**Stelian STANCU, PhD (corresponding author)**
stelian.stancu@csie.ase.ro
Bucharest University of Economic Studies, Romania

**Andreea PERNICI, PhD Candidate**
andreea.pernici@csie.ase.ro
Bucharest University of Economic Studies, Romania

**Ion-Florin RADUCU, PhD Candidate**
florin.raducu@csie.ase.ro
Bucharest University of Economic Studies, Romania

**Daniela-Elena MARINESCU, PhD**
daniela.marinescu@csie.ase.ro
Bucharest University of Economic Studies, Romania

# Exploring Spatial Analysis Methods for Socio-Economic Profiling in the European Union Nuts2 Level

**Abstract.** *Socioeconomic profiling remains an ongoing priority for both researchers and public institutions, especially in the highly volatile context of the European Union. Therefore, in the current paper, we aim to compute several spatial analysis methods to extract observations regarding the performance of different development axes, at the NUTS2 regional level. For that objective, we will define the socioeconomic profile through five dimensions: economy, labour, science and technology, demographics, and education, each of them being represented by a specific indicator: GDP per capita, the unemployment rate, GERD, the median age of the population, and the participation rate in education. We consider this selection to be adequate for constructing a robust perspective, that can encapsulate the main characteristics and dynamics at the regional level, while also remaining concise. In terms of methodology, we will employ three methods, starting from Moran's I test for spatial autocorrelation, and then proceeding to spatial clustering through the K-Means algorithm. As a last step, we will also compare a linear regression model to a geographically weighted one, to decipher whether the spatial factor plays a role in defining the relationship between variables. Finally, throughout the application, our focus will be on interpreting the results and identifying local specificities, as well as differentiation points, which can then be integrated into the broad endeavour of constructing a socio-economic profile of the European Union regions.*

**Keywords**: *spatial autocorrelation, Moran's I, Geographically Weighted Regression, K-Means Clustering, socio-economic, NUTS2, European Union.*

**JEL Classification:** C21, C50, C38, E70.

DOI: 10.24818/18423264/58.2.24.05

## 1. Introduction

Socio-economic profiling has always been a necessity, as well as a priority, for both public institutions and researchers. Its main objective will be to identify regional specificities, which can be interpreted as a first milestone in designing the way forward, especially for international cooperation organisations. Thus, to synchronise the progress and inflict a real positive impact on the current economic structures, a highly personalised approach needs to be implemented, which takes into account the present dynamics of each area, as well as the general layout.

A similar exercise will be proposed in the current paper, through the illustration of a socio-economic profile of the European continent, using several elements. Considering this, we chose to focus on the European Union regions at a NUTS2 level, since it can provide a higher level of differentiation between the main nucleus of performance. We have then selected five socio-economic dimensions, which we consider comprehensive for the goal at hand, each of those being represented by a specific indicator. Starting from the economic dimension, we have included the GDP per capita, followed closely by labour, through the unemployment rate, and finally by science and technology, characterised by the gross expenditure in R&D. In terms of the social axis, we have turned our attention to two aspects: demographics, represented by the median age of the population, and education, defined through the participation rate.

Next, using these variables, we will follow a three-stage methodological journey. First of all, we will graphically represent the values registered for each NUTS2 location, making a first step in extracting significant socio-economic specificities. After that, since we are working with geographical data, we aim to assess the level of spatial autocorrelation presented by each indicator, defined as the measure of similarity between nearby observations. Our objective will be to evaluate whether closer regions tend to have homogeneous characteristics and, therefore, can be clustered together. This analysis will be done by computing the Moran’s I statistical test, as well as a Moran scatterplot. Afterward, since each indicator will present a certain level of positive autocorrelation, we will employ a *K-Means* algorithm on the aggregated dataset, designed to complete the regional picture and contribute to the socio-economic profile. Lastly, our aim is to determine if there are any strong connections between the indicators, and for that, we will employ two multivariate regression models, one linear and one using the geographically weighted process.

Finally, in terms of limitations, the most impactful one will be the lack of more recent NUTS2 data, which has caused us to analyse the 2021 year. At the same time, the datasets will be different for each indicator, with the aggregated one being slightly reduced, since there will be some missing values for several regions. Despite this, we consider the current model to be a good starting point in the socio-economic profiling of the region since it will integrate multiple development axes, as well as methodological instruments.

## 2. Related literature

In terms of the related literature, we have focused our attention on the NUTS2 spatial researches, to highlight what are the main subjects and insights collected. Therefore, we could find plenty of articles that tacked economic aspects through different concepts, such as economic growth (Jankiewicz, 2023), (Formánek, 2019a), macroeconomic dynamics (Formánek, 2019b), and economic convergence (Arbia et al., 2008). Moreover, the regional development overview will be completed by socio-economics topics such as unemployment (Filenta & Kydros, 2023a), tertiary education (Filenta & Kydros, 2023b), migration flows (Prada & Cimpoeru, 2023), agriculture productivity (Smit et al., 2015), science (Andersson et al., 2020), innovation (Żółtaszek & Olejnik, 2021), or even road safety (Wachnkicka et al., 2021). However, what is interesting is the fact that the vast majority of these references will be recent, so the current interest in this type of analysis is high, stimulated also by European mechanisms and strategies. At the same time, we can pinpoint all the socio-economic dimensions that we also integrated into the analysis, so we can consider the current approach as a catalyst that aggregates several perspectives into a concise overview.

Lastly, after gaining an understanding of the main subjects studied, we also want to illustrate the main methodologies involved. Many of the references employed different complex econometric models, such as those specific to panel data, which take into account both spatial and temporal dependencies (Jankiewicz, 2023), (Formánek, 2019a), or the geographically weighted regression, a frequently computed method (Wachnkicka et al., 2021), (Prada & Cimpoeru, 2023). Other methodological choices are the spatial Mankiw-Romer-Weil model (Smit et al., 2015) or the social network analysis in the context of spatial data (Filenta & Kydros, 2023a), (Filenta & Kydros, 2023b). However, we could not find any combinations focused on clustering, so we can consider this a new contribution to the field. At the same time, this paper adds to our previous profiling models, such as the one based on green energy and digitalisation (Pernici & Stancu, 2023).

## 3. Methodology

### *3.1 Dataset and NUTS2 framework*

Going forward, as previously mentioned, the current paper aims to construct a socio-economic profile of the European Union, at the NUTS 2 level, based on 5 dimensions: economy, labour, science, and technology, demographics, and education. Each dimension will be represented by the indicators described in detail in Table 1.

**Table 1. Dimensions and indicators definition**

| Dimension | Name | Definition | U.m. | Eurostat Code |
|---|---|---|---|---|
| Economy | **GDP** | GDP per capita | Euro per inhabitant | nama_10r_2gdp |
| Labor | **UNEMP** | Unemployment Rate, all ISCED 2011 levels | % | tgs00010 |
| Science and Technology | **GERD** | Gross expenditure on Research & Development | Euro per inhabitant | rd_e_gerdreg |
| Demographics | **AGE** | Median age of the population | Years | demo_r_pjanind 2 |
| Education | **EDU** | Participation rate in education or training | % | trng_lfse_04 |

*Source*: Authors 'own creation.

In terms of the NUTS2 classification, we will use the latest one, published in 2021, which includes 242 regions from the European Union, plus several other countries. The framework has been extensively used by all EU institutions, with Eurostat defining it as a hierarchical system created to divide the EU and UK economic territories for several purposes: the collection, development, and harmonisation of regional statistics, the construction of socio-economic analyses, or the consequent creation of regional policies.

Next, in terms of the methodological journey, we will describe three instruments that will be useful in our research objective: the Moran's I method for identifying spatial autocorrelation, the K-Means clustering method, and both linear and geographically weighted regression. As a last mention, all of the methods have been computed using different R packages.

### *3.2 Spatial autocorrelation. Moran's I Test*

When discussing spatial analysis, the first notion that needs to be studied is the spatial autocorrelation. In a simple definition, this concept will evaluate how similar or dissimilar objects are in comparison with their neighbours, or the degree of alignment between the relative magnitude of their values (Griffth & Chun, 2018). Thus, one of the most used methods that will evaluate spatial autocorrelation is Moran's I analysis, developed by Moran (1950) and then widely applied to a multitude of domains and models.

Therefore, the first step in this computation will be constructing the *spatial weights matrix of neighbours*, which will represent the spatial relationships in a geographical dataset. In our case, we will use it based on *binary contiguity*. The functionality behind it is that the neighbours will be identified based on a 1-unit buffer around each polygon, and if two polygons share boundaries within this buffer, they are considered neighbours, with a weight set to 1. If they do not share a boundary, then the weight is set to zero, thus resulting in the $W$ binary matrix, similar to the one below, where the diagonal elements will be set to zero.

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

$$\text{where } w_{ij} = \begin{cases} 1 \; if \; locations \; i \; and \; j \; are \; contiguous \\ 0, \; otherwise \end{cases} \quad (1)$$

Next, after computing the spatial weights matrix, we will apply Moran's I measure, based on the formula described in equation (2). The elements will be as follows: $n$ is the total number of observations, $x_i, x_j$ will be the values of the spatial units indexed by $i$ and $j$, $\bar{x}$ is the mean of the variable of interest, $W$ is the sum of the spatial weights (or all the elements in matrix $W$) and $w_{ij}$ are the spatial weights between observations $i$ and $j$.

$$I = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (2)$$

In terms of interpretation, Moran's I method will identify three possible scenarios: positive spatial autocorrelation (measure close to 1), meaning that the values will be similar to their neighbours and that certain clusters can be formed on the map, zero or no spatial autocorrelation (measure near 0), meaning that there will be no clear pattern visually and that random values will be given to close objects, and negative spatial autocorrelation (measure close to -1), in which values are dissimilar to their neighbours, patterns are dispersed, and the map tends to show checker-style polygons.

Lastly, Moran's I measure will be transposed into a statistical significance test, with the null hypothesis being that the variable of interest shows no spatial autocorrelation, or shows a random spatial pattern, a fact that can be either confirmed or denied. For the computation to be complete, we will also illustrate a Moran scatterplot, which will visually translate the method. Shortly, the graph will display the values of the variable of interest on the abscissa, while the ordinate will show the spatial lag or the values averaged over the neighbouring locations. To interpret it, we will look at the points distribution; if those are scattered around the diagonal, a positive spatial autocorrelation is represented.

Therefore, by using all these elements, we will be able to assess whether there will be spatial relationships between our observations in the case of each indicator, helping us to proceed further with the aggregation of the dataset.

### *3.3 K-Means Clustering*

After the individual study of each indicator has been completed, we will aggregate the 5 variables and use the newly created dataset to apply the K-Means algorithm. This method is one of the most used ones for clustering exercises, having the potential to be applied to a multitude of datasets and conditions. In our case, we will compute the function using Euclidean and Ward's distances (3), along with an optimised initial configuration.

$$W = \sum_{k=1}^{g} W_k\,, \quad where\ W_k = \sum_{i=1}^{n} \sum_{m=1}^{p} (x_{i,m} - \bar{x}_{k,m})^2 \quad (3)$$

For the optimal allocation scenario, a map will be constructed that will show the regional patterns and clusters. Based on the indicators' averages, we will then define each cluster in terms of socio-economic performance.

### *3.4 Linear and Geographically Weighted Regression*

Finally, in the last section of our paper, we will present two regression models, to mark the main connections between the five indicators selected. Since we are working with spatial data, we will explore the comparison between linear and geographically weighted regression (GWR). Thus, in terms of linear regression, the method is widely known, so we will define only its equation and main components, in a multivariate formula (4). Thus, $Y$ will represent the dependent variable, $X_i$ the independent ones, $B_i$ the coefficients, and $\varepsilon$ the residuals.

$$Y = B_0 + B_1X_1 + \cdots + B_nX_n + \varepsilon \quad (4)$$

Next, in the case of GWR, this technique will be a spatial analysis method used to explore and model varying geographical relationships, being seen as an extension of traditional linear regression. Therefore, the main difference will be that the independent variables will now account for spatial dependencies since they will be computed using the geographical weights. Thus, the equation for each location $i$ becomes (5).

$$Y_i = B_0 + B_1WX_{i1} + \cdots + B_nWX_{in} + \varepsilon_i,$$
$$where\ WX_{ij} = \sum_{k=1}^{n} w_{ik}hX_{kj} \quad (5)$$

In the formula above, $h$ will define the bandwidth or kernel width, one of the most important elements in the computation of GWR models. In other words, the bandwidth will be the size of the local neighbourhood used to estimate the coefficients. Since we will use R, we can employ the *gwr.sel* function, which performs a cross-validation process to find the optimal bandwidth that minimises the prediction error.

Afterward, we will interpret the results using mainly the *quasi-global R squared*. It is important to note that in GWR, the relationship between the dependent variable and the independent ones varies across space, and thus, a single value may not adequately capture the model's performance. Thus, the *quasi-global R squared* is an attempt to address this issue by considering the overall explanatory power of the model over the entire region. In terms of formula, the indicator will be expressed in equation (6), where $\hat{y}_{ij}$ $\bar{y}_{ij}$ will be the predicted, respectively mean values of the dependent variable at the $j$ observation of location $i$.

$$R^2_{quasi-global} = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{ij})^2} \quad (6)$$

## 4. Results

Finally, the results section will be described individually for each of the five dimensions. As mentioned briefly in the introductory part, the number of NUTS2 regions included and the graphical representations will be slightly different for each indicator, based on the data available.

### *4.1 Economy*

First, the economic dimension will be represented by the GDP per capita, with values found for 273 NUTS2 units, some of them outside the EU border. The graphical representation is visible in Figure 1, where a more intense colour emphasises a higher level of the gross domestic product.

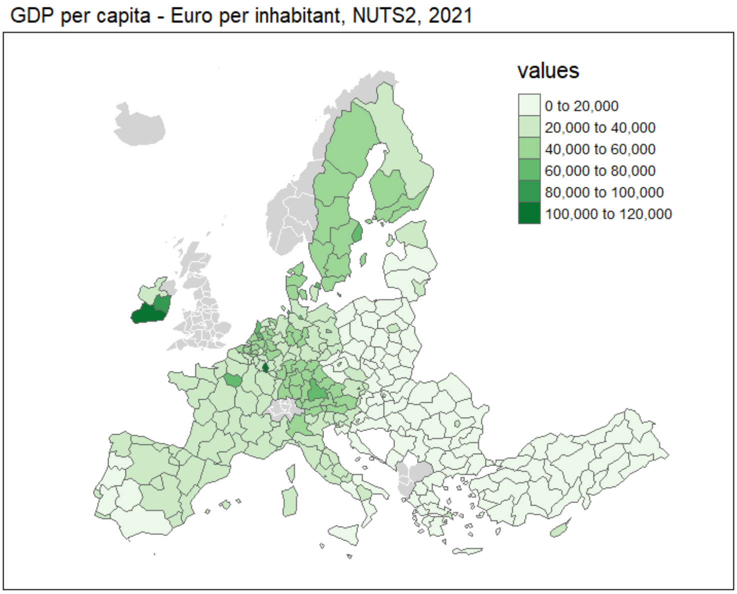


**Figure 1. GDP per capita distribution, NUTS2, 2021**
*Source*: Authors 'own creation.

Therefore, in terms of performance, we can see a certain homogeneous distribution, except for certain regions that can be considered poles of wealth. For example, Luxembourg and Southern Ireland have a GDP per capita of over 100.000 euro per inhabitant, while Eastern and Midland Ireland, Hovedstaden in Denmark, Stockholm in Sweden, and BE10 Region de Bruxelles in Belgium will exceed the 70.000 euro threshold. This can be explained by a multitude of factors, starting with the existence and development of important industrial centres around capitals such as Dublin, Copenhagen, or Stockholm. In time, these centres managed to attract foreign direct investments and become technology and innovation headquarters, while also providing an enhanced quality of life.

After we have seen the graphical distribution, we can now apply Moran's I test to observe whether the GDP per capita indicator will show any spatial relationships between the NUTS2 regions. Therefore, in Table 2 we can see the results, with the Moran's I statistic registering a value of 0.65 and the *p-value* being significantly lower than the 0.05 threshold, concluding that the series presents a moderate positive autocorrelation and that certain patterns can be identified in neighbouring observations. The same conclusion can be extrapolated from the scatterplot represented in Figure 2, where we can see the observations close to the slope.

**Table 2. Moran's I Test, GDP per capita**

| | |
|---|---|
| Moran's I Statistic | 0.65 |
| p-value | < 2.2e-16 |

*Source*: Authors 'own creation.

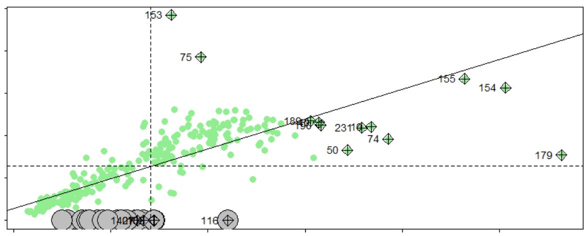


**Figure 2. Moran scatterplot, GDP per capita, NUTS2**
*Source*: Authors 'own creation.

### *4.2 Labour*

Going forward to the labour dimension, represented by the unemployment rate, this time we could find 259 NUTS regions. Thus, in Figure 3 we can see the European distribution. In terms of patterns, one thing we can observe is that the Southern regions of Europe seem to have higher unemployment values. In this category, we can mention the Spanish regions Ciudad de Ceuta, Carairas, or Adalucia, which will all register worrying values above 25% percent. At the same time, certain Greek and Italian regions will showcase a low performance, such as Dytiki Makedonia, Campania, Notio Aigaio, or Sicilia. This can be explained by a multitude of factors, such as the economic reliance on extremely volatile sectors, for example, tourism or services. Also, it is important to remember that the studied year is 2021, so the COVID-19 pandemic effects surely affected this indicator.

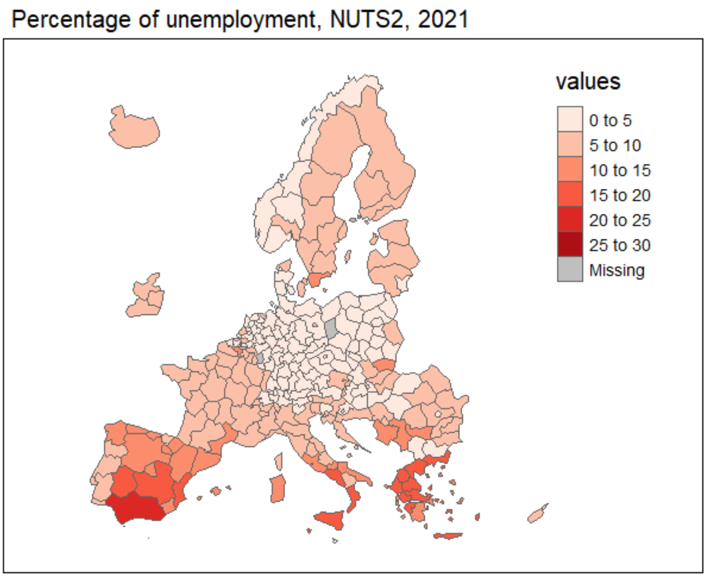


**Figure 3. Unemployment Rate distribution, NUTS2, 2021**
*Source*: Authors 'own creation.

However, another observation is that although the high unemployment regions are easy to place geographically, they will not be neighbouring, so we should expect a lower Moran I statistic value. This is confirmed by the results illustrated in Table 3 and Figure 4, where we can see a significant *p-value,* but a weaker positive spatial autocorrelation.

**Table 3. Moran's I Test, Unemployment Rate**

| | |
|---|---|
| Moran's I Statistic | 0.46 |
| p-value | < 2.2e-16 |

*Source*: Authors 'own creation.

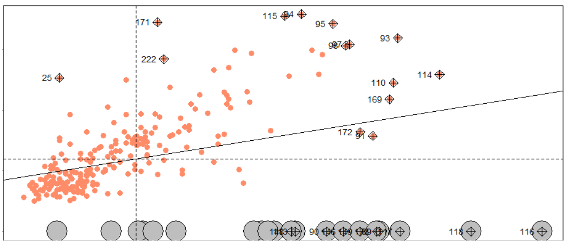

**Figure 4. Moran scatterplot, Unemployment Rate, NUTS2**
*Source*: Authors 'own creation.

### *4.3 Science and technology*

Next, in terms of science and technology, we have chosen the GERD per capita indicator, or the gross domestic expenditure on research and development, which will sum up all budgets allocated on R&D made by business enterprises, higher education institutions, governments, and private non-profit organisations. Therefore, we could find 268 NUTS2 regions, graphically represented in Figure 5.

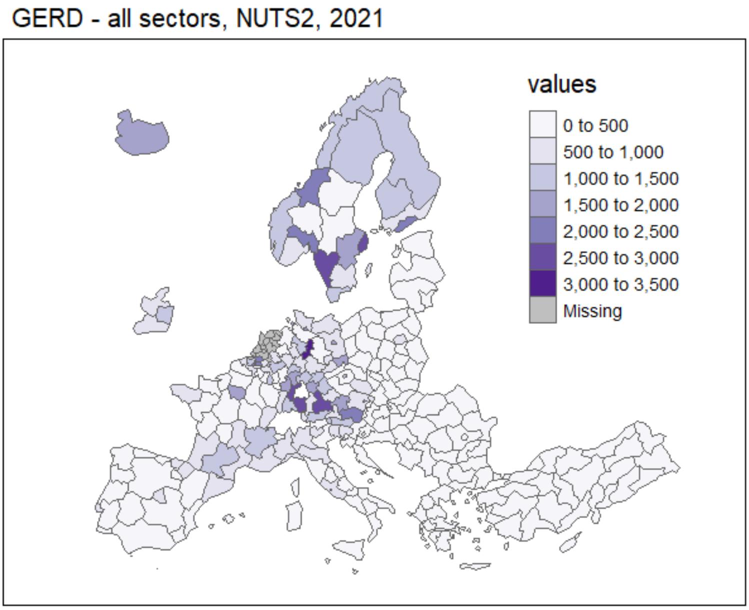


**Figure 5. Gross Expenditure on R&D, NUTS2, 2021**
*Source*: Authors 'own creation.

Therefore, this time, we can profile a category with high expenditures, in which we see the prevalence of German regions, such as Braunschweig, Oberbayern, or Tubingen, as well as some Sweden locations: Vastsverige and Stockholm. We should also note that we excluded the Belgium region Brabant Wallon and the German Stuttgart from the graphical representation, since they were clear outliers with values of over 3.700 euros per inhabitant allocated to R&D. The good performance of the mentioned areas can be explained by the prioritisation of science and technology as a development axis, with many research institutions and universities continuously growing throughout the last years. However, in addition to those, we can see that the rest of the continent will register similar and considerably small values, up to 1.000 euros per inhabitant.

Another interesting aspect that prevails is the fact that there will be no clear clusters of performance, with a more puzzled aspect, so we can expect a lower Moran I value. This will be confirmed by the results presented in Table 4 and Figure 6, with the lowest measure calculated so far, however, still significant.

**Table 4. Moran' s I Test, GERD**

| | |
|---|---|
| Moran's I Statistic | 0.4 |
| p-value | < 2.2e-16 |

*Source*: Authors 'own creation.

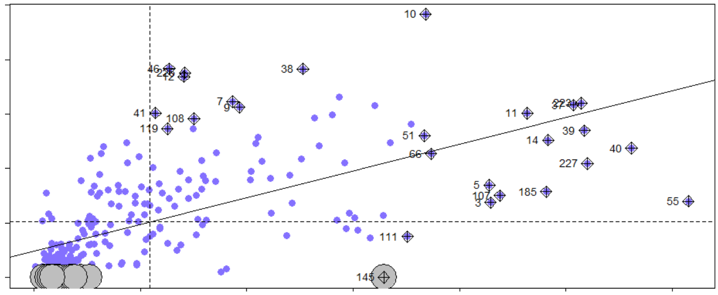

**Figure 6. Moran scatterplot, GERD, NUTS2**
*Source*: Authors 'own creation.

### *4.4 Demographics*

Next, in terms of demographics, represented by the median age of the population, we have the most complete dataset, composed of 286 European NUTS2 regions. Moreover, the general distribution seems to be much more homogeneous (Figure 7), with most of the regions registering values between 40 and 50 years old. However, certain regions will fall into the aging population category, most of them from Germany (Chemnitz, Sachsen-Anhalt, or Mecklenbur-Vorpommern), Italy (Liguria), or Spain (Principado de Asturias). One of the youngest populations from the European Union will be living in France Mayotte or Guyanne. Outside of the community block, we can observe several Turkish regions with a population age of below 30 years old.

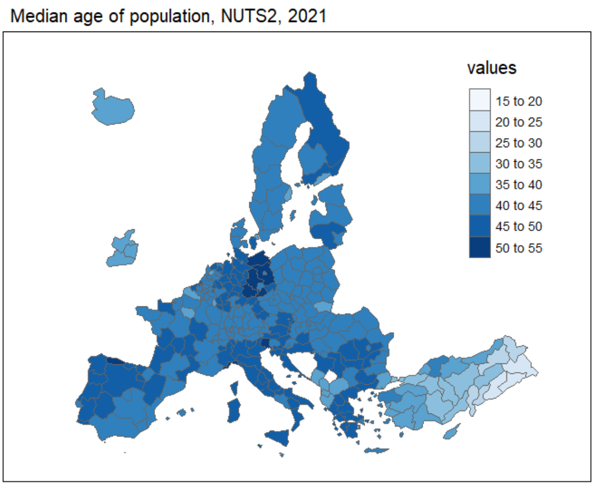


**Figure 7. Median age of population, NUTS2, 2021**
*Source*: Authors 'own creation.

At the same time, the values registered for the Moran I test will be significant and will show a moderate positive autocorrelation, as observed in Table 5 and Figure 8.

**Table 5. Moran' s I Test, Median Age**

| | |
|---|---|
| Moran's I Statistic | 0.62 |
| p-value | < 2.2e-16 |

*Source*: Authors 'own creation.

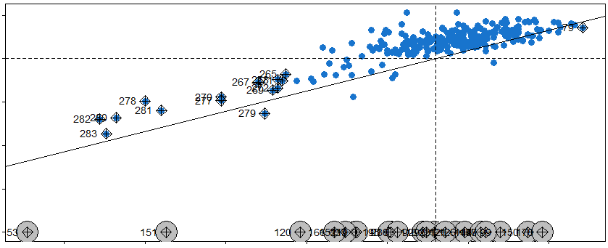

**Figure 8. Moran scatterplot, Median Age, NUTS2**
*Source*: Authors 'own creation.

### *4.5 Education*

Lastly, in terms of education performance, defined by the participation rate in education or training, we have a similar number of NUTS2 regions, namely 259. Thus, in Figure 9 we can observe a cluster of regions that stands out, namely Sweden, where the indicator averages will be over 30%.

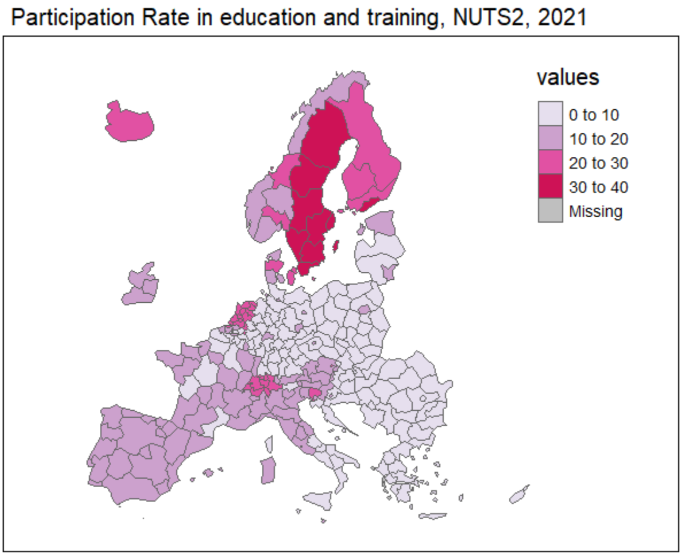


**Figure 9. Participation Rate in education distribution, NUTS2, 2021**
*Source*: Authors 'own creation.

Next in line, we have several regions from Finland (Helsinki-Uusimaa, Länsi-Suomi, or Etelä-Suomi), the Netherlands (Utrecht, Groningen, Noord-Holland), or

Switzerland (Zürich), but not limited to them, with values in the 20% to 30% interval. It's important to mention that this indicator will be representative of socio-economic development since education is a continuous process that should not end when certain levels (such as secondary or tertiary) are achieved. The Nordic countries, and especially Sweden and Finland, have understood that, having now one of the best education systems in the world, internationally recognised by multiple evaluation frameworks (for example, PISA).

At the same time, in terms of the Moran I test, we can observe the highest value registered until now, so there will be a more intense positive autocorrelation, with several patterns that can be identified throughout the spatial data (Table 6 and Figure 10).

**Table 6. Moran' s I Test, Participation Rate**

| | |
|---|---|
| Moran's I Statistic | 0.77 |
| p-value | < 2.2e-16 |

*Source*: Authors 'own creation.

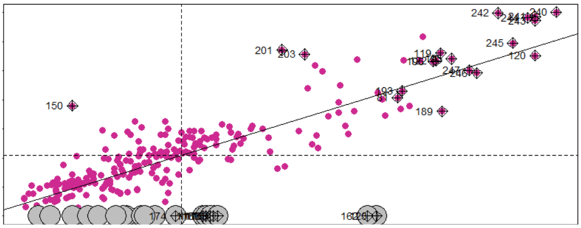

**Figure 10. Moran scatterplot, Participation Rate, NUTS2**

### *4.6 K-Means Clustering*

After the individual study, we can now perform the *K-Means* algorithm on the aggregated data. Thus, when merging the 5 indicators, we generated a new dataset composed of 219 NUTS2 regions, after the elimination of all missing observations. Therefore, after we have decided that the optimal number of clusters is 4 (by looking at the quality of the partition), we can use the averages available in Table 7 to define the clusters and ultimately profile the regions (Figure 11).

**Table 7. Average Indicators by Cluster**

| Cluster | No. | GDP per Capita | Unemployment Rate | GERD | Median Age | Participation Rate |
|---|---|---|---|---|---|---|
| **Cluster 1** | 38 | 54.357 € | 6.1% | 1.819 € | 42.1 years | 17.8% |
| **Cluster 2** | 77 | 32.971 € | 6.2% | 604 € | 47 years | 10.7% |
| **Cluster 3** | 75 | 18.049 € | 5.4% | 231 € | 42.7 years | 6.3% |
| **Cluster 4** | 29 | 16.462 € | 16.8% | 148 € | 44.8 years | 7.1% |

*Source*: Authors 'own creation.

Therefore, the first cluster will gather 38 NUTS2 regions with the highest GDP per capita, GERD, and participation rate, while also registering the lowest unemployment rates and having the youngest population. In terms of distribution in the European Union, the vast majority of Sweden and Finland will be integrated here, as well as some locations in Central Europe. In terms of the second cluster, which is also the most populated (77 regions), we can see a high performance, but much more moderate versus cluster 1. The GDP and GERD will be considerably higher than the ones in clusters 3 and 4, while the median age of the population tends to be higher. On the map, we can observe the South of France, along with the North of Italy and Spain, and the vast majority of Central Europe as being part of this structure.

Going forward, the third cluster will gather the Eastern European countries, with a similar profile, while the last cluster will show the regions with the lowest performance, generated mainly by the high unemployment rate (Greece, Southern Spain, and Italy).

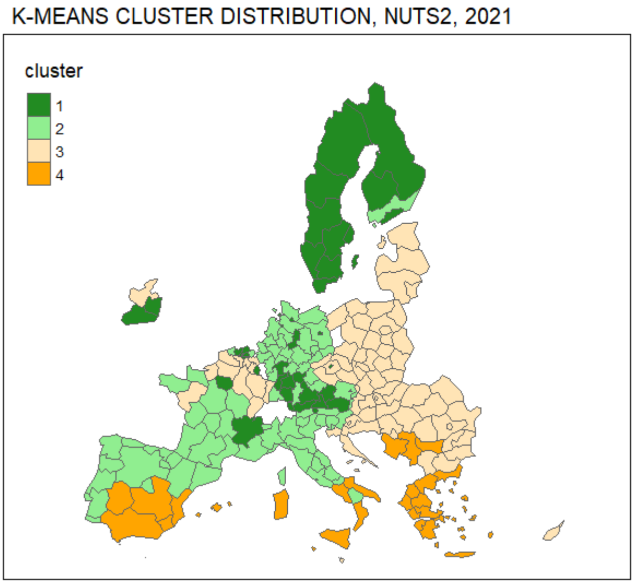


**Figure 11. Cluster distribution, EU, NUTS2**
*Source*: Authors 'own creation.

### *4.7 Regression Models*

Finally, in the last part of our analysis, we will present a comparison between a simple regression model and a geographically weighted one. Thus, the strongest relationship we could find in the aggregated dataset was between the level of GDP per capita and both the expenditure on R&D and the participation rate. Therefore, we have computed a linear regression model, in which we can observe that the relationship between the variables will be direct, significant, and moderate since all the coefficients are positive and R-squared takes a value of 54% (Table 8).

**Table 8. Linear Regression Results**

| Model 1 | Estimate | Significance level |
|---|---|---|
| Intercept | 14.396 | *** |
| GERD | 12.7 | *** |
| EDU | 708.6 | *** |
| Multiple R-Squared = 54% | | |
| F-stat = 124.9 *** | | |

*Source*: Authors 'own creation.

However, when we take into consideration the spatial factor and use the geographically weighted regression, we generate different results, with a more explicative model (Table 9). This is proven by the higher quasi-global R-squared value, of almost 80.2%.

**Table 9. Geographically Weighted Regression Results**

| Model 1 | Minimum | Median | Maximum | Global |
|---|---|---|---|---|
| Intercept | -21,228.71 | 10,202.88 | 36,184.61 | 14,396.34 |
| GERD | -0.157 | 13.2 | 64.45 | 12.66 |
| EDU | -239.77 | 639.49 | 6,621.45 | 708.52 |
| Quasi-global R-Squared = 80.2% | | | | |

*Source*: Authors 'own creation.

Lastly, we can also look at the local R-squared distribution and pinpoint the places where the relationship between GDP, education, and GERD is stronger, such as Finland or the Baltic countries (Figure 12).

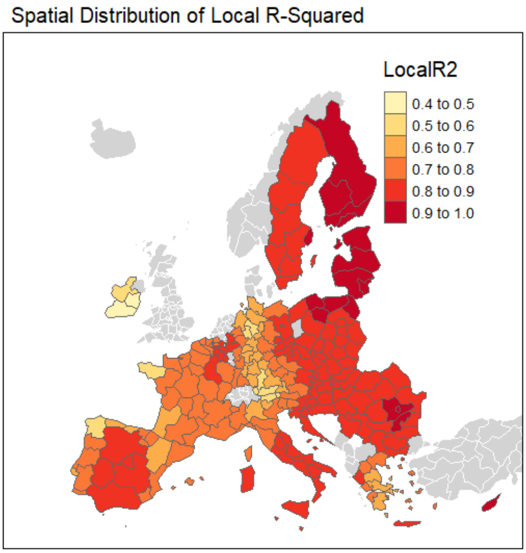


**Figure 12. Spatial Distribution of Local R-Squared, NUTS2, 2021**
*Source*: Authors 'own creation.

This distribution will confirm the previously mentioned insights, where we could observe the high levels registered by these areas for the education, economy, or science and technology dimensions. Moreover, the profile can now be completed through the validation of this relationship, since a higher level of education, corroborated with higher R&D investments, will indeed stimulate the gross domestic product and, therefore, the economy as a whole.

## 5. Conclusions

To conclude, we have seen throughout the paper that spatial analysis is extremely useful in identifying regional specificities, especially in the highly dynamic area of the European Union. Using instruments such as Moran's I spatial autocorrelation test, geographically weighted regression, or K-Means clustering, we have managed to analyse the socio-economic profile based on five dimensions, starting from the economy, labour, and science and technology, and then proceeding to demographics and education.

In terms of regional insights, we have seen how, in general, the Central or Nordic European countries tend to have improved performance for all indicators, especially in terms of GDP per capita and participation in education. At the same time, many of the South European regions will be characterised by high levels of unemployment, while in terms of age, we could see the most homogeneous distribution. Strongly linked, for spatial autocorrelation, the highest values were registered by the education and economy dimensions, a fact that was afterward transposed into the clustering distribution. Lastly, we have seen that the geographically weighted regression has been more efficient in defining the relationship between GDP per capita, GERD, and education, with very good results in terms of global determination. Therefore, in many EU regions, there will be an intense connection between the economy, education, and science, adding to the general socio-economic profile.

Thus, although succinct, this type of analysis can be useful for both researchers, as well as public institutions, since it can help design the way forward. The issues identified could be tackled through synchronised action plans, as well as national initiatives, while the potential of some regions or hubs of wealth can be extrapolated to neighbouring locations. At the same time, best-practice examples could be replicated, therefore increasing the overall socio-economic regional development.

## References

[1] Andersson, D.E., Andersson, Å.E., Hårsman, B., Yang, X. (2020), *The geography of science in 12 European countries: a NUTS2-level analysis. Scientometrics,*124, 1099-1125.

[2] Arbia, G., Le Gallo, J., Piras, G. (2008), *Does evidence on regional economic convergence depend on the estimation strategy? Outcomes from analysis of a set of NUTS2 EU regions. Spatial Economic Analysis*, 3(2), 209-224.

[3] Filenta, P., Kydros, D. (2023), *Economic and regional development through SNA: the case of the unemployment rate in NUTS 2 regions of the EU. Eastern Journal of European Studies*,14(1).

[4] Filenta, P., Kydros, D. (2023), *The Application of Social Network Analysis to Economic and Regional Development: Tertiary Educational Attainment by Sex and Nuts 2 Regions. Studies in Business and Economics*, 18(2), 124-139.

[5] Formánek, T. (2019), *GDP per capita in selected EU countries: Economic growth factors and spatio-temporal interactions examined at the NUTS2 level. Journal of International Studies,* 12(1), 119-133.

[6] Formánek, T. (2019), *Spatial econometric analysis with applications to regional macroeconomic dynamics. Habilitation thesis,* Prague University of Economics and Business.

[7] Griffith, D.A., Chun, Y. (2018), *GIS and Spatial Statistics/Econometrics: An Overview. Earth Systems and Environmental Sciences,* 1-26.

[8] Jankiewicz, M. (2023), *Regional economic growth and unemployment in the European Union–a spatio-temporal analysis at the NUTS-2 level (2013–2019). Hungarian Geographical Bulletin,* 72(2), 179-192.

[9] Moran, P.A P. (1950), *Notes on Continuous Stochastic Phenomena*. *Biometrika,* 37 (1), 17-23.

[10] Pernici, A., Stancu, S. (2023), *Performing Different Clustering Methods for Mapping the European Union Member States using Green Energy, Digitalization, and R&D Indicators: A Five-Year Comparison (2016-2020). In Science and Information Conference, Springer Nature Switzerland,* 440-461.

[11] Prada, E.M., Cimpoeru, S. (2023), *Mapping and modelling the main determinants of Migration Flows at the NUTS2 Level in European Union using Spatial Data Analysis Techniques. Management & Marketing,* 18(4), 594-607.

[12] Smit, M.J., van Leeuwen, E.S., Florax, R.J., de Groot, H.L. (2015), *Rural development funding and agricultural labour productivity: A spatial analysis of the European Union at the NUTS2 level*. *Ecological Indicators,* 59, 6-18.

[13] Żółtaszek, A., Olejnik, A. (2021), *Regional effectiveness of innovation: leaders and followers of the EU NUTS 0 and NUTS 2 regions. Innovation: The European Journal of Social Science Research,* 1-22.

[14] Wachnicka, J., Palikowska, K., Kustra, W., Kiec, M. (2021), *Spatial differentiation of road safety in Europe based on NUTS-2 regions. Accident Analysis & Prevention,* 150, 105849.