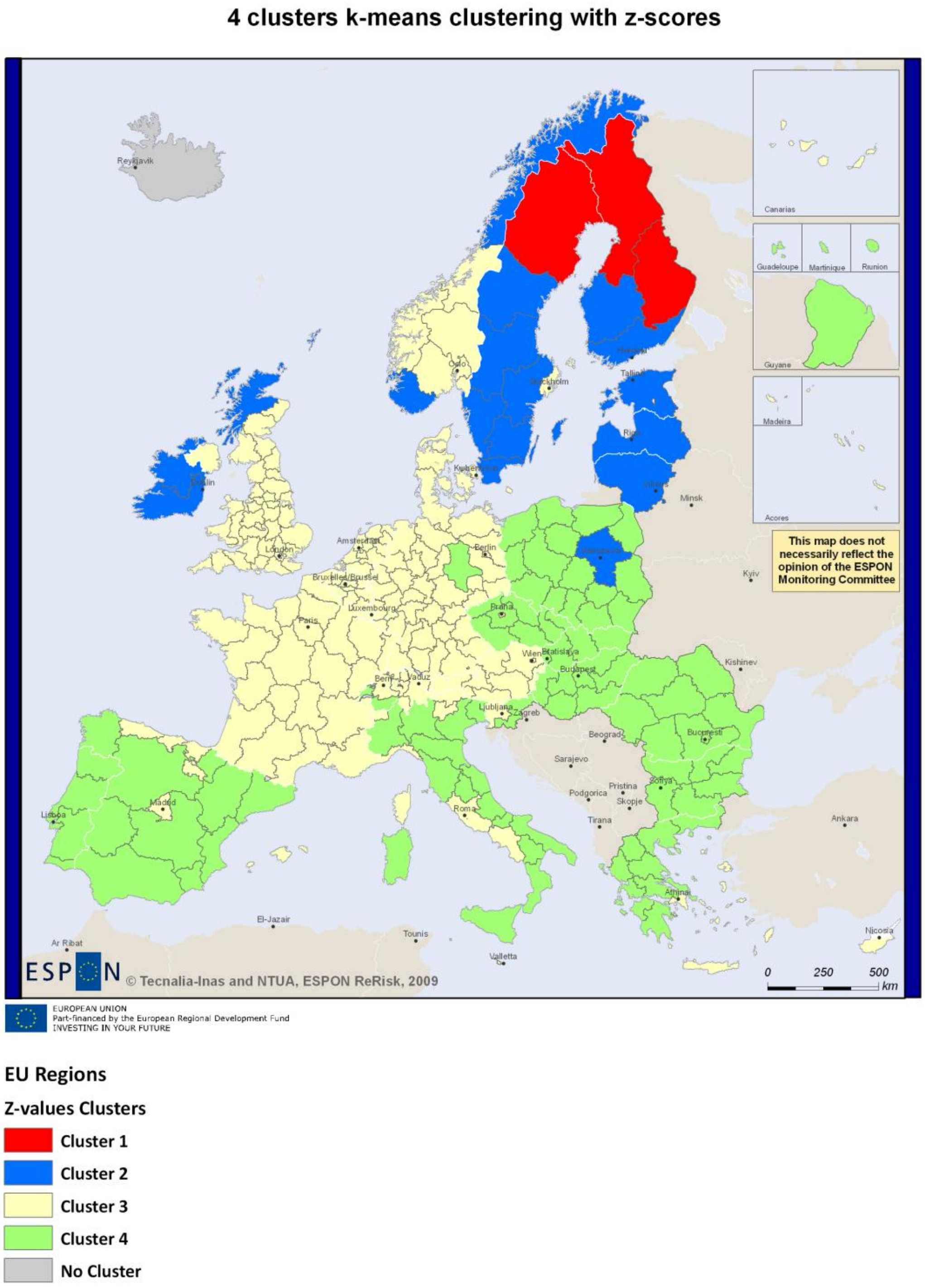
Figure 3.21 Cluster centres spider graph: Production potential or renewables

Description of the characteristics of each cluster

This time regions have been classified based on the z-scores of the indicators. In this case the variance of each variable is 1. This ensures a more balanced influence of each variable in the convergence of the clustering algorithm. Based on the cluster membership, the cluster centres, i.e. the mean value for each indicator in each cluster, of the original data have been computed. The following description of clusters is based on the latter.

Cluster 1 consists of three regions in the Nordic counties with very low values in mean January temperature and long term unemployment and very high wind power potential. This time there are more variables that characterise this cluster. The proportion of workers commuting is in lower levels than EU average and the proportion of employment in industries with high energy purchases rather higher. Cluster 2 has several similarities with Cluster 1 except for mean January temperatures (higher, similar to the mean) and wind energy potential (lower, but still high compared to the mean). Most of the Cluster 2 regions are located in the Nordic countries, the Baltic Sea countries, Ireland and Scotland. The latter is in line with the similarities of clusters centres of the two clusters.

Geographically, Cluster 3 appears to represent the central EU regions and statistically the EU regions average in terms of most of the 9 indicators except for transport dependency variables and social dimension variables. Clearly (Figure 3.19) these are regions of household with higher than average disposable incomes and very high levels of commuting. Cluster 4 differentiates significantly from Cluster 3 in most indicators, especially, economic structure (higher), transport dependency (lower commuting and double freight costs), social dimension (higher long term unemployment rate and half household disposable income) and to a lesser extent climate (higher mean July temperatures). The regions classified in Cluster 2 are mainly located in South and East Europe and have higher PV potential than Europe's average region.



Source: Own elaboration based on Eurostat data

Map 3.3 K-means clustering membership of EU regions (NUTS II): 4 Clusters, Z values

3.3.4 A 4 clusters k-means of the normalized and weighted values of the original indicators

The same k-means clustering procedure has been also repeated here. The number of clusters is 4 but the instead of the original variables, their q-scores (the sum based normalised values documented in Section 3.2) where included in the analysis. Furthermore, each variable was weighted with the weighted presented in Table 8. The weighted was based on the expert's opinions about the appropriateness of each indicator and their ranking of the indicator's importance in terms of the policy implications of the results of this project (Workshop III). Since there are two variables in each category but Economic structure, the original weight for % employment in industries with high energy purchases was doubled from 2.50 to 5.00.

Table 8 Indicator's weights

Indicator	Weight
Climate conditions	
Mean maximum temperature July	1.86
Mean minimum temperature January	2.00
Economic structure	
% employment in industries with high energy purchases	5.00
Transport dependency	
Fuel costs of freight transport	2.43
% workers commuting	2.21
Social dimension	
Long-term unemployment rate	2.64
Disposable income in households	2.36
Production potential of renewables	
Wind power potential	1.86
PV potential	2.14

The cluster centres of the original variables based on the resulted membership are presented in Table 9 along with the number of regions that were assigned to each cluster. Figures 3.22 – 3.3.28 present the data of Table 5 for each group of variables by mean of spider graphs. Map 3.4 shows the membership of each EU Region in one of the four clusters in difference colour.

Table 9 Final cluster centres

	Mean	*Cluster Centres*			
		1	**2**	**3**	**4**
Maximum temperature July	31.62	31.70	31.12	33.43	26.73
Minimum temperature January	-8.26	-7.36	-5.99	-10.06	-17.23
% employment in industries with high energy purchases	4.42	3.44	3.69	8.59	6.38
Fuel costs of freight transport	2.53	2.31	1.74	3.89	2.57
% workers commuting	9.46	7.23	45.87	5.50	3.53
Long-term unemployment rate	39.22	40.11	37.54	43.39	21.58
Disposable income in households	13316.31	14036.55	15752.46	8595.01	11321.29
Wind power potential	142525.07	114226.80	81414.17	55296.27	809093.41
PV potential	979.24	982.25	902.82	1045.55	815.14
Number of Cases		191	27	52	17

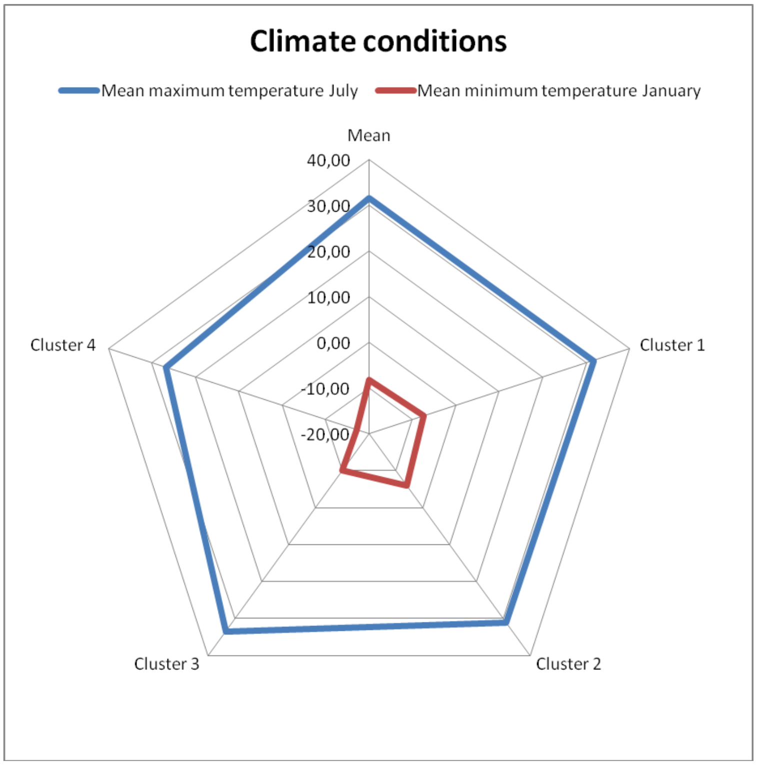


Figure 3.22 Cluster centres spider graph: Climate Conditions

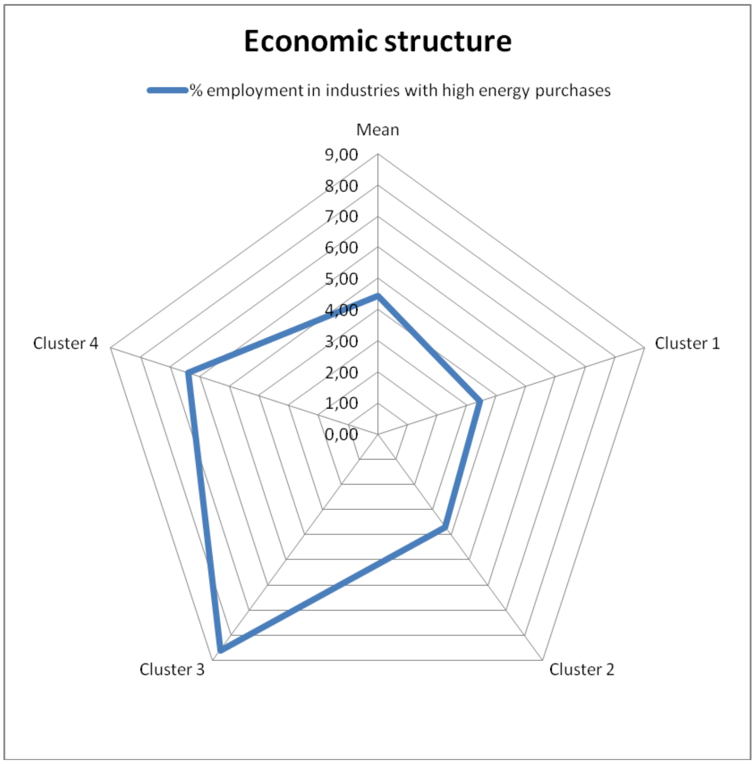


Figure 3.23 Cluster centres spider graph: Economic Structure

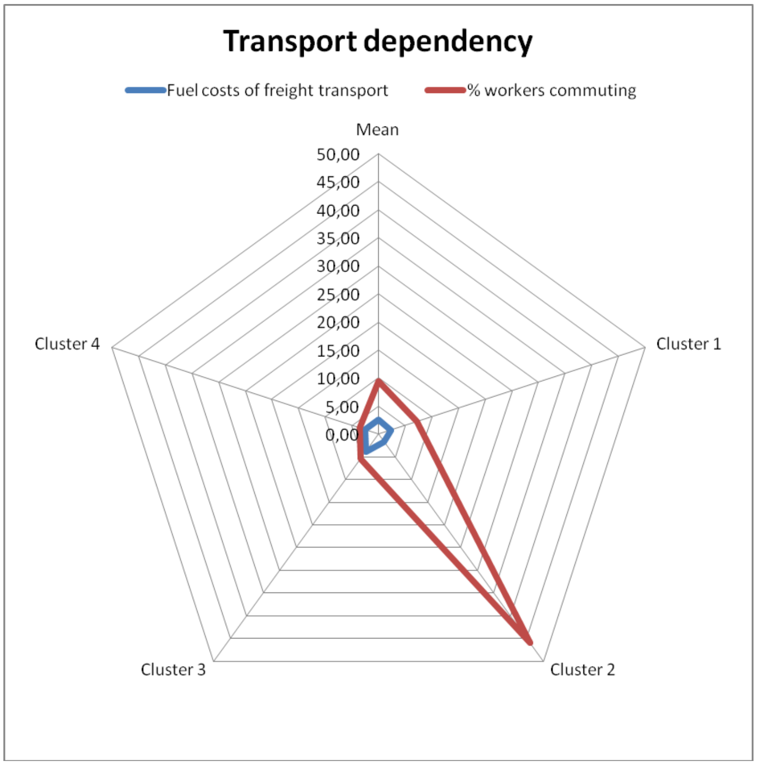


Figure 3.24 Cluster centres spider graph: Transport dependency

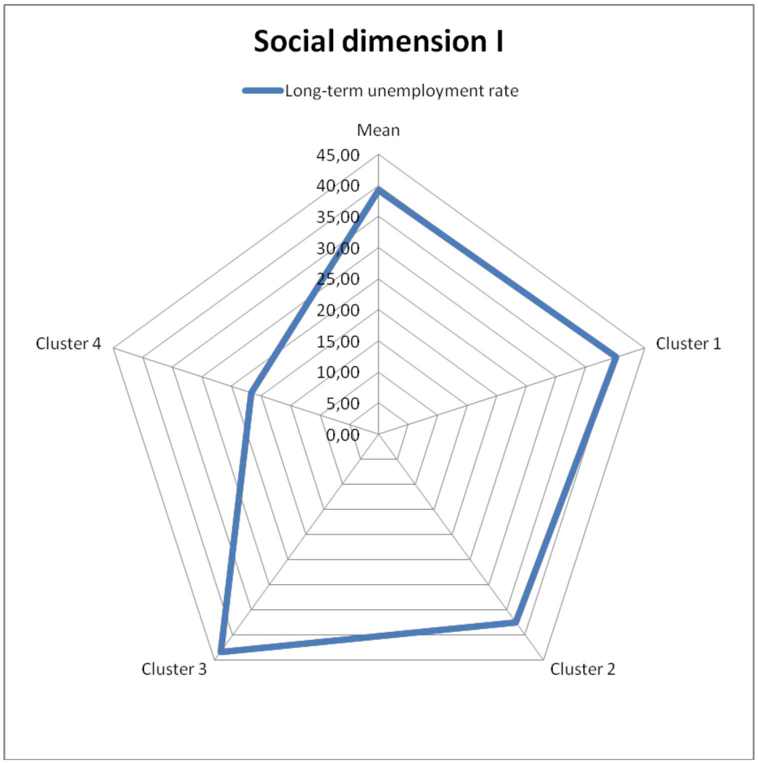


Figure 3.25 Cluster centres spider graph: Social Dimension

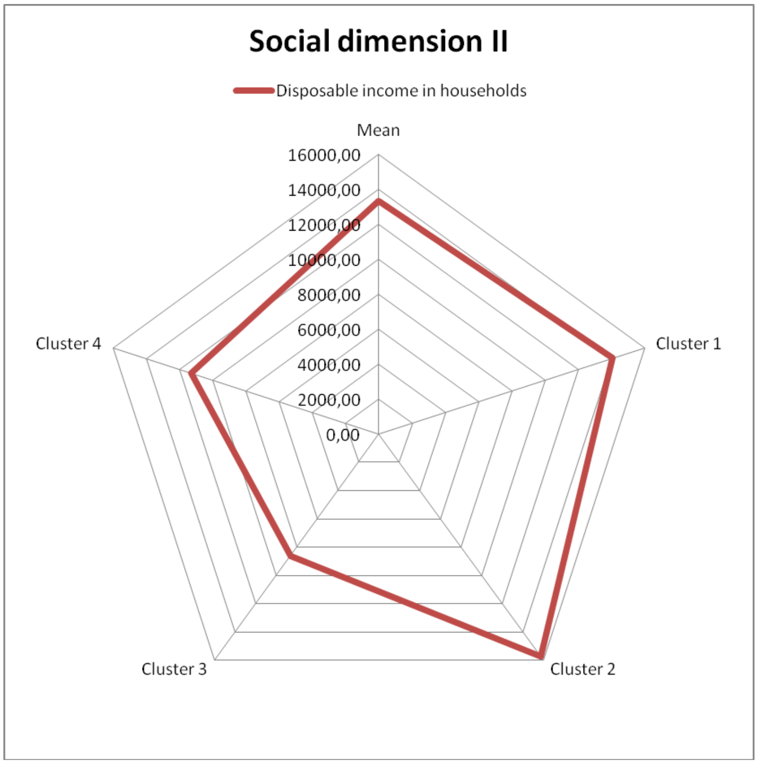


Figure 3.26 Cluster centres spider graph: Social Dimension

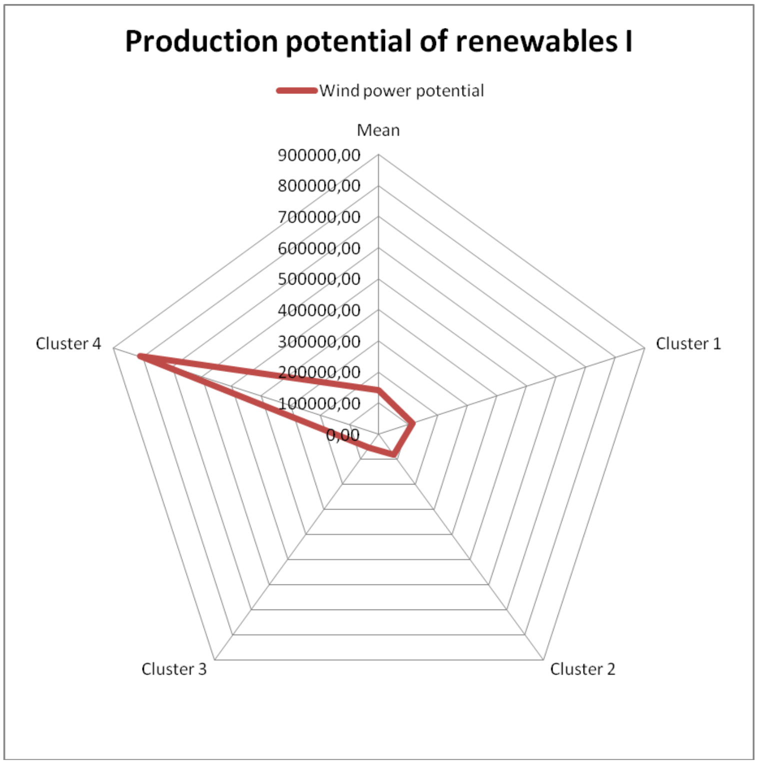


Figure 3.27 Cluster centres spider graph: Production potential or renewables

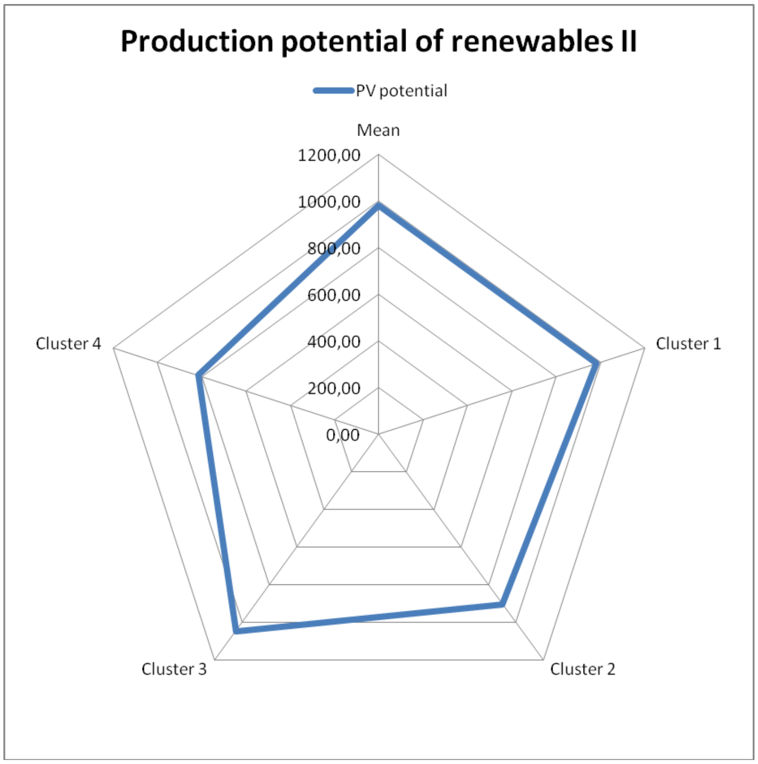


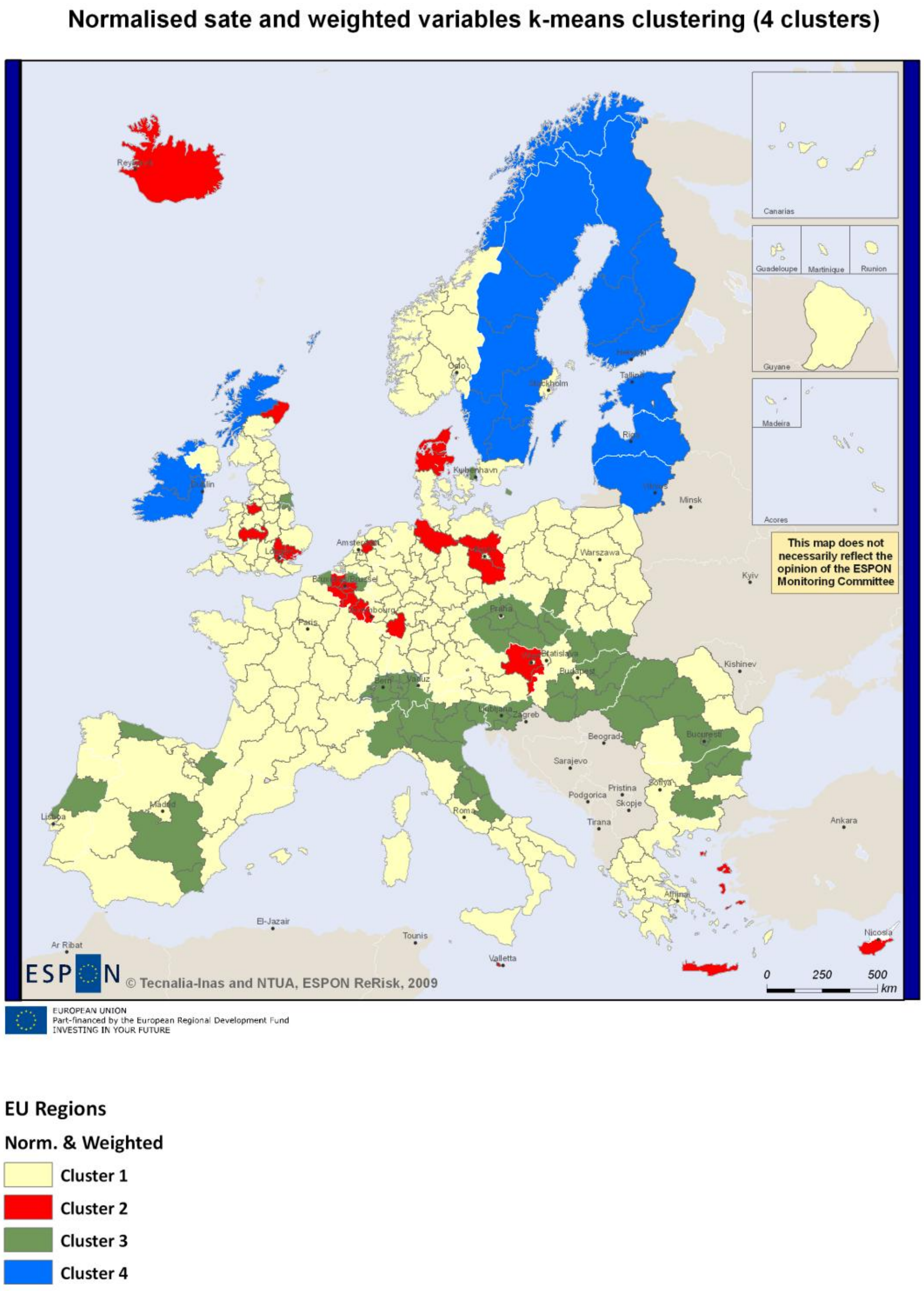
Figure 3.28 Cluster centres spider graph: Production potential or renewables

Description of the characteristics of each cluster

This time regions have been classified based on the weighted q-scores of the indicators (normalised based on the indicators sum). In this case the sum of each variable equals its weight with the higher being the one for the % employment in industries with high energy purchases. It could be expected that this variable with have higher effect on the convergence of the clustering algorithm. Again, based on the cluster membership, the cluster centres, i.e. the mean value for each indicator in each cluster, of the original data have been computed. The following description of clusters is based on the latter.

The results are interesting. Clusters 1 & 2 in the previous clustering exercise are now classified in a single cluster; this is Cluster 4. This cluster consists of 17 regions in the Nordic counties, the Baltic Sea countries, Ireland and Scotland. The differences of these regions to the EU average include lower mean January temperatures, much lower % commuting, lower social dimension, lower PV potential and extremely high wind power potential. Two in three regions are classified in Cluster 1 and as such Cluster 1 appears to represent the EU average region. Indeed the cluster centres much most of the 9 indicators mean values except for small differences in economic structure and transport dependency variables. Cluster 2 has several similarities with Cluster 1 except for transport dependency and social dimension (lower, but still high compared to the mean) indicators. Members of this cluster exhibit high household disposable income and extremely high levels of commuting (50%). Geographically speaking, Cluster 2 regions are scattered across Europe.

The regions classified in Cluster 3 are mainly located in South and East Europe. Regions in this cluster are characterised by low income, high long-term unemployment rate, high PV potential and high % employment in industries with high energy purchases.



Source: Own elaboration based on Eurostat data

Map 3.4 K-means clustering membership of EU regions (NUTS II): 4 Clusters, Normalised and weighted values

4. Conclusions

In this report the classification procedure of the EU Regions into groups of regions with similar characteristics has been presented. Area classification (typologies) is a helpful tool in the study of the risk of energy poverty in Europe. This classification has resulted in regional typologies that should assist policy makers in to understand the picture of Europe in various aspects and form their policy agenda accordingly.

It is necessary to note that although the classification is sensitive to the input data, there are distinct groups of regions that are very dissimilar to each other. Sometimes this dissimilarity is based on only a few indicators such as wind power potential (Nordic Regions), commuting (Central European Regions) and social dimension (South and East European Regions).

Typologies suggested in the beginning of this report were meant to be such that allow the assessment of risk for of energy poverty as well as the spatial differentials of this risk. This would help policy makers to make informed decisions. One could perhaps assign Type 4, this is Lagging Regions (regions with current or eminent energy poverty), to Cluster 3 of the last clustering (Section 3.3.3) or Cluster 4 in the previous (Section 3.3.4). However, we felt that the policy makers and other readers of this report should not be bound to the stereotype typologies but understand the performance of each region based on different criteria.

Obviously this type of research is an on-going process. Future attempts should need more indicators and a finer geographical scale, e.g. NUTS III. For a geodemographic-like area classification that seems to be gaining a lot of credit in the literature in answering deprivation related research questions, detailed data on the demographic and socio-economic structure of the population and business is required.

5. Acknowledgements

It is important to thank everybody who helped with his/her valuable comments in the preparation of this report, especially Edurne Magro. It is also necessary to acknowledge our use of the software GeoDa version 0.9.5-i. GeoDa was developed by Luc Anselin at the Spatial Analysis Laboratory (SAL) of the Department of Geography at the University of Illinois, Urbana-Champaign (https://www.geoda.uiuc.edu).

6. References

- Anselin, L. (2003) An Introduction to EDA with GeoDa, Spatial Analysis Laboratory (SAL), Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.
- Anselin, L. (2004) GeoDa 0.95i Release Notes, Spatial Analysis Laboratory (SAL), Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.
- Batagelj, V., Bock, H.-H., Ferligoj, A., Žiberna, A. (2006) Data Science and Classification, Series: Studies in Classification, Data Analysis, and Knowledge Organization (Berlin: Springer-Verlag).
- Cliff, A.D., and Ord, J.K. (1973) Spatial autocorrelation (London: Pion).
- Cliff, A.D., and Ord, J.K., (1981) Spatial processes: models and applications (London: Pion).
- Harris, R., Sleight, P., Webber, R. (2005) Geodemographics, GIS and Neighbourhood Targeting (London: Wiley).
- Hartigan, J.A., (1975) Clustering Algorithms (New York: Wiley).
- Hartigan, J.A., and Wong, M.A. (1979) A K-means clustering algorithm, Applied Statistics, 28, pp. 100 – 108.
- Kalogirou, S. (2003) The Statistical Analysis And Modelling Of Internal Migration Flows Within England And Wales, PhD Thesis, School of Geography, Politics and Sociology, University of Newcastle upon Tyne, UK.
- Milligan, G.W., and Cooper, M.C. (1988) A Study of Standardization of Variables in Cluster Analysis, Journal of Classification, 5, pp. 181 – 204.
- Moran, P.A.P. (1948) The interpretation of statistical maps, Journal of the Royal Statistics Society, Series B (Methodological), 10, 2, 243 – 251.
- Singleton, A. and Longley, P. (2008) Creating Open Source Geodemographics - Refining a National Classification of Census Output Areas for Applications in Higher Education, Papers in Regional Science, 88 (3), pp. 643 – 666.
- Su, C., Zhan, J. and Sakurai, K. (2009), Importance of Data Standardization in Privacy-Preserving K-Means Clustering In L. Chen et al. (Eds.): DASFAA 2009 Workshops, LNCS 5667 (Berlin: Springer-Verlag), pp. 276–286.
- Vickers, D., Rees, P. (2007). Creating the National Statistics 2001 Output Area Classification, Journal of the Royal Statistical Society, Series A, 170 (2), pp. 379–403.
- Webber, R. J. and Craig, J. (1978) Socio-economic classification of local authorities. In Studies on Medical and Population Subjects No 35, (London: HMSO).

Annex I. Figures

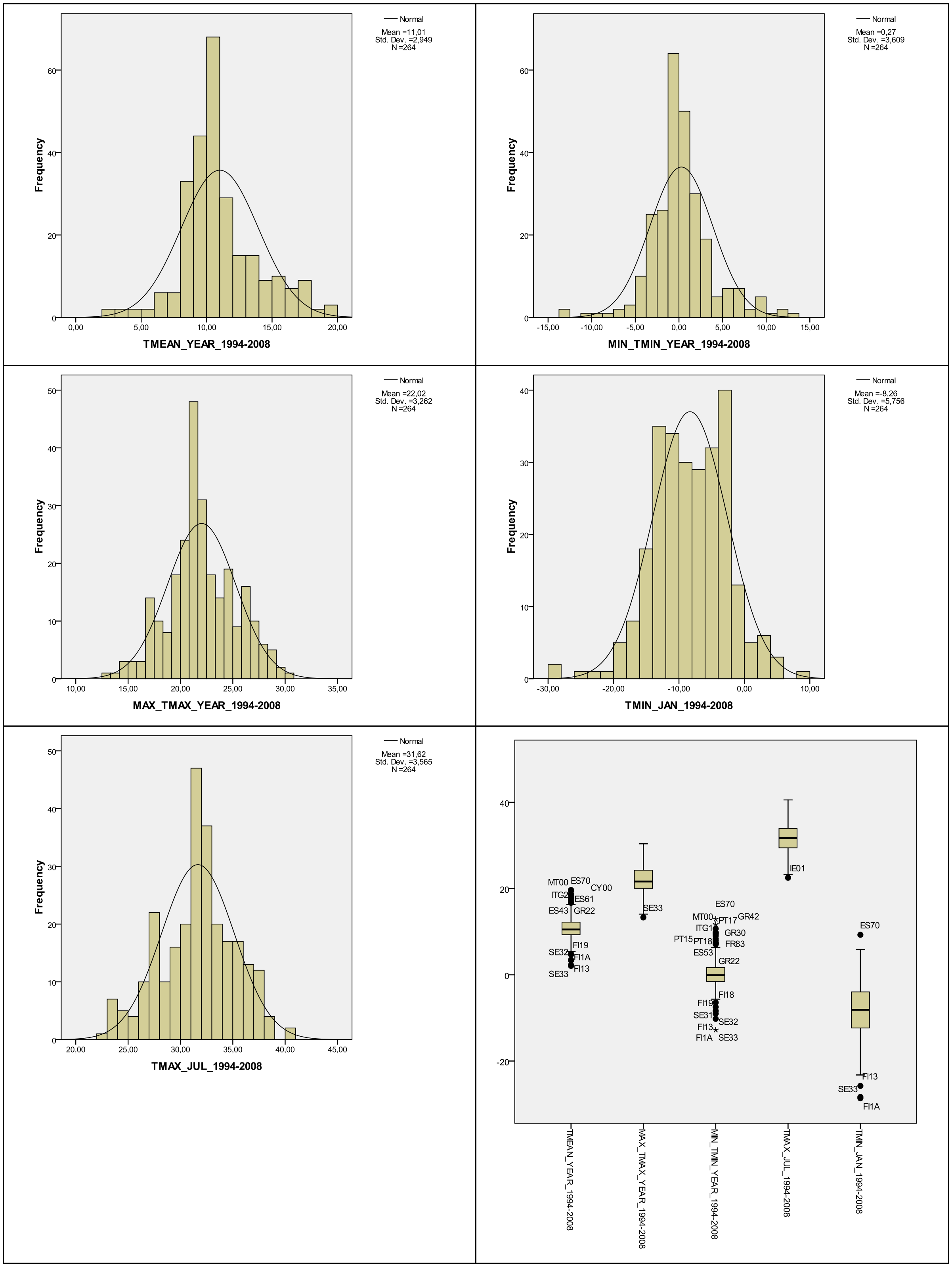


Figure 1 Frequency histogram and boxplot of the climate conditions indicators

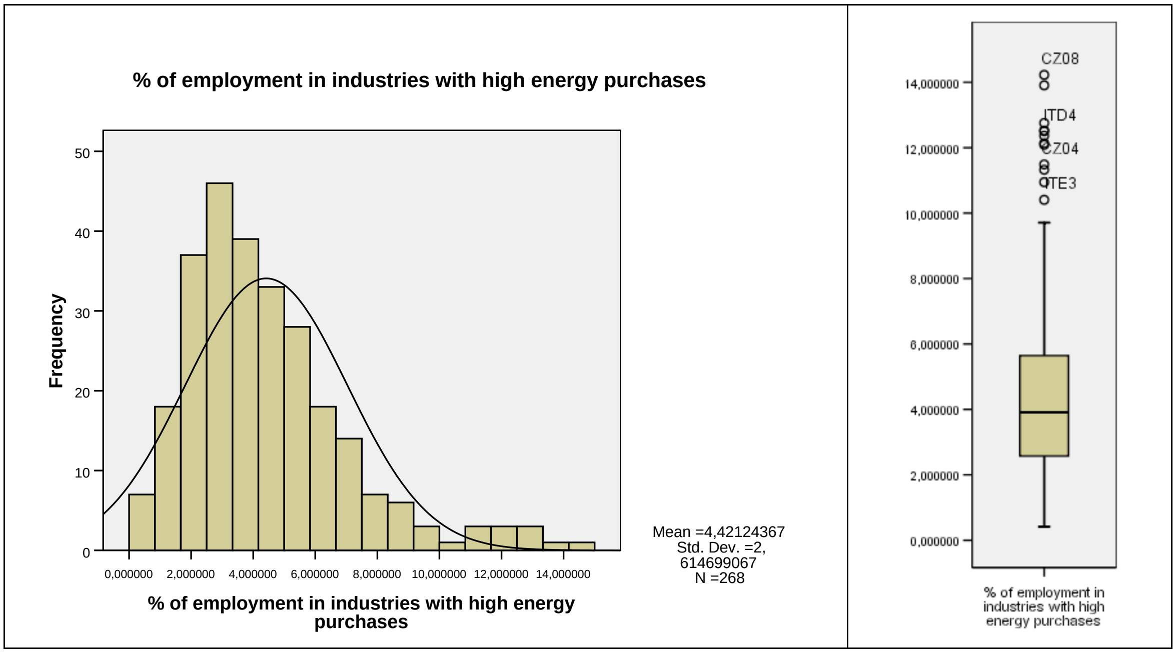


Figure 2 Frequency histogram and boxplot of the percentage of employment in industries with high energy purchases

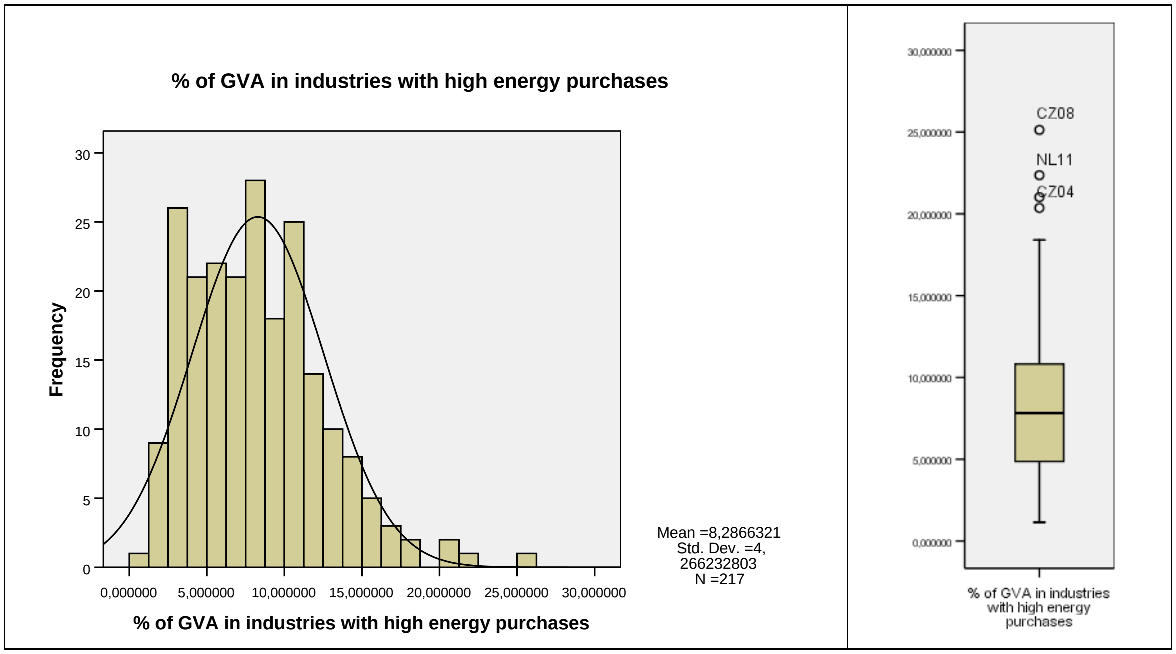


Figure 3 Frequency histogram and boxplot of the percentage of GVA in industries with high energy purchases

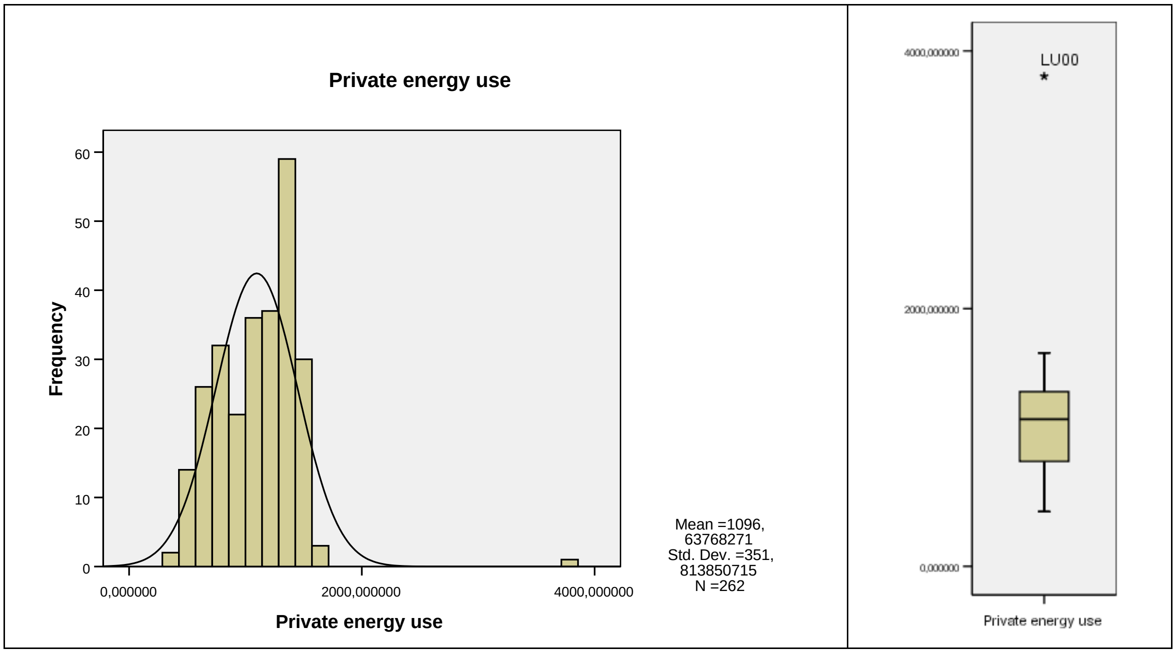


Figure 4 Frequency histogram and boxplot of the private energy use

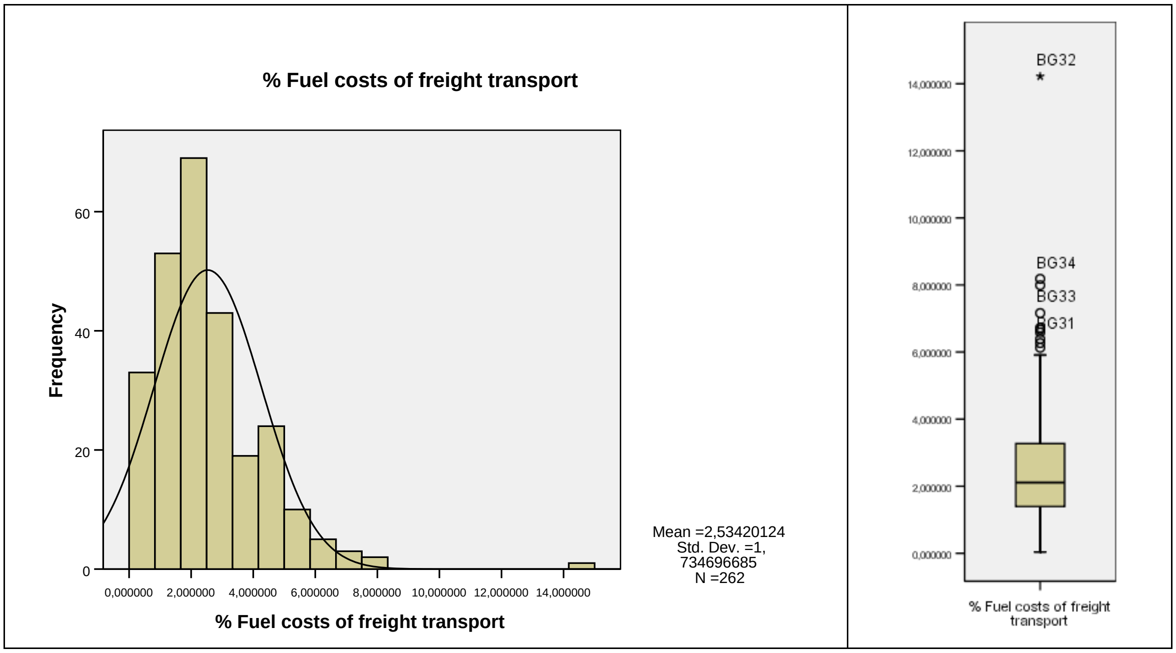


Figure 5 Frequency histogram and boxplot of fuel costs

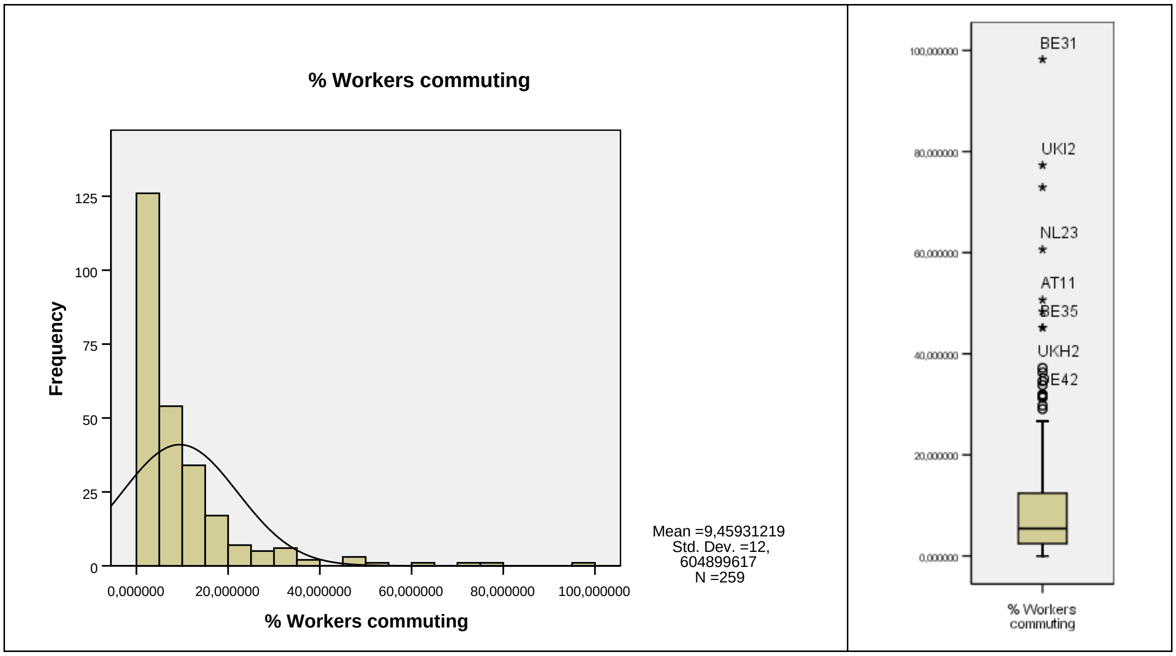


Figure 6 Frequency histogram and boxplot of the proportion of workers commuting

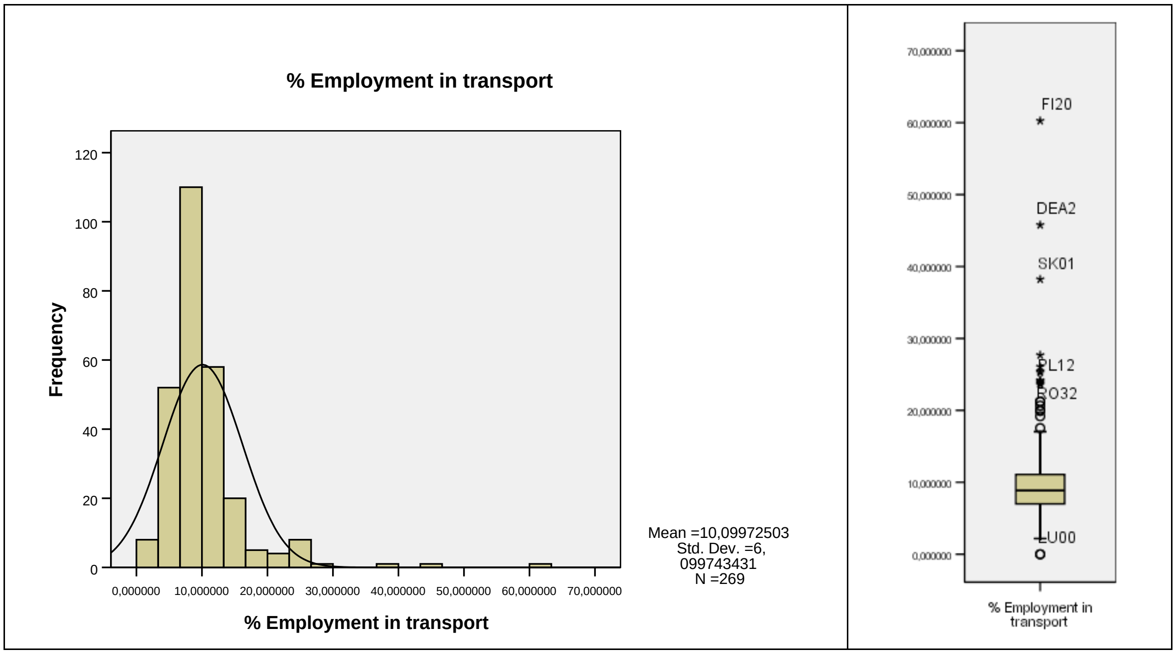


Figure 7 Frequency histogram and boxplot of the proportion of employment in transport

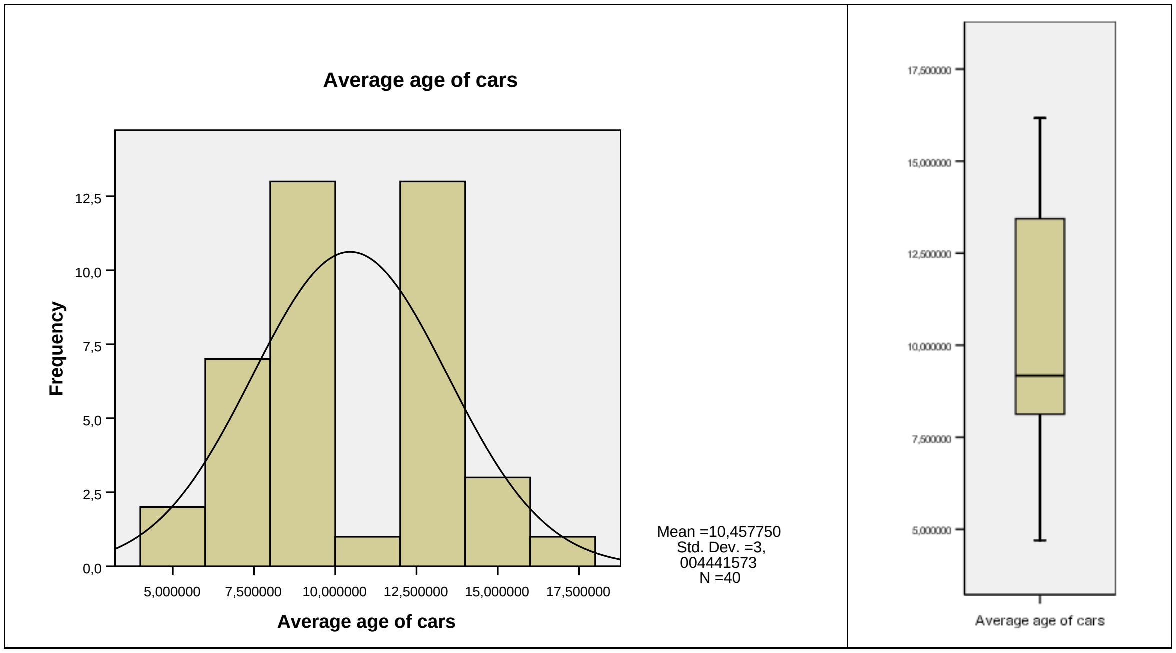


Figure 8 Frequency histogram and boxplot of average age of cars

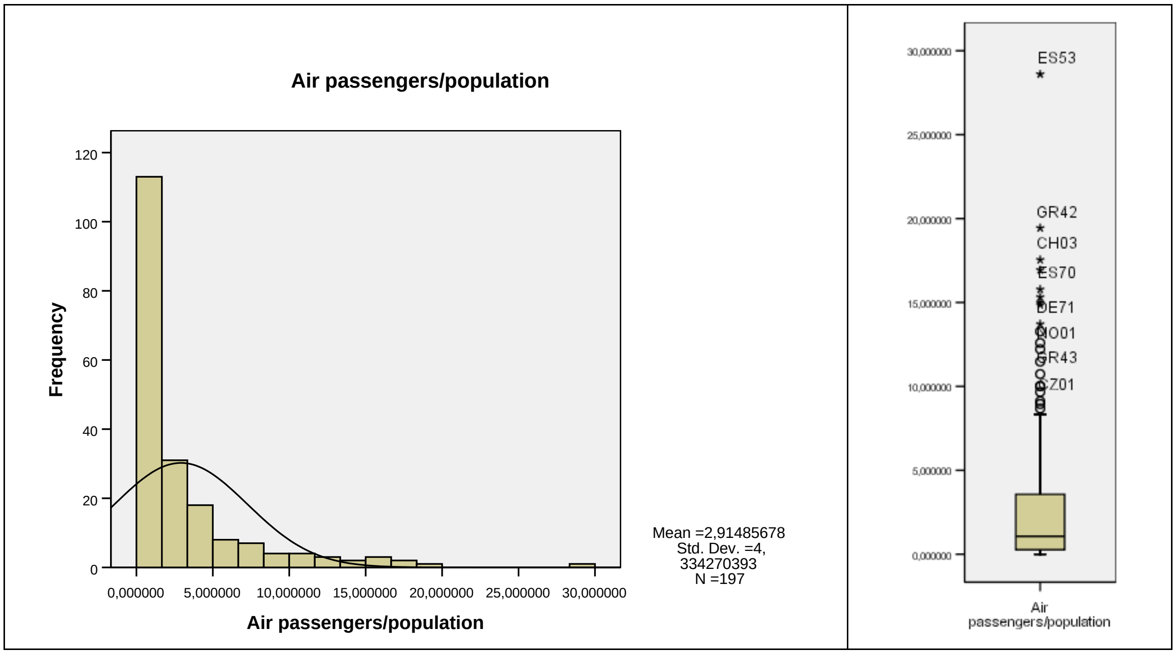


Figure 9 Frequency histogram and boxplot of the proportion of air passengers

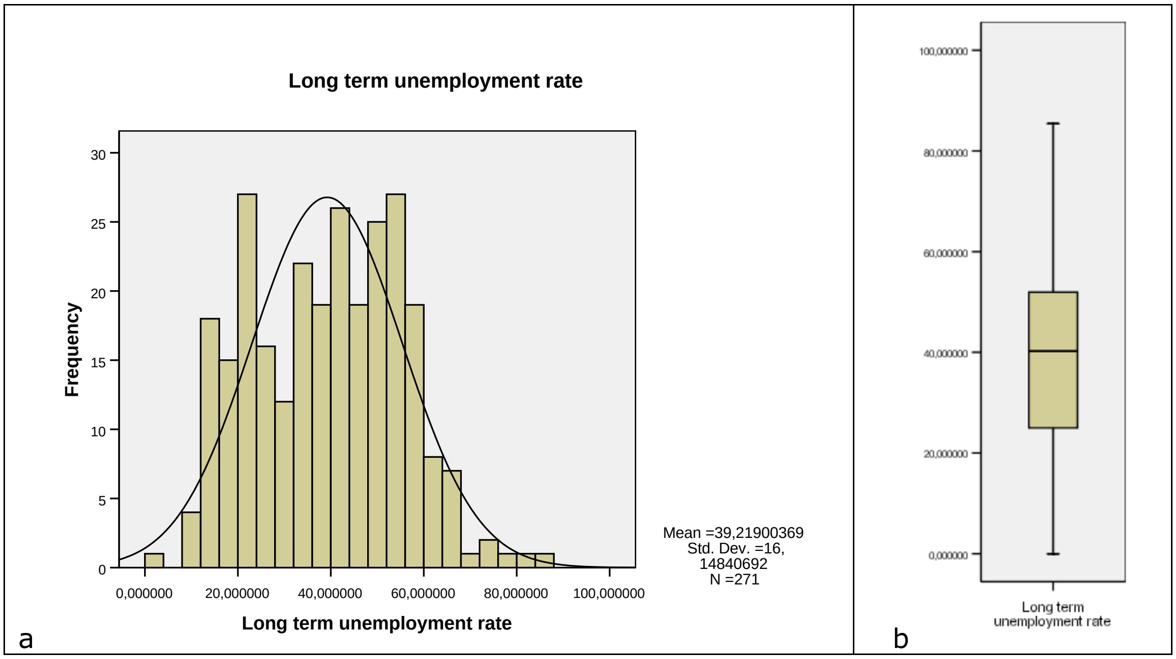


Figure 10 Frequency histogram and boxplot of the long-term unemployment rate

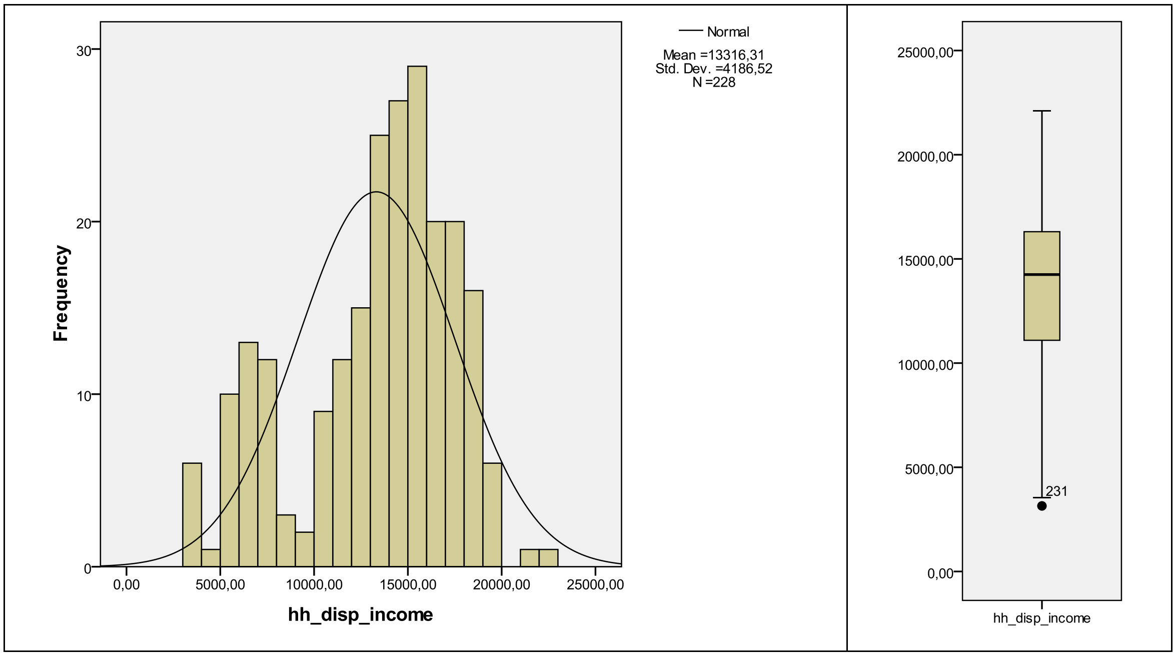


Figure 11 Frequency histogram and boxplot of the disposable income in households

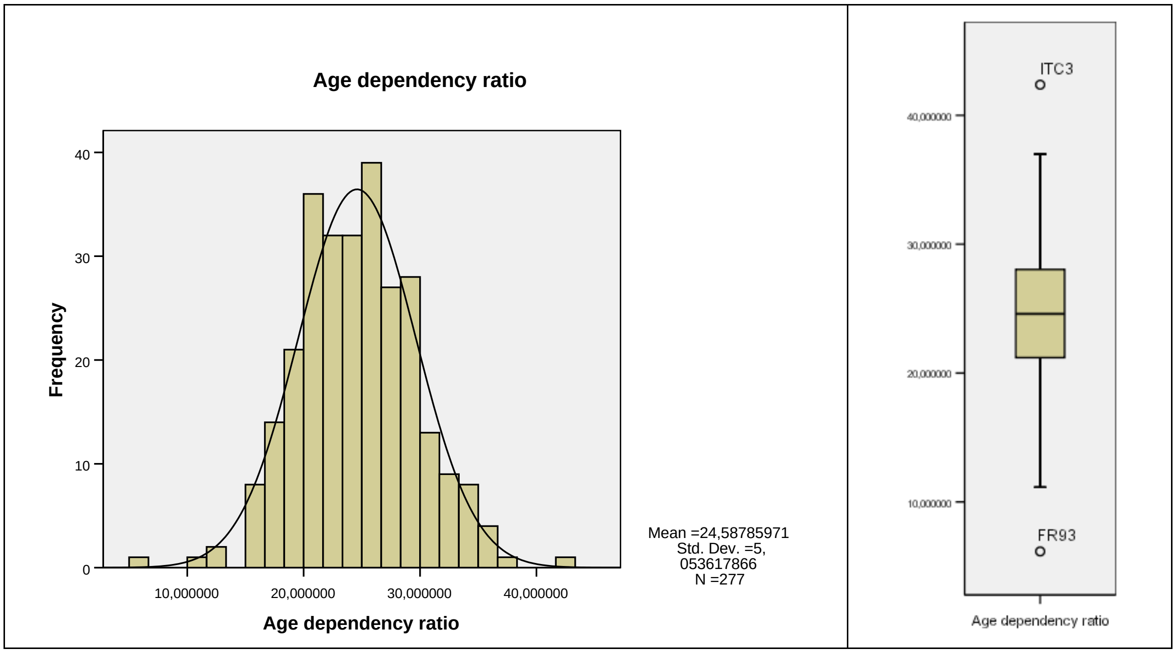


Figure 12 Frequency histogram and boxplot of age dependency ratio

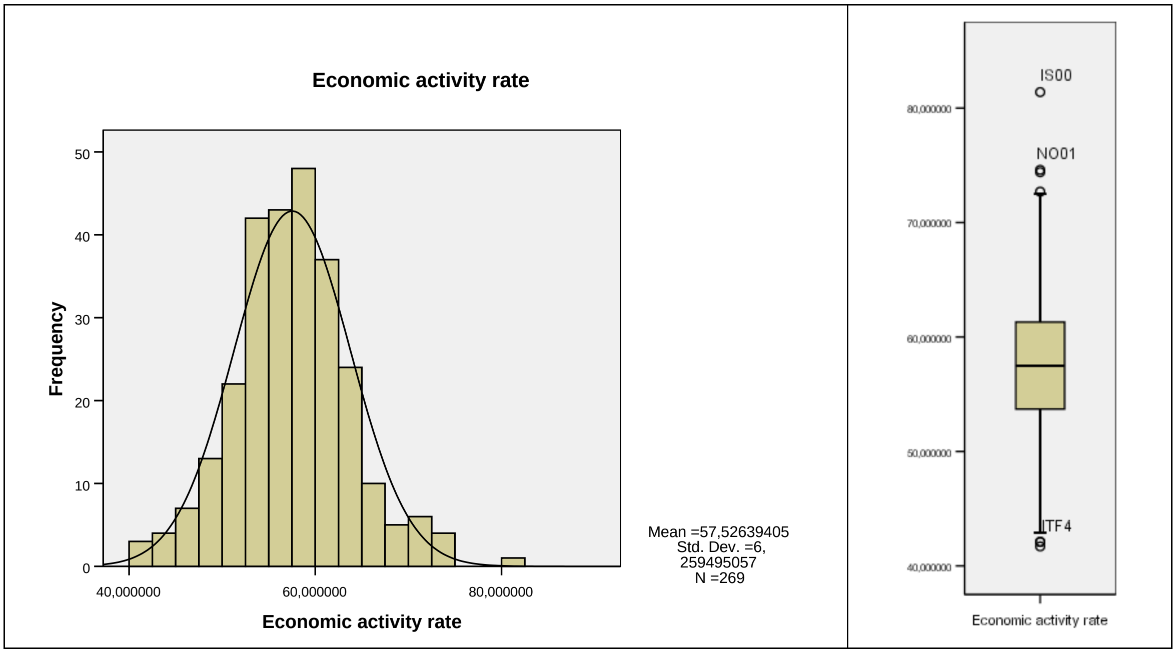


Figure 13 Frequency histogram and boxplot of economic activity rate

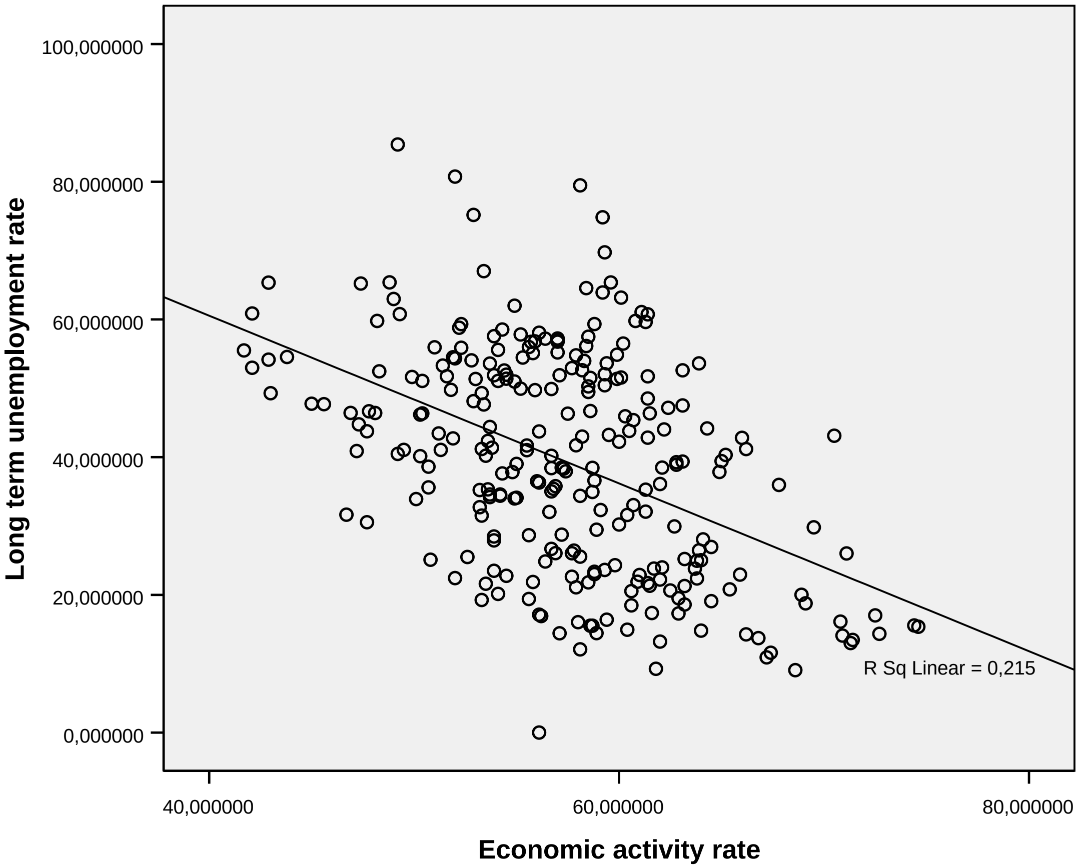


Figure 14 Scatter plot of the Economic activity rate vs. the Long term unemployment rate

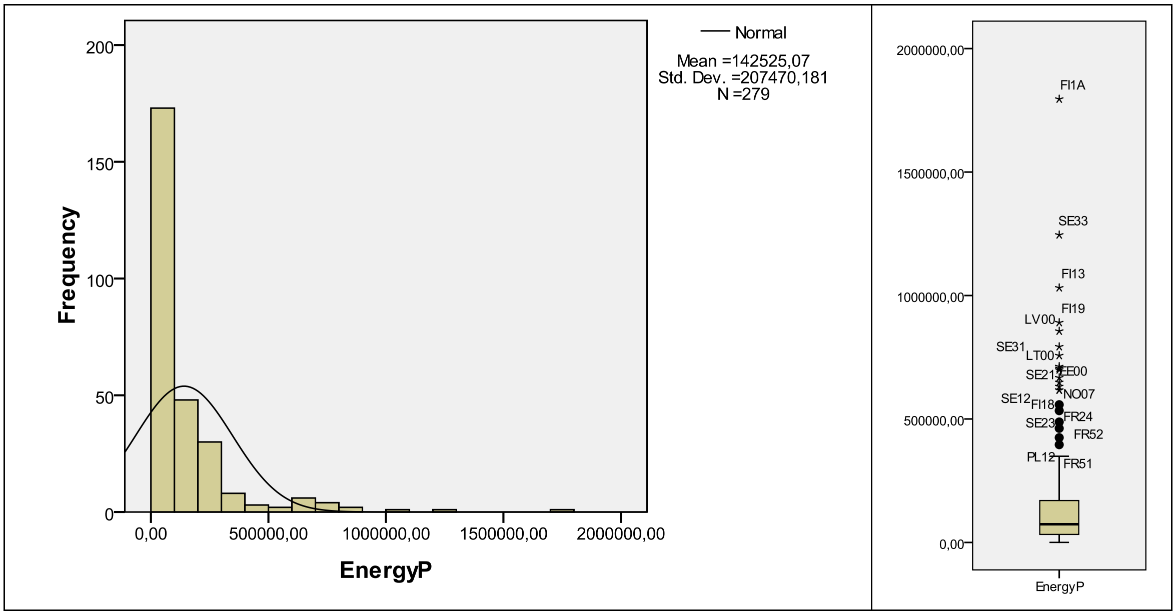


Figure 15 Frequency histogram and boxplot of the Wind Power Energy Potential

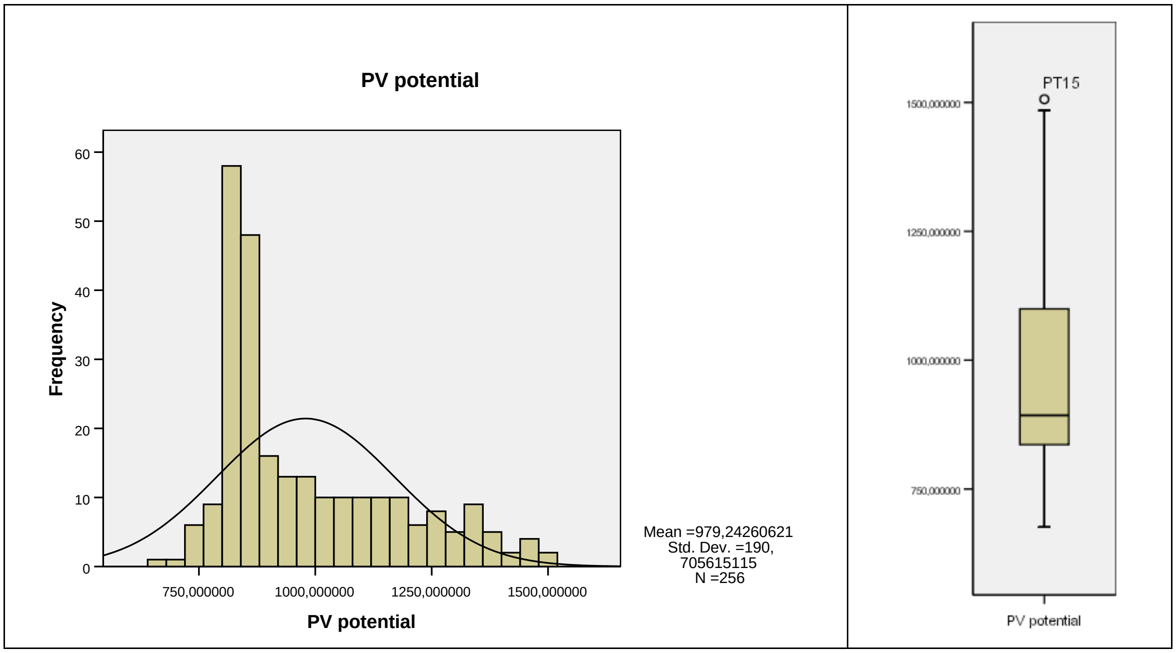


Figure 16 Frequency histogram and boxplot of PV potential

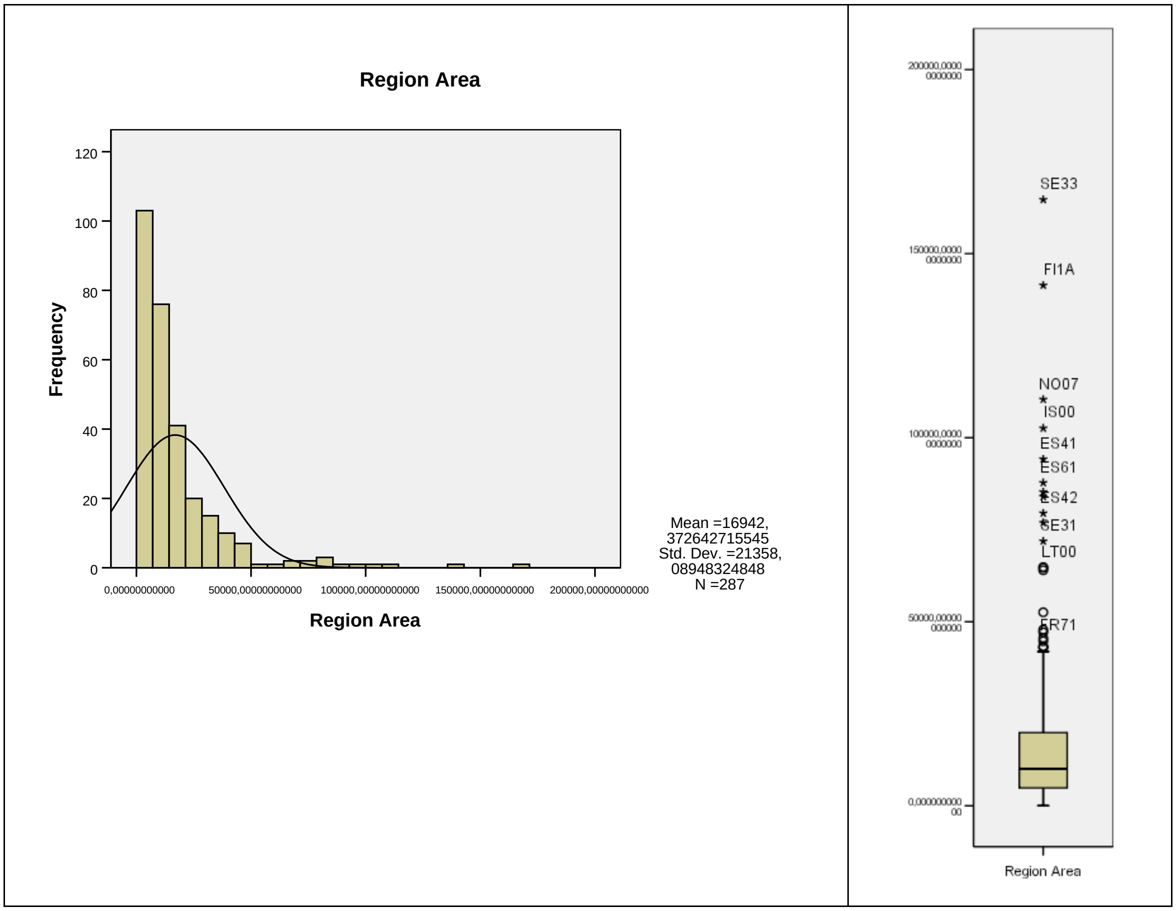
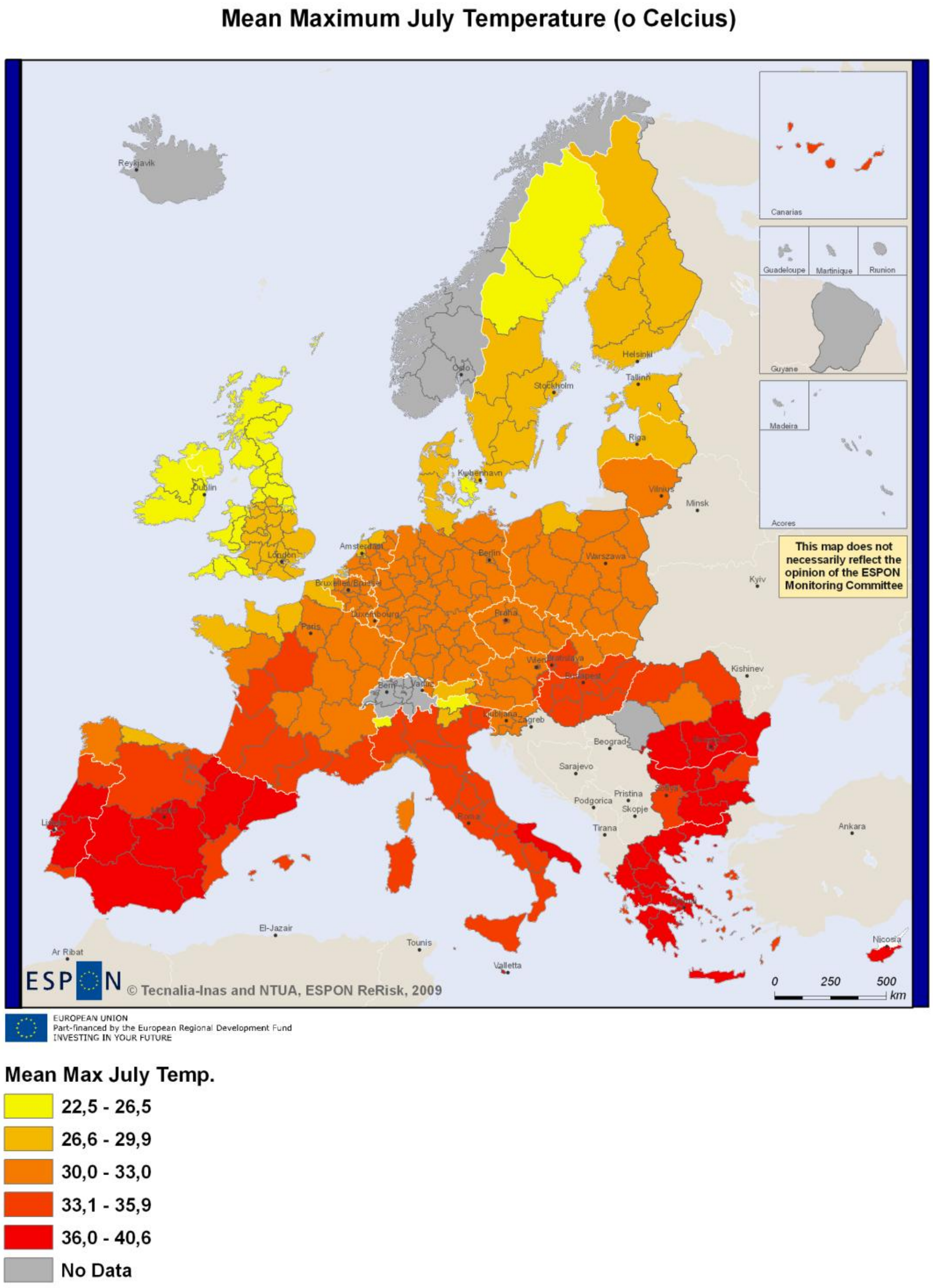


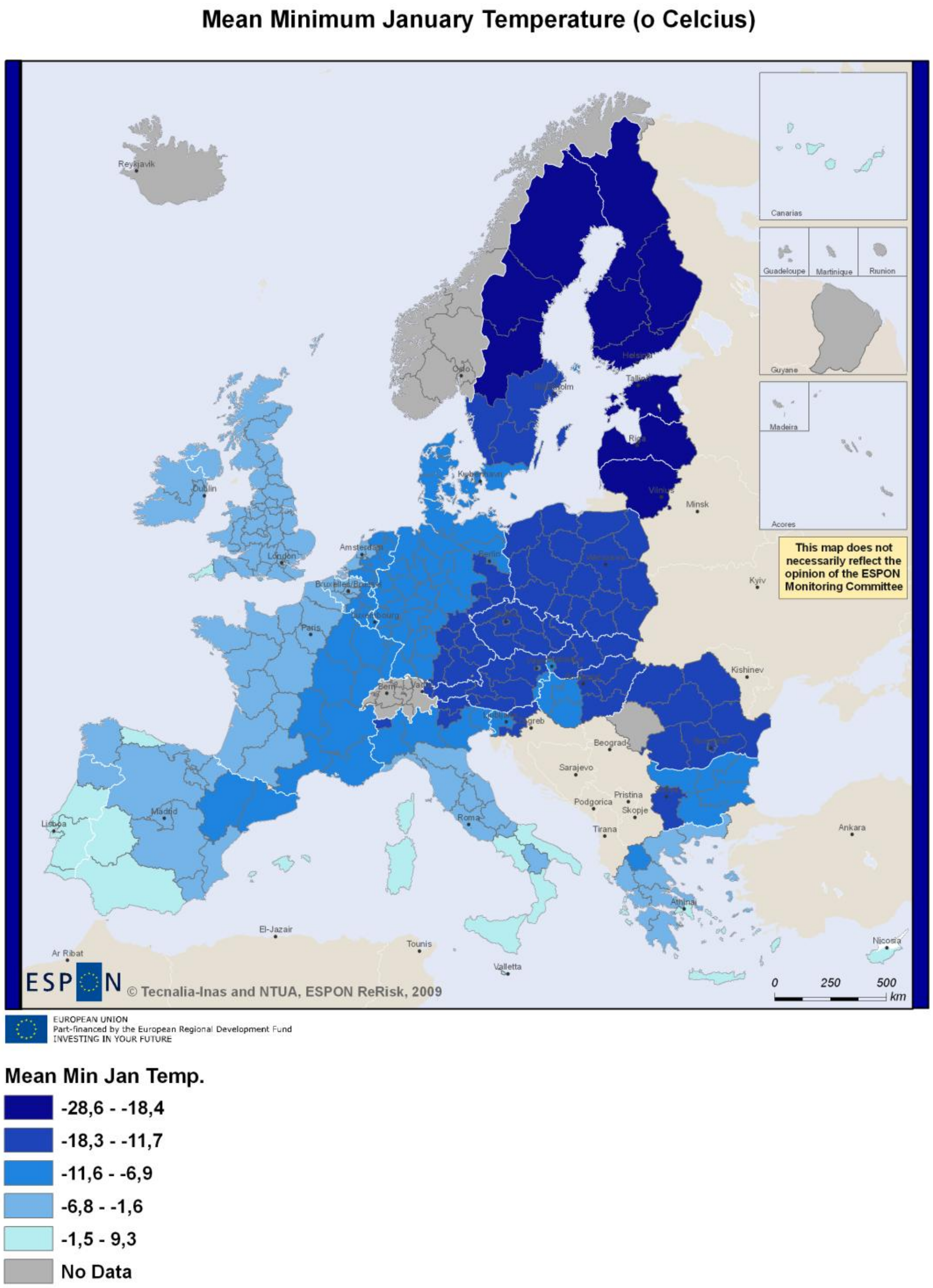
Figure 17 Frequency histogram and boxplot of the regions' area

Annex II. Maps



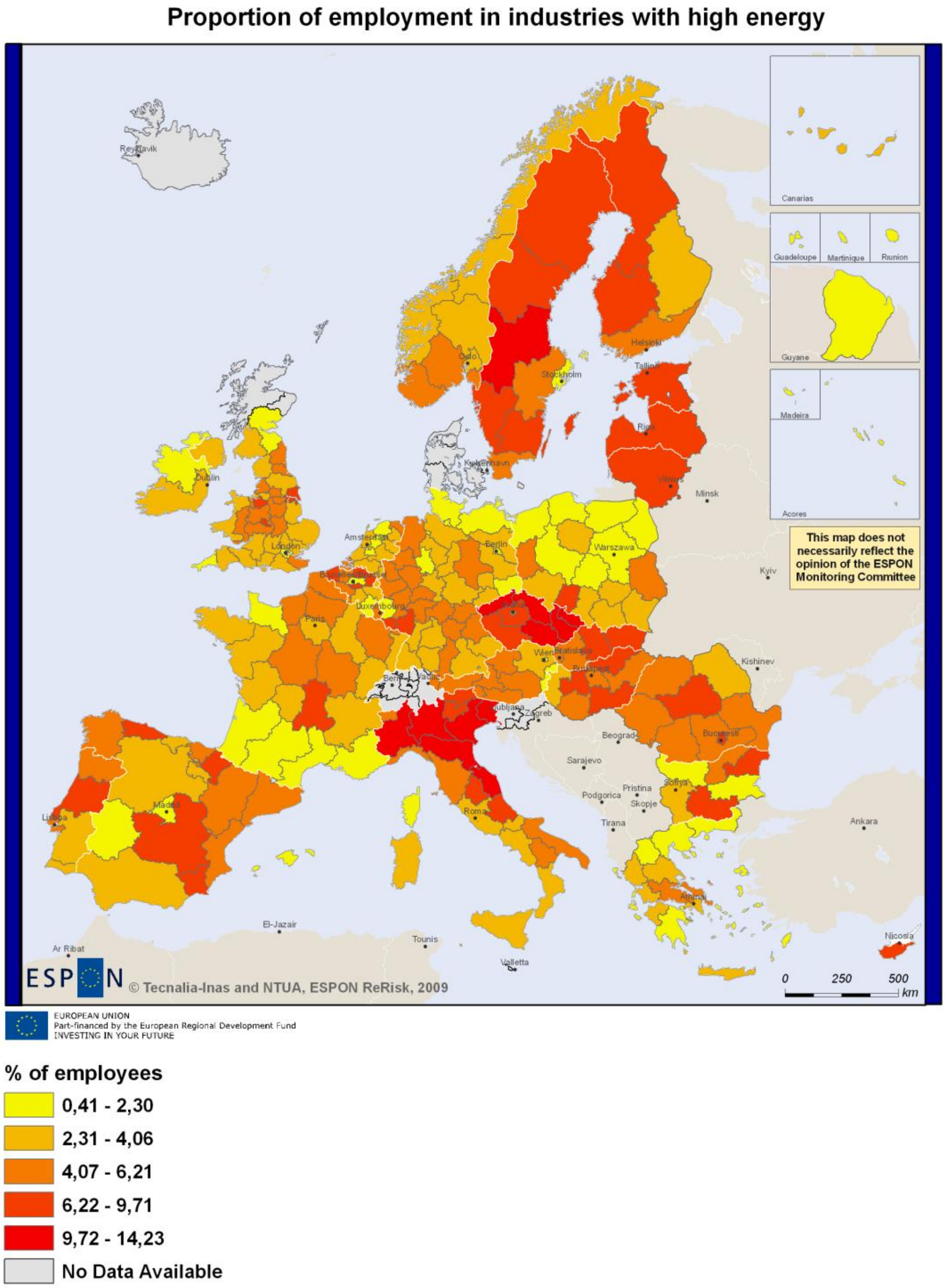
Source: Joint Research Centre, Ispra - IPSC -MARS Unit

Map 1 Mean maximum July temperature in the EU regions (NUTS II)



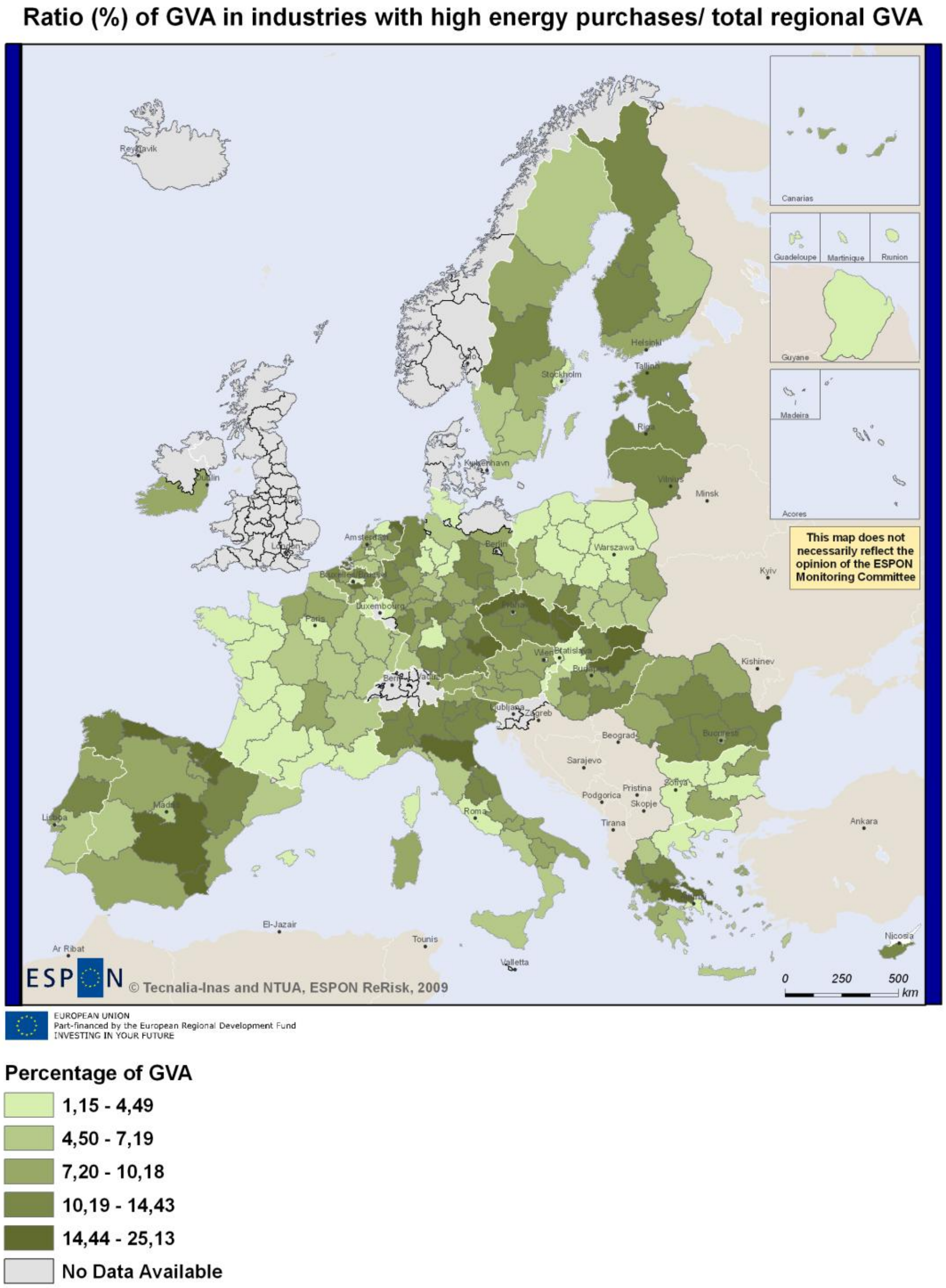
Source: Joint Research Centre, Ispra - IPSC -MARS Unit

Map 2 Mean minimum January temperature in the EU regions (NUTS II)



Source: Own elaboration based on Eurostat data

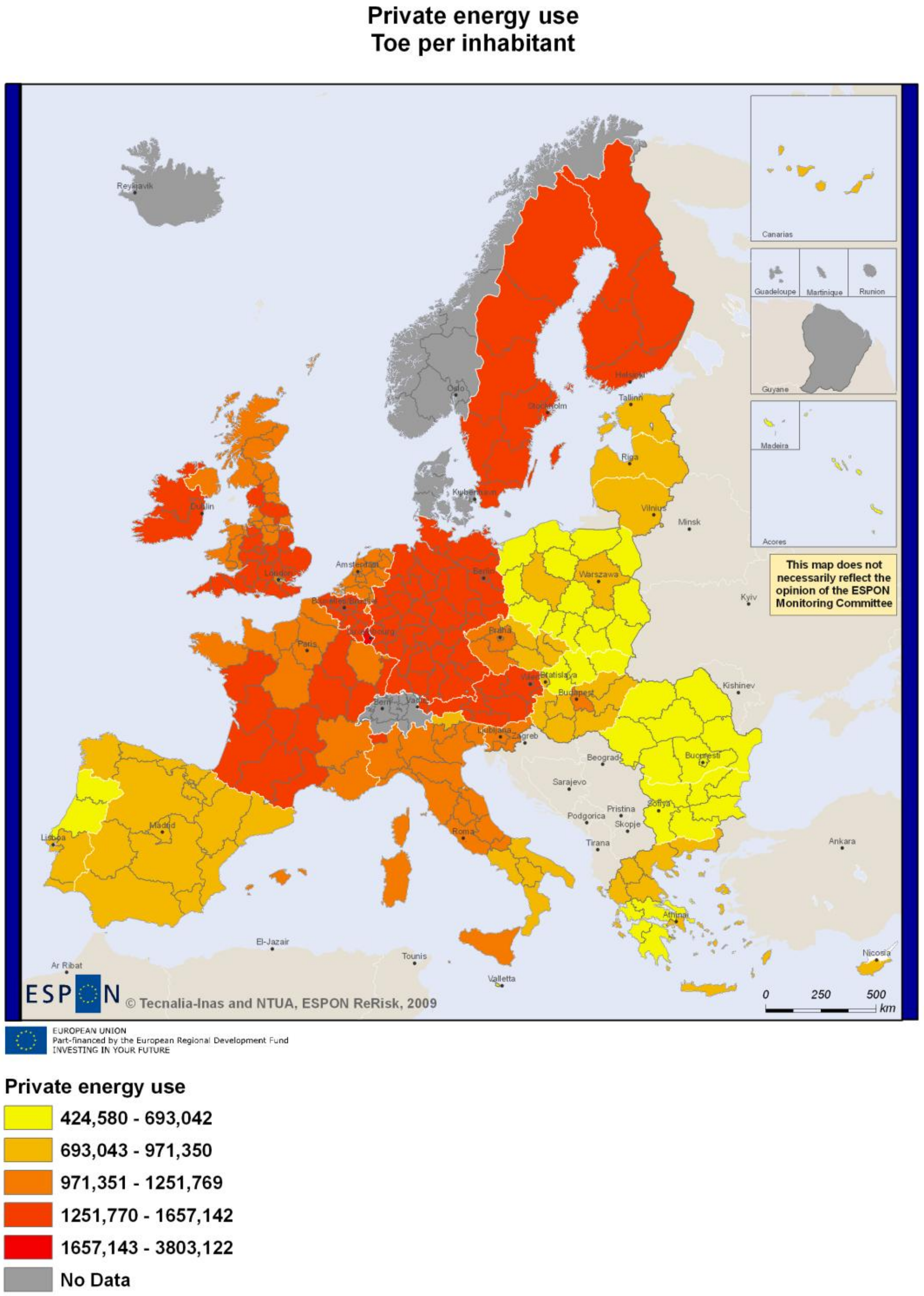
Map 3 % of employment in industries with high energy purchases in the EU regions (NUTS II)



Ratio (%) of GVA in industries with high energy purchases/ total regional GVA

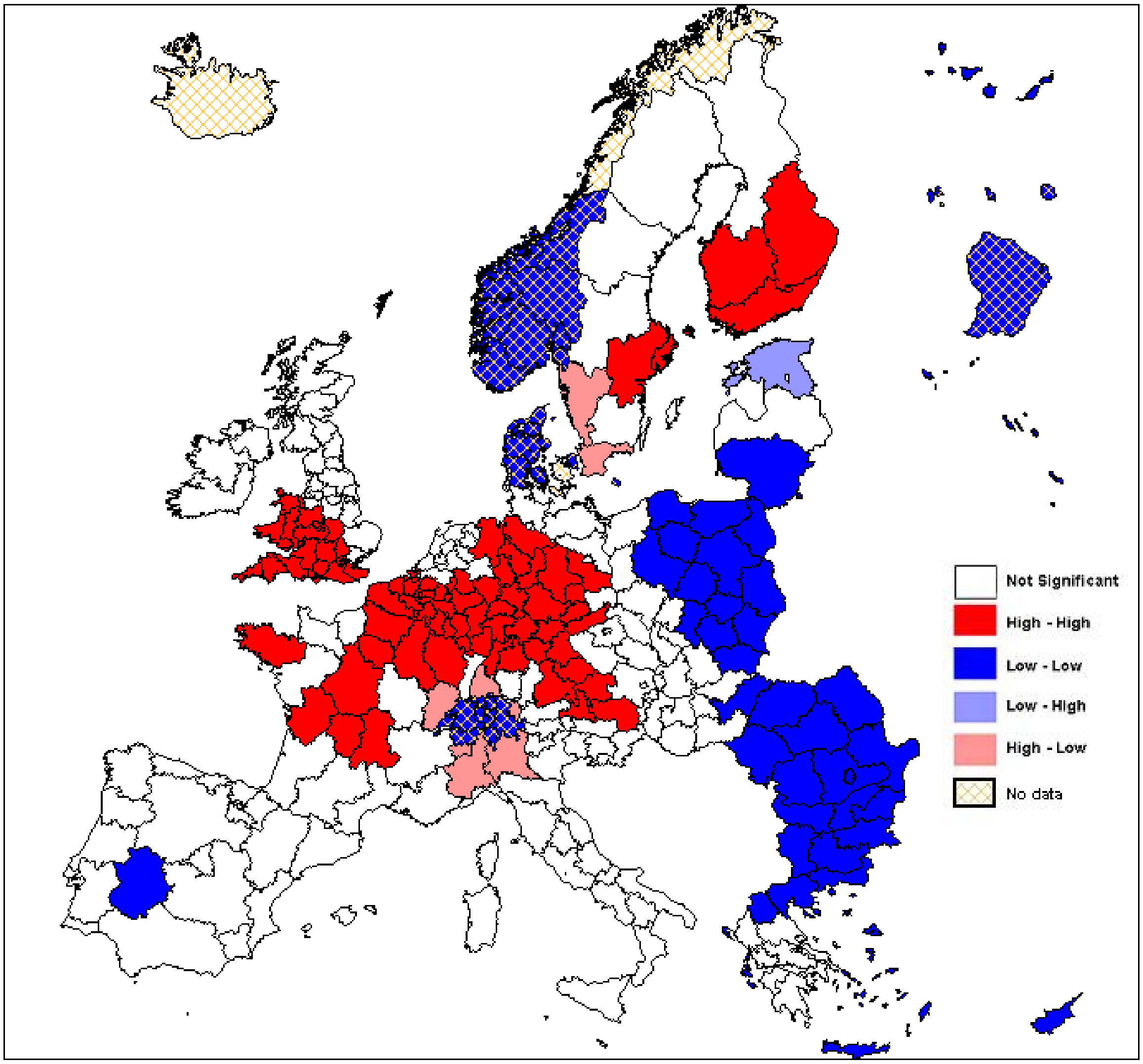
Source: Own elaboration based on Eurostat data

Map 4 Proportion of GVA in industries with high energy purchases in the EU regions (NUTS II)



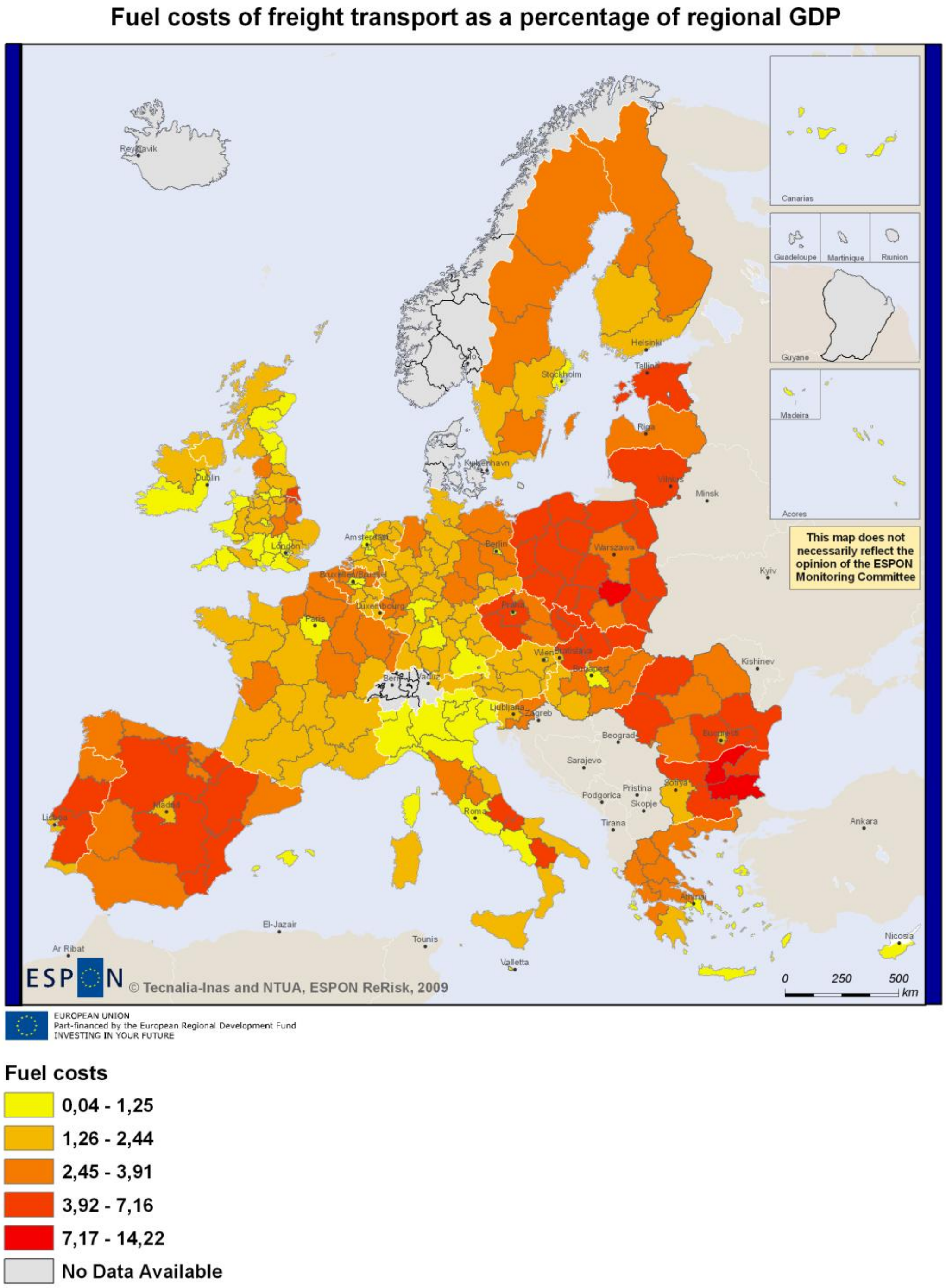
Source: Own elaboration based on Eurostat data

Map 5 Private energy use in the EU regions (NUTS II)



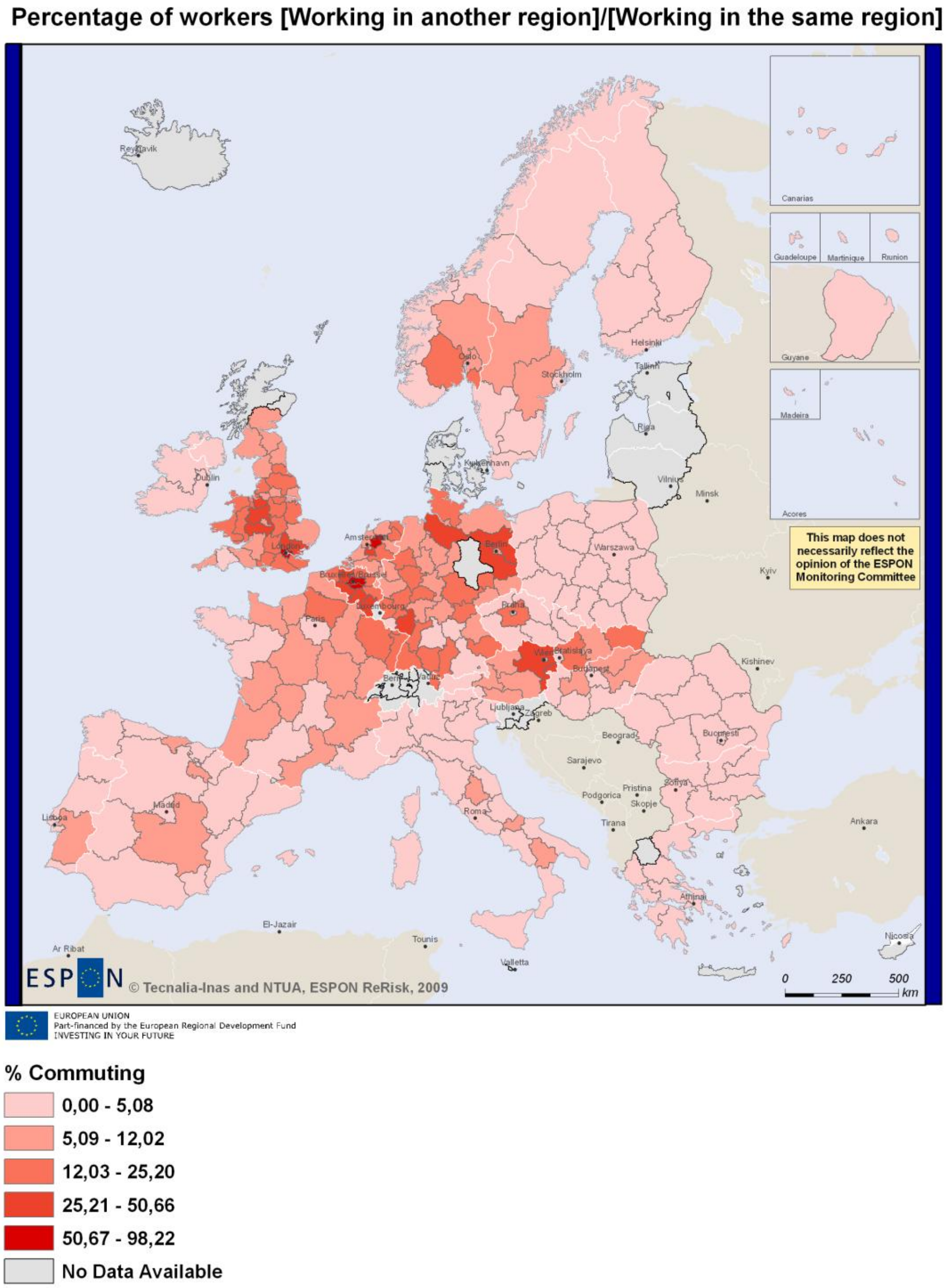
Source: Own elaboration based on Eurostat data

Map 6 LISA cluster map of the private energy use



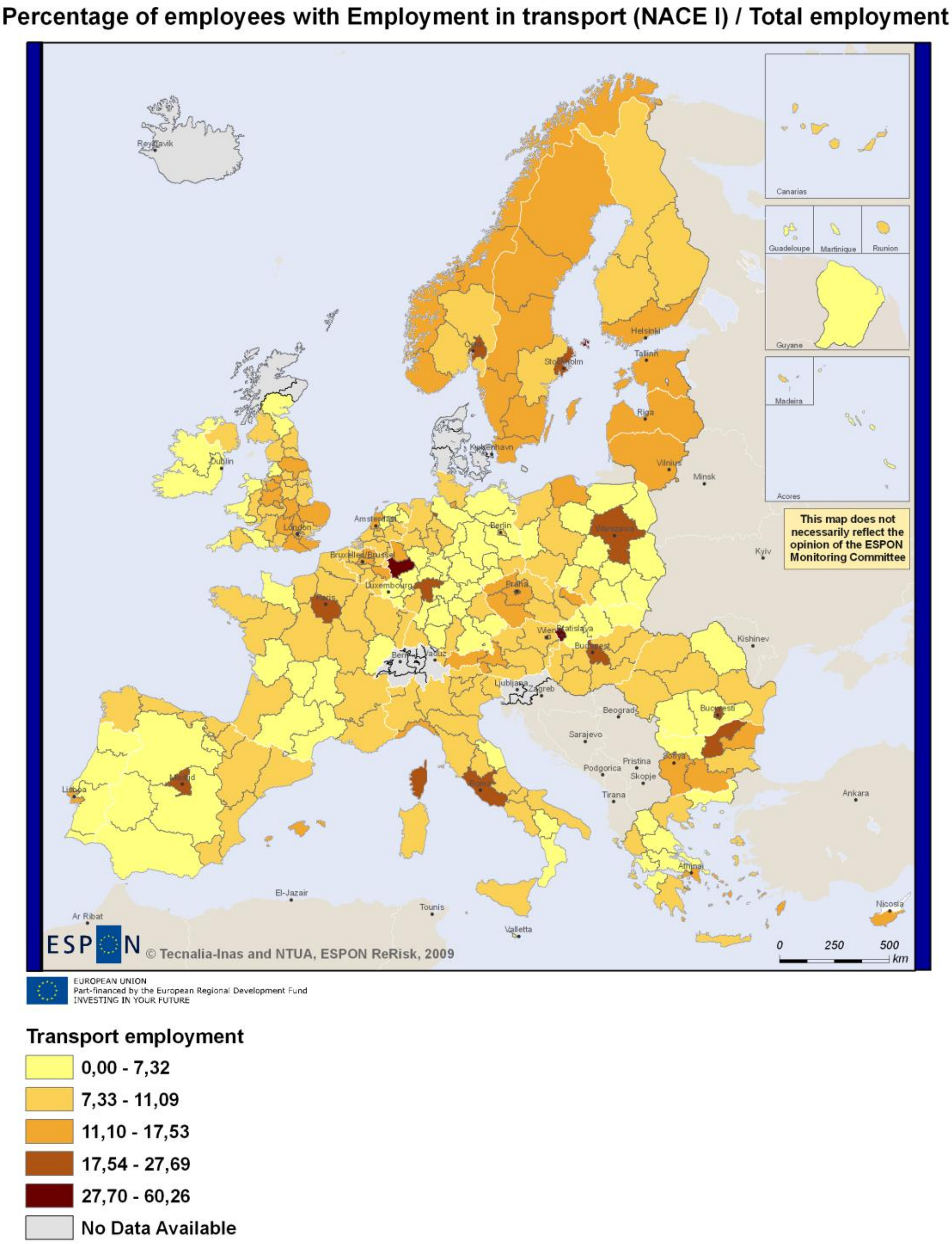
Source: DG Regio

Map 7 Fuel costs in the EU regions (NUTS II)



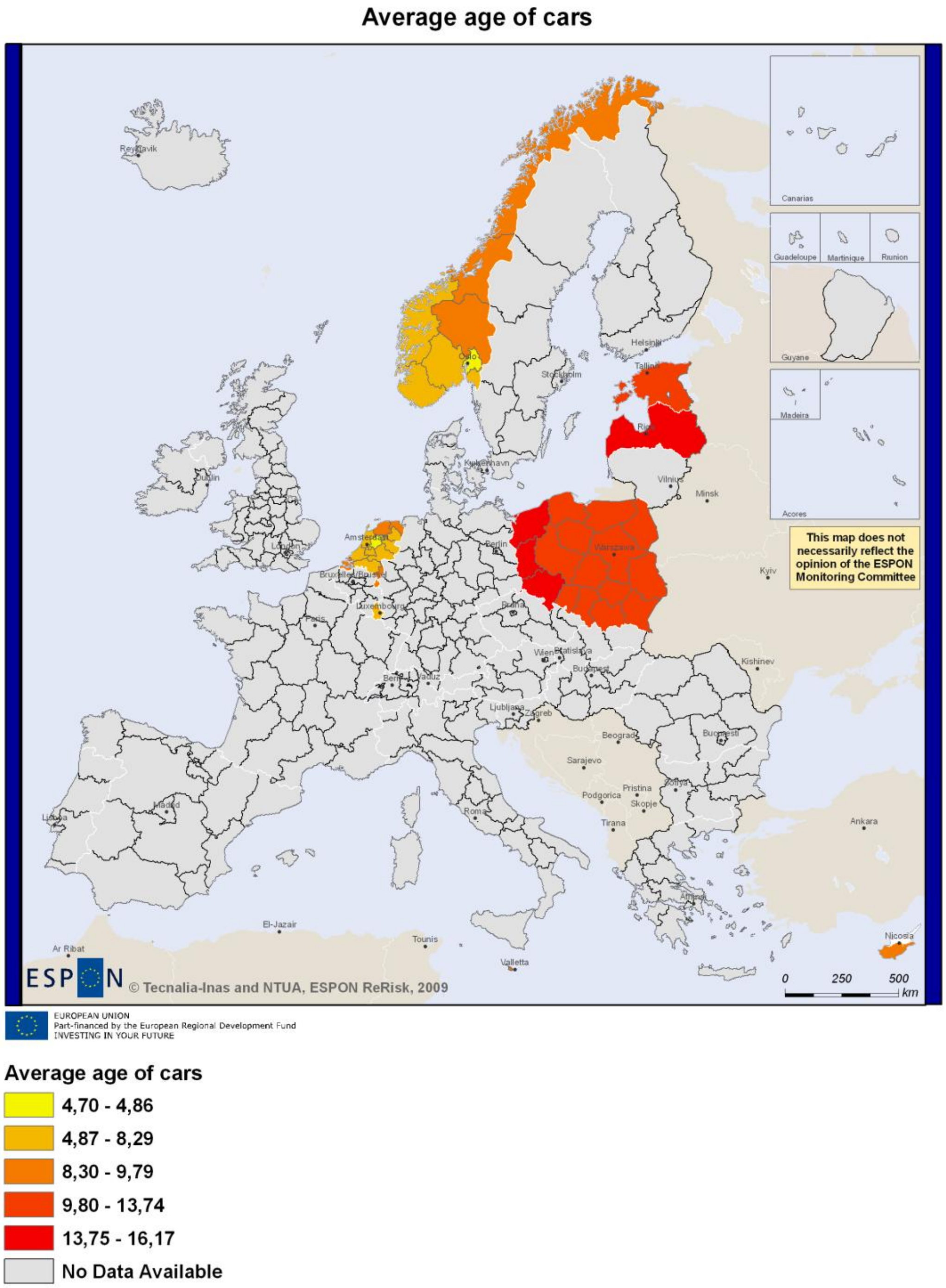
Source: Own elaboration based on Eurostat data

Map 8 Percentage of workers commuting to another region in the EU regions (NUTS II)



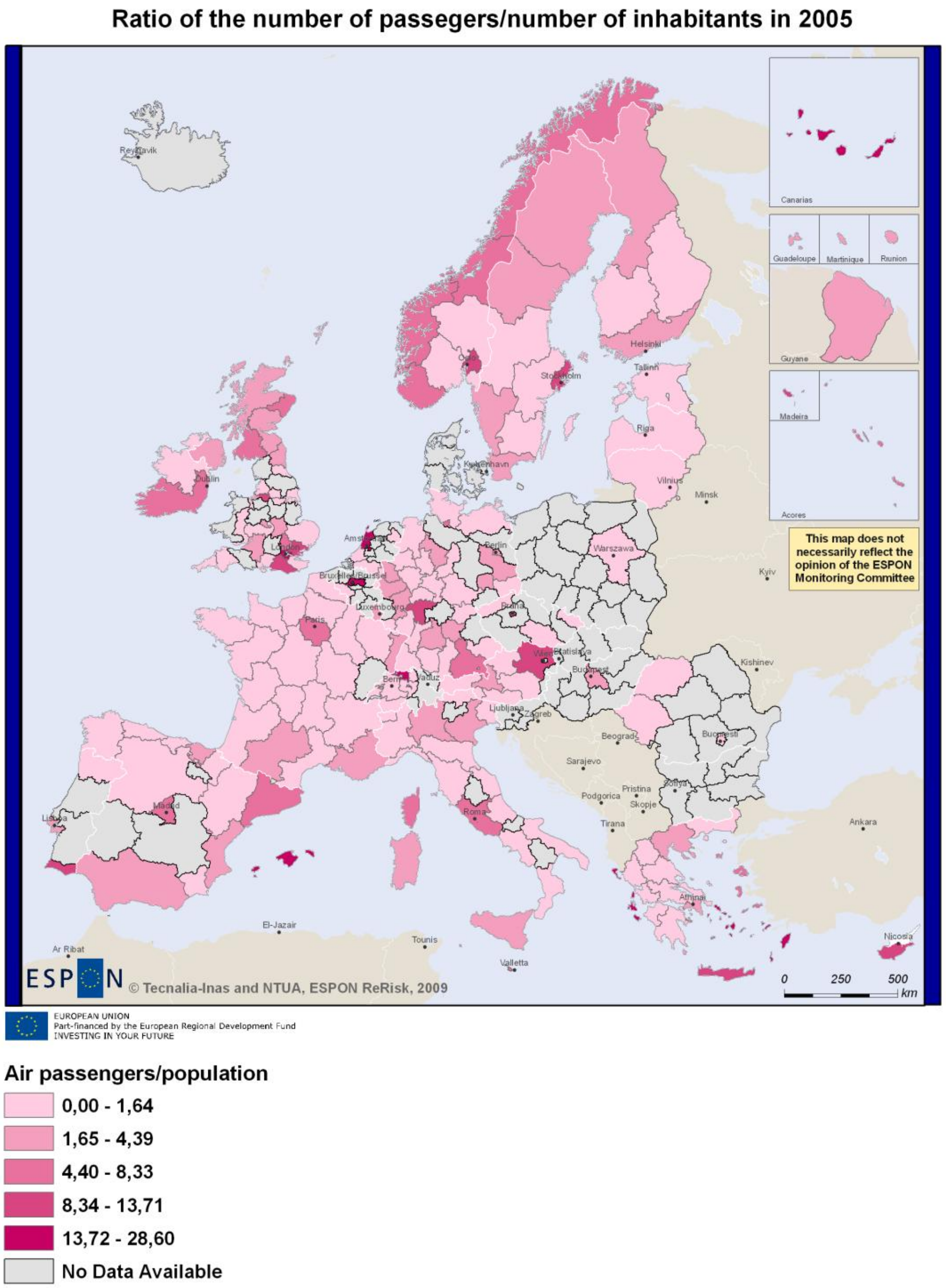
Source: Own elaboration based on Eurostat data

Map 9 Percentage employment in transport in the EU regions (NUTS II)



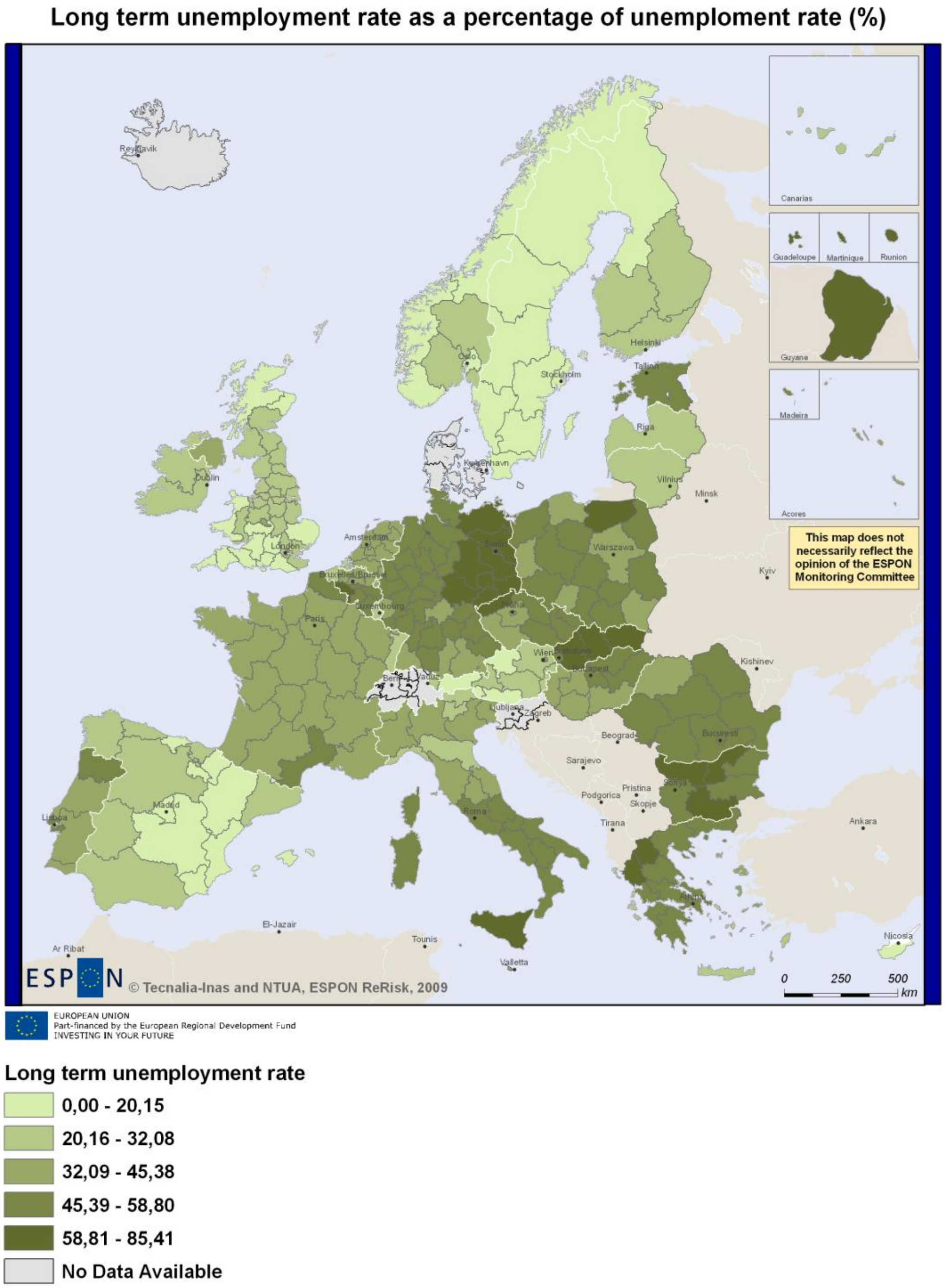
Source: Own elaboration based on Eurostat data

Map 10 Average age of cars in the EU regions (NUTS II)



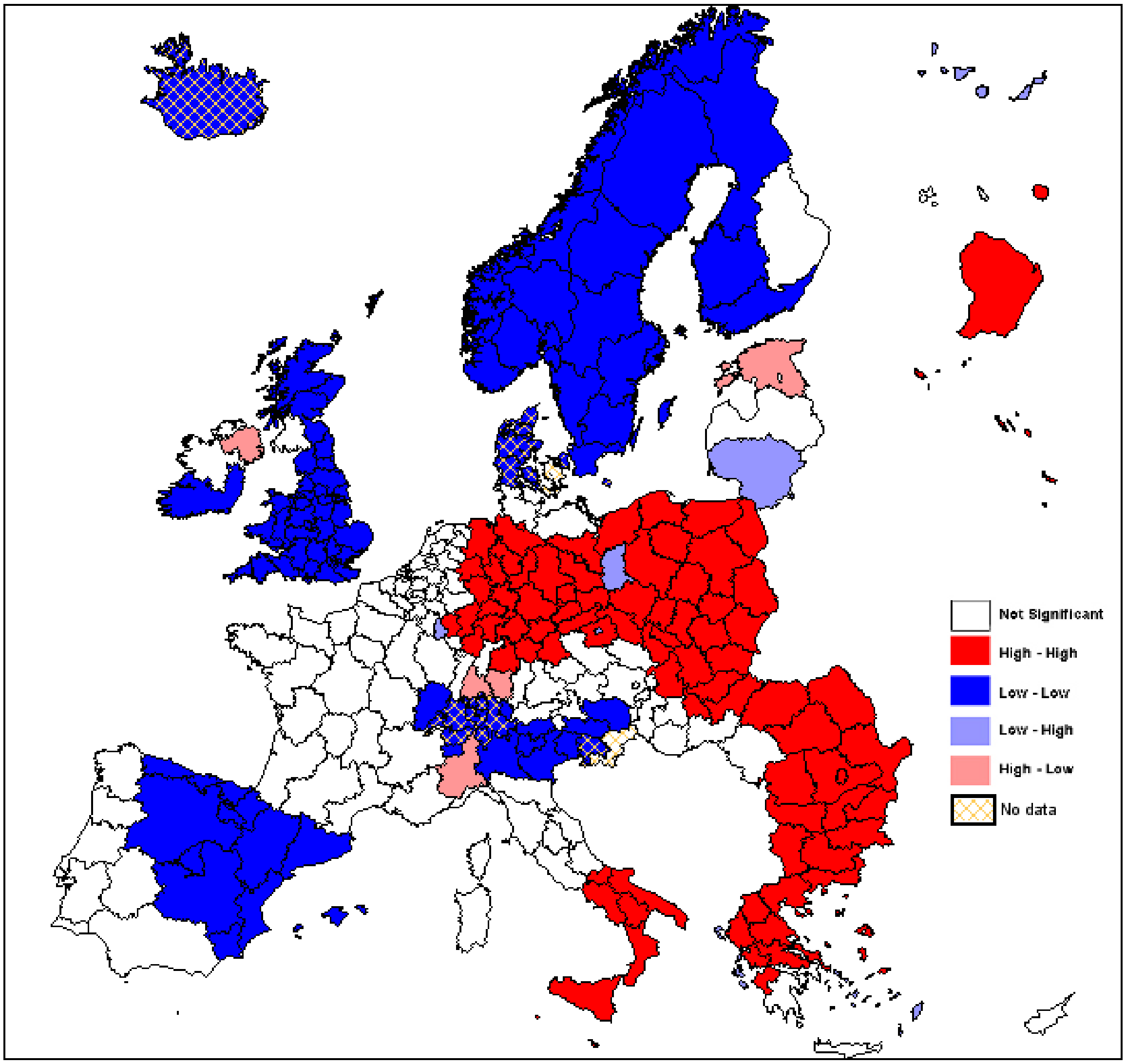
Source: Own elaboration based on Eurostat data

Map 11 Air passengers / population in the EU regions (NUTS II)



Source: Own elaboration based on Eurostat data

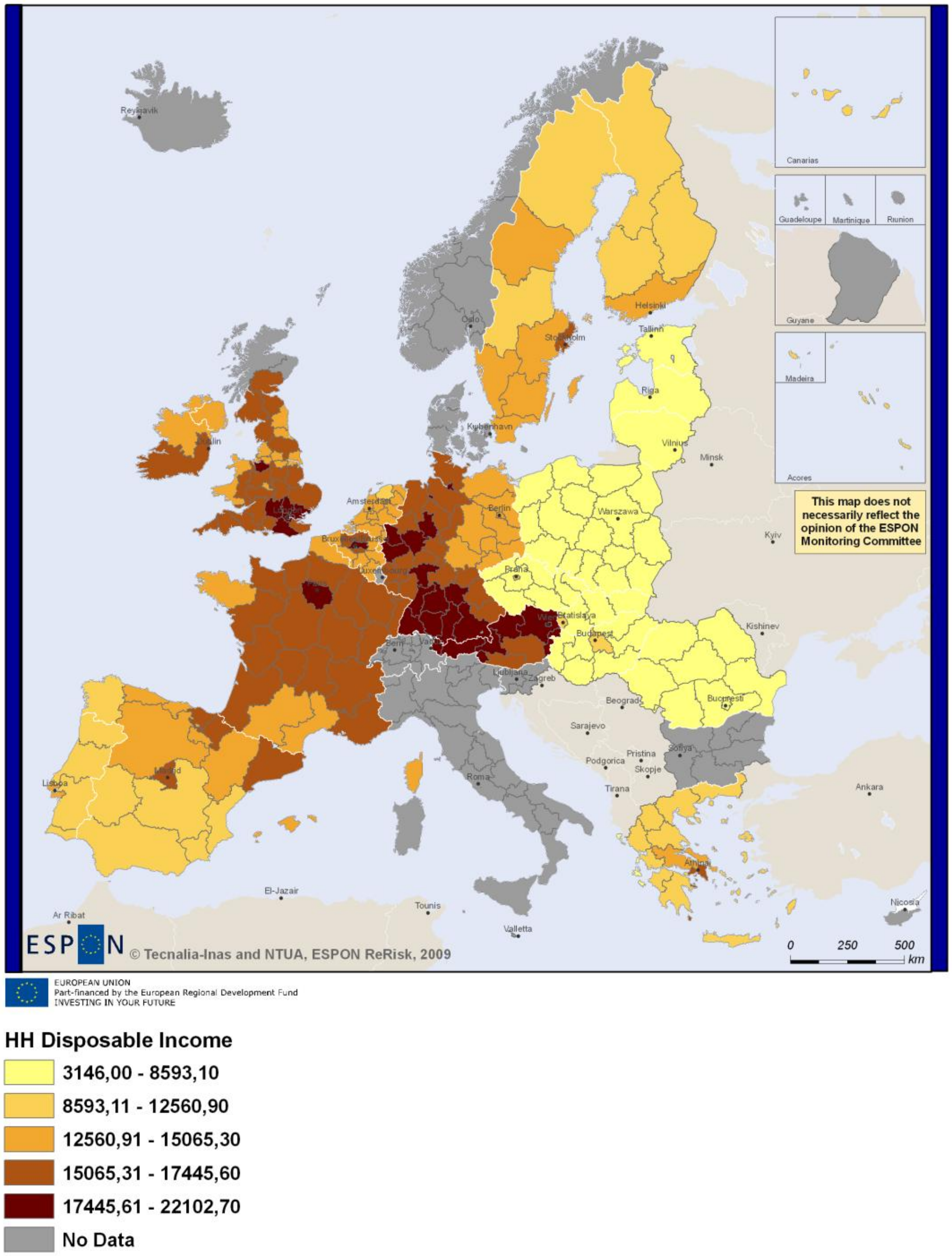
Map 12 Long term unemployment rate in the EU regions (NUTS II)



Source: Own elaboration based on Eurostat data

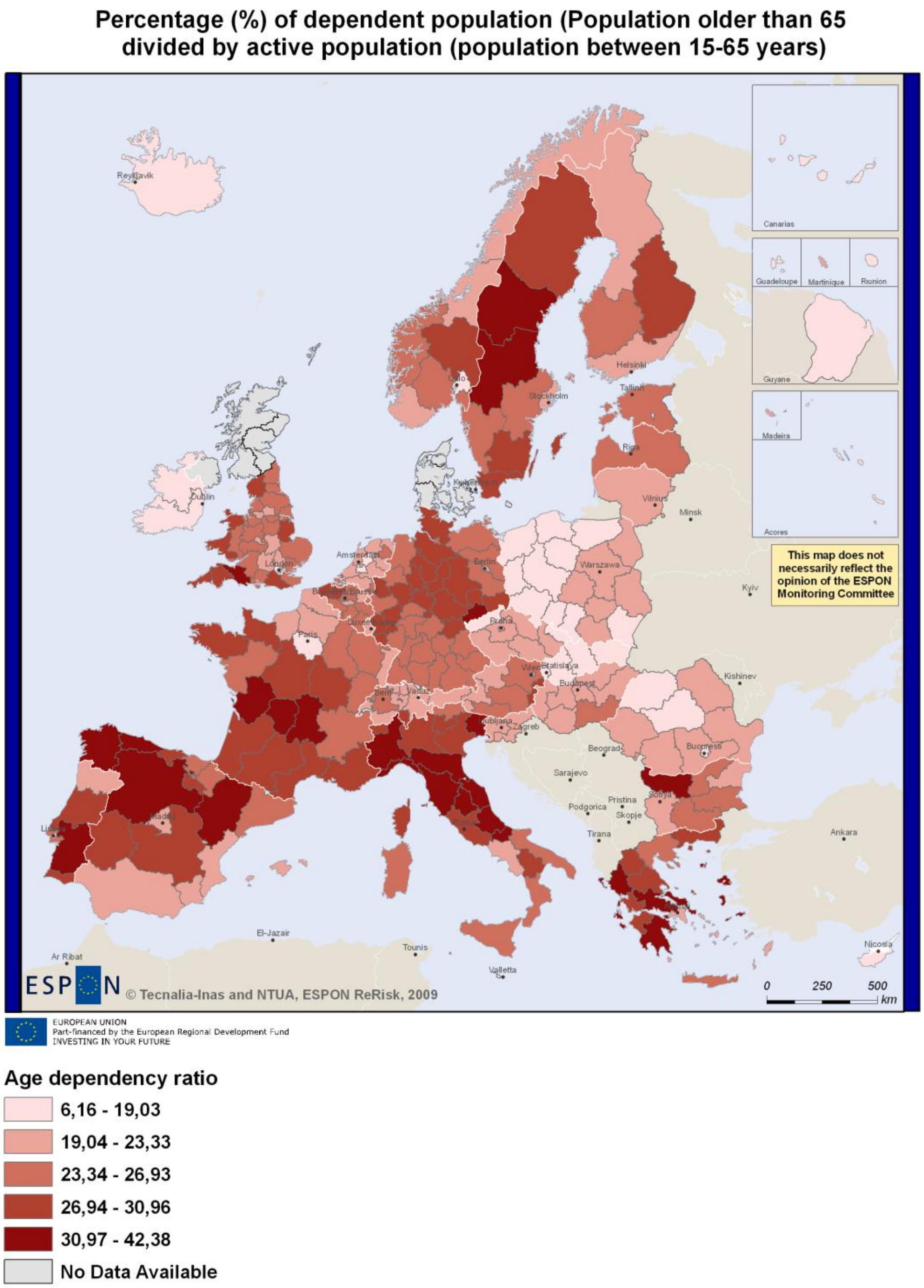
Map 13 LISA cluster map of the long-term unemployment rate

Disposable income in households



Source: Own elaboration based on Eurostat data

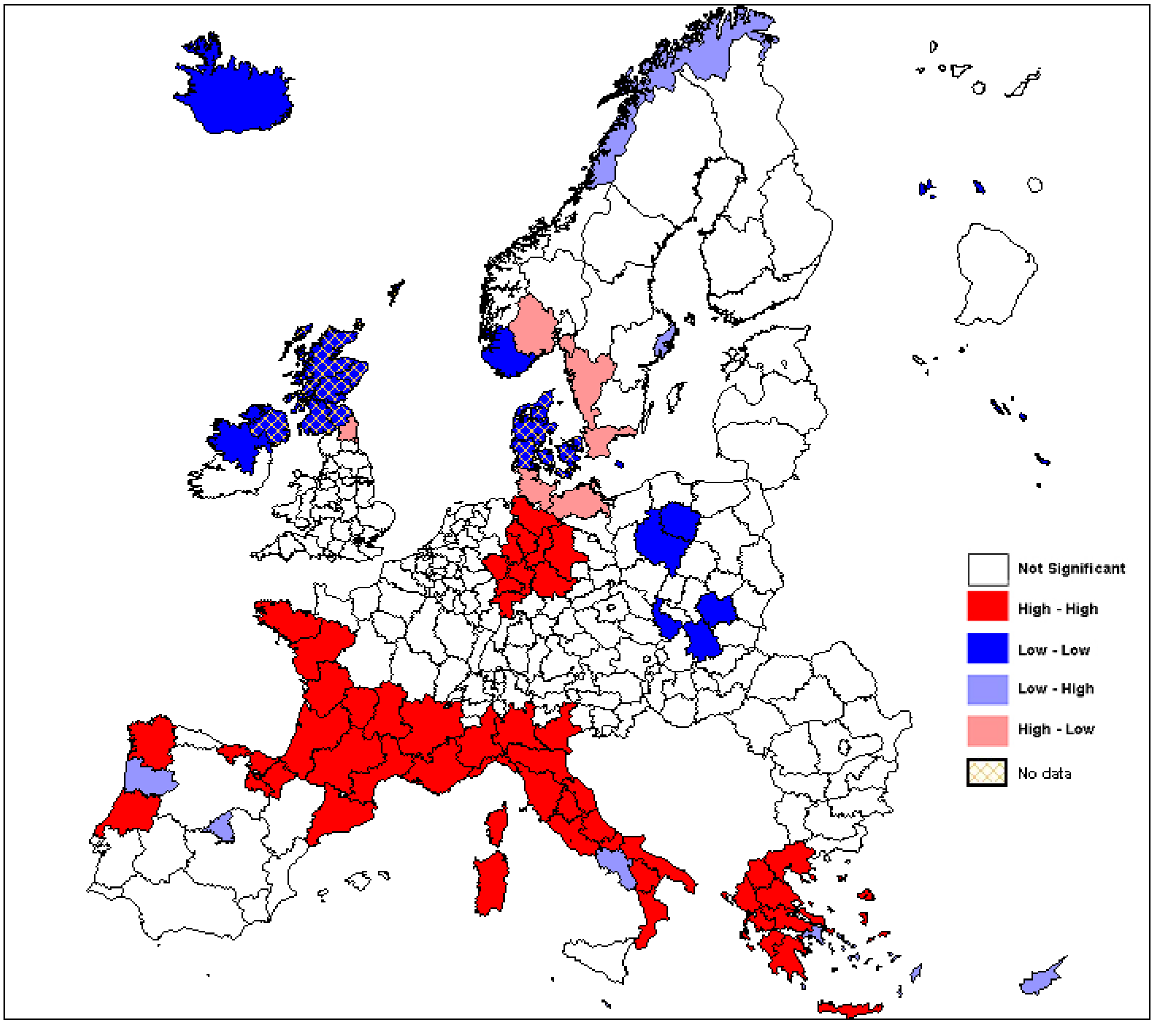
Map 14 Disposable income in households in the EU regions (NUTS II)



Percentage (%) of dependent population (Population older than 65 divided by active population (population between 15-65 years)

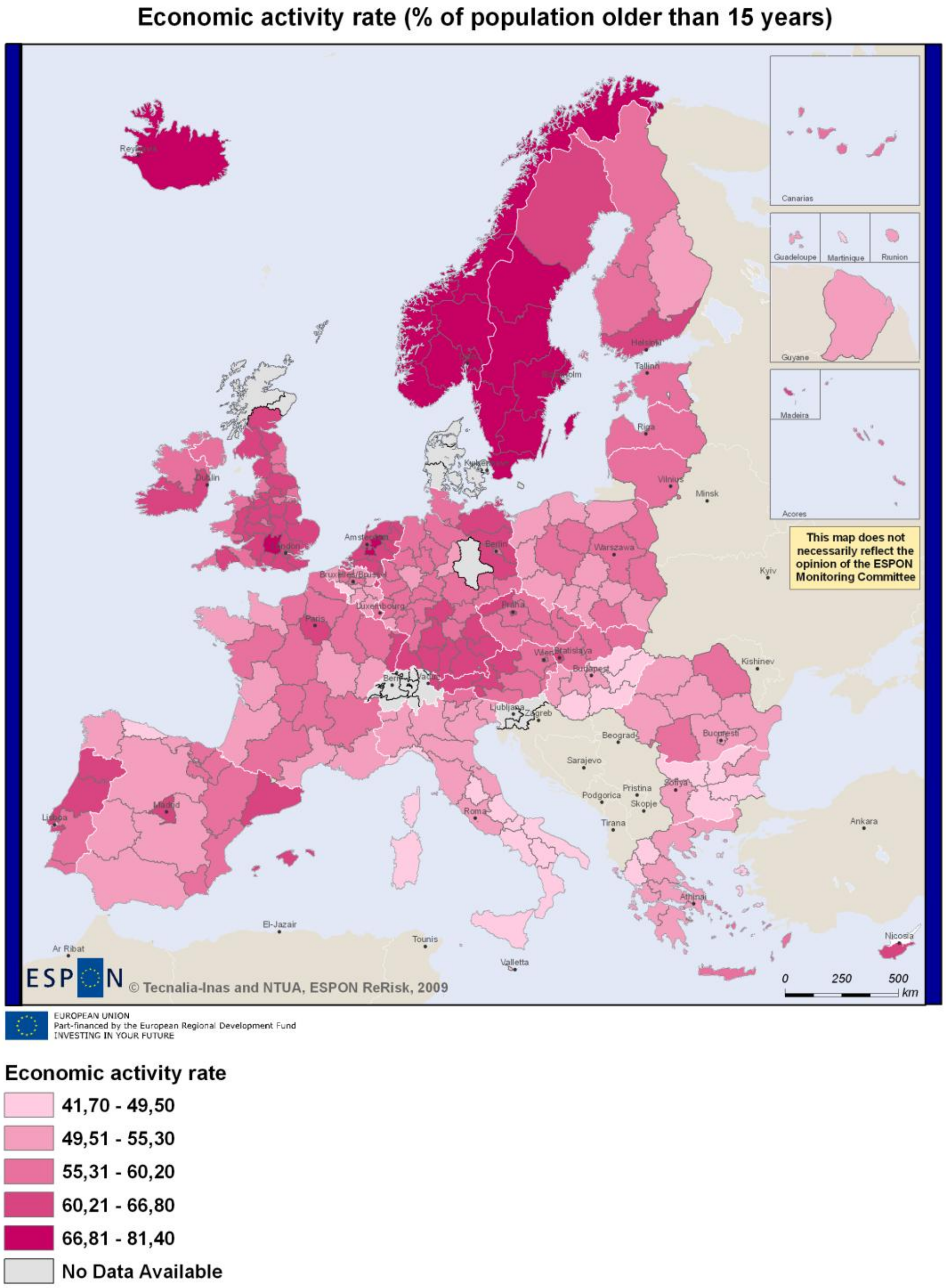
Source: Own elaboration based on Eurostat data

Map 15 Age dependency ratio in the EU regions (NUTS II)



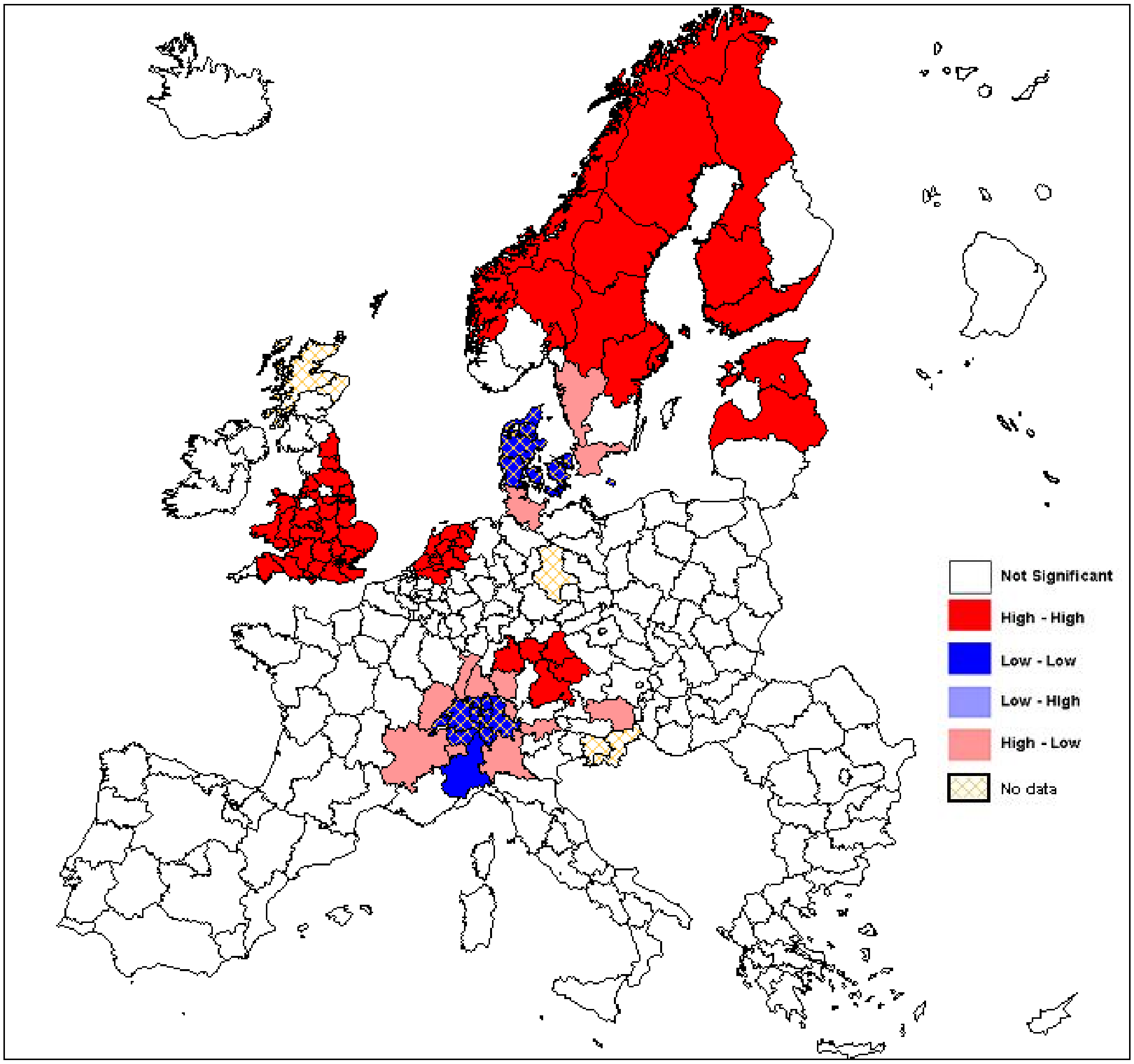
Source: Own elaboration based on Eurostat data

Map 16 LISA cluster map of the age dependency ratio



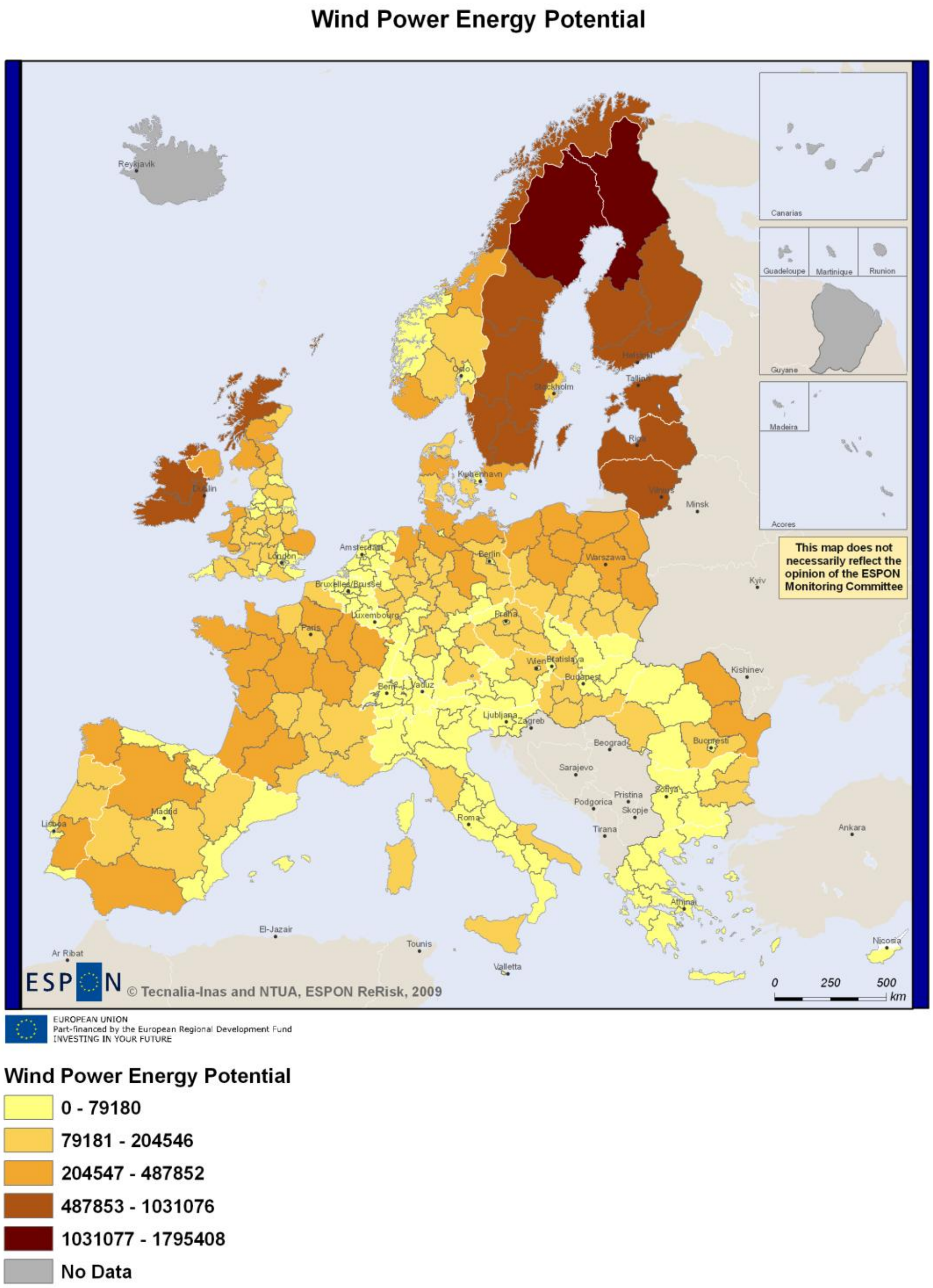
Source: Own elaboration based on Eurostat data

Map 17 Economic activity rate in the EU regions (NUTS II)



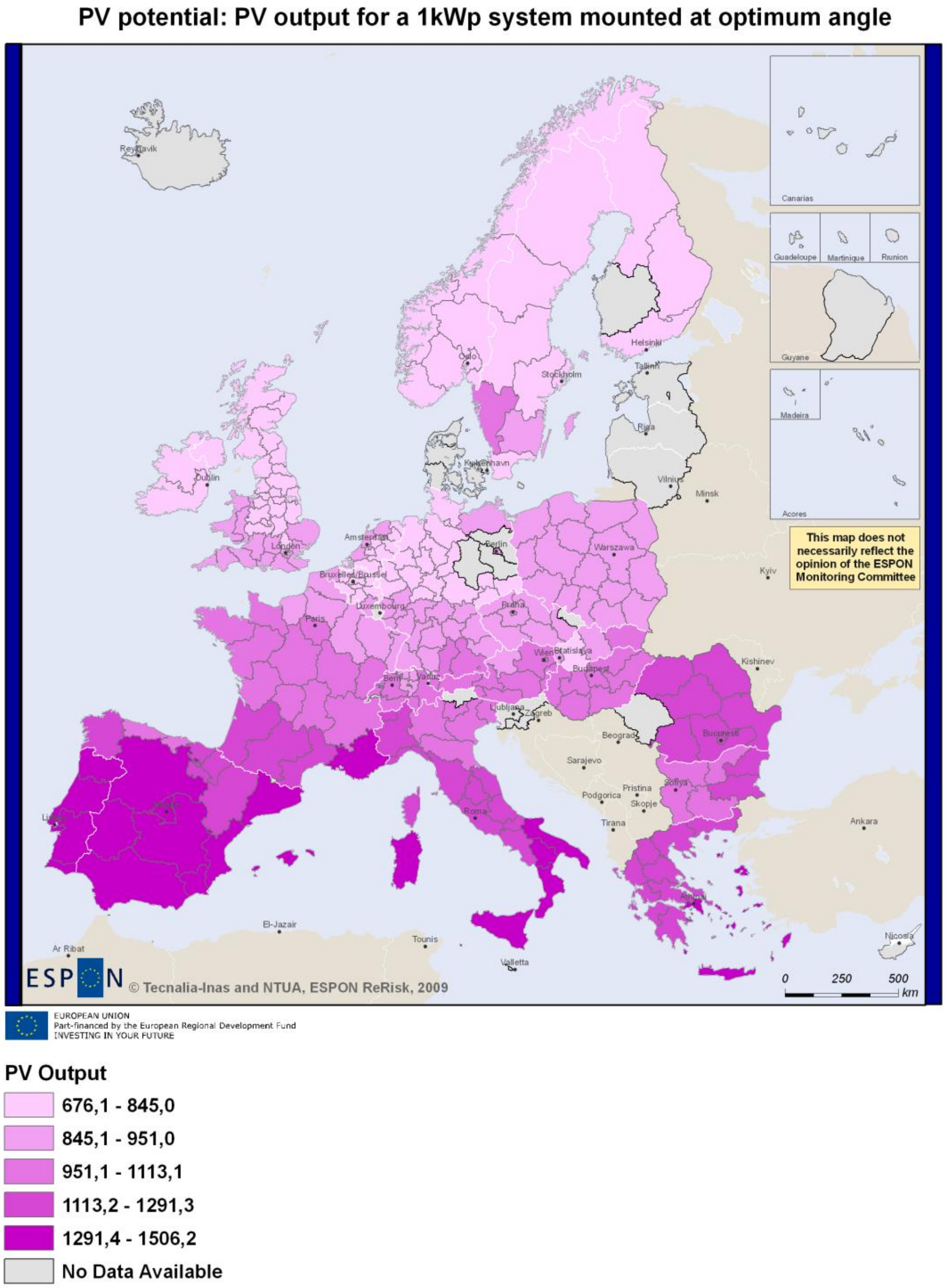
Source: Own elaboration based on Eurostat data

Map 18 LISA cluster map of economic activity rate



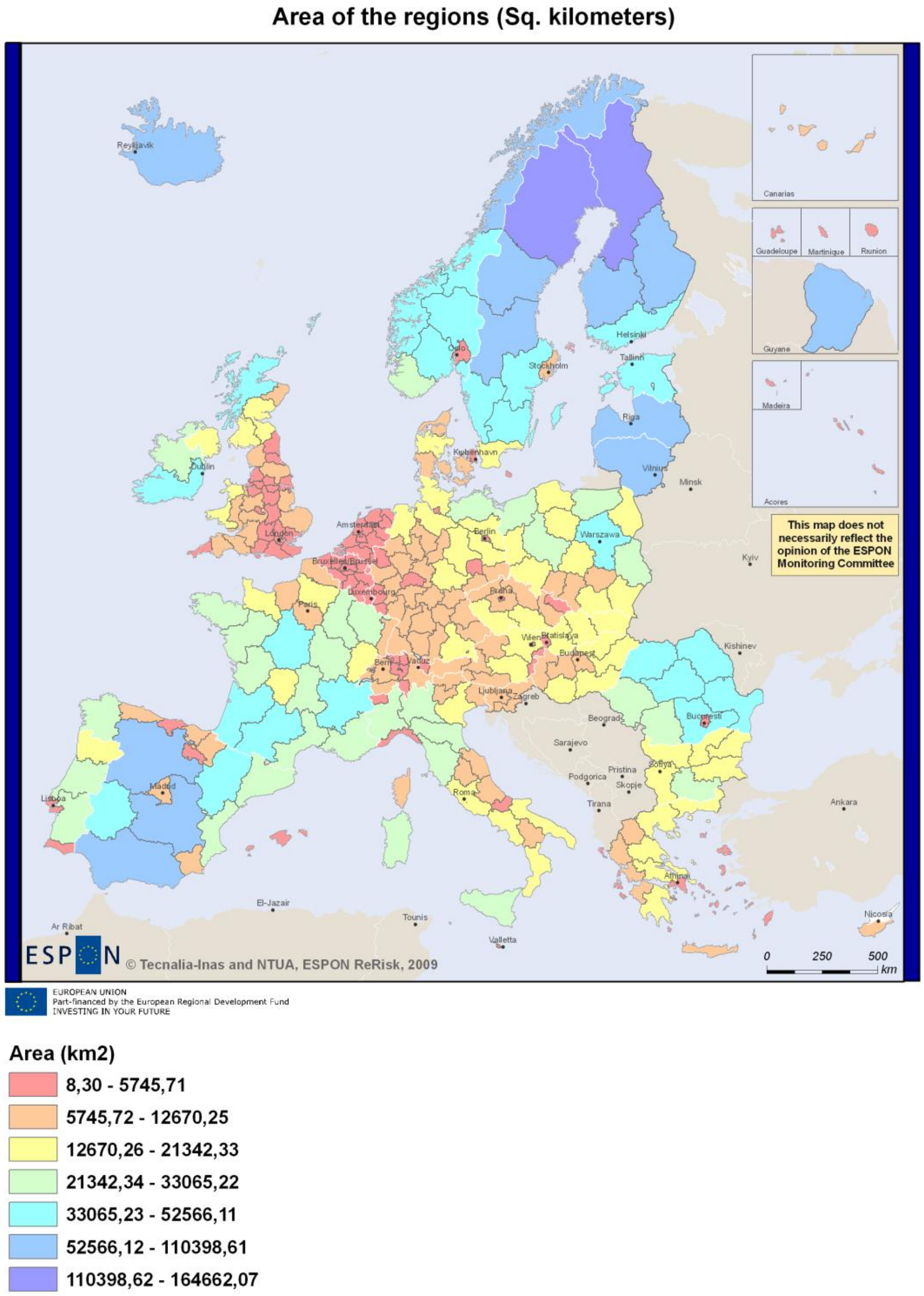
Source: Own elaboration based on European Topic Centre on Air and Climate change (ETC/ACC) data on wind intensity

Map 19 Wind Power Energy Potential in the EU regions (NUTS II)



Source: Joint Research Centre, Renewable Energies Unit

Map 20 PV potential in the EU regions (NUTS II)



Source: Own elaboration based on Eurostat data

Map 21 The area of the EU regions (NUTS II) in sq. kilometres