Received March 18, 2021, accepted April 18, 2021, date of publication April 23, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075175

Insights Into Energy Indicators Analytics Towards European Green Energy Transition Using Statistics and Self-Organizing Maps

CRISTIAN BUCUR[1], BOGDAN GEORGE TUDORICĂ[1], SIMONA-VASILICA OPREA[2], DUMITRU NANCU[3], AND DOREL MIHAIL DUŞMĂNESCU[1]

[1]Department of Cybernetics, Economic Informatics, Finance and Accountancy, Petroleum-Gas University of Ploiesti, 100680 Ploiesti, Romania
[2]Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania
[3]Department of Finance and Accounting, Ovidius University of Constanta, 900470 Constanta, Romania

Corresponding author: Bogdan George Tudorică (btudorica@upg-ploiesti.ro)

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI-UEFISCDI (Executive Unit for Financing Higher Education, Research, Development and Innovation), project number 462PED/28.10.2020, project code PN-III-P2-2.1-PED-2019-1198, within PNCDI III (the National Plan for Research-Development and Innovation for the 2015 - 2020 period).

ABSTRACT The more frequent meteorological anomalies and climate changes push us to consider green sustainable energy as a chance to slow down such issues. Thus, we should introspect the correlations between indicators over time and understand the underneath of their meaning. Large volumes of data regarding energy are provided by Eurostat and other official data sources that require data analytics to extract valuable insights from energy indicators and indices to better understand the dynamics towards a green energy transition of the European Union State Members (EU-SM). In this paper, we analyze several energy indicators calculated for a 12-year time span with statistics and machine learning techniques, such as an unsupervised clustering algorithm with Self-Organizing Maps (SOM). Grouping the EU-SM by energy indicators from the beginning years to the end of the analyzed interval reveals differences and similarities in their efforts, shifted trends, influencing power and tendencies towards a green energy transition. The results of our analyses can be further used to assess the efficiency of stimuli for green energy generation and improve the policymakers' strategies.

INDEX TERMS Artificial neural networks, correlation, energy management, renewable energy sources, self-organizing feature maps, statistics, machine learning.

I. INTRODUCTION

In contemporary society, green energy generation became a necessity, because of the evermore higher pressure, to conform to environmental requirements. In the fight against climate change, the modern economy must adapt and be oriented towards those sustainable energy sources which can offer not only certainty in provisioning but also a low impact on the environment. Several studies [1]–[3] proved that there is a tight correlation between economic growth and energy production and consumption, and new energy solutions, including green energy, can become powerful stimuli for rising competitivity and creating high-quality jobs in the European economy.

Green energy production is a growing industry these days. Imperatives such as continuous climate change and general population health issues related to air pollution are restrictive for more traditional means of energy generation, such as coal and gas-based generation. Other power generation means, such as nuclear power, have a public perception problem related to near-past large-scale accidents, and a higher cost problem. Many states and other state-like entities, such as the EU, have large-scale policies for supporting green energy production [1].

Energy production transition towards green energy demands a profound revision of both production and consumption economic models. Decarbonization of the economy imposes the transition to energy sources with low pollution and reduces the dependency on the primal energy sources. The development of the green energy production system asks for promoting dispersed local energy sources that can provide a major contribution to energy potential valorization. The development and diversification of green energy sources have a significant influence on achieving the decarbonization of the economy and economic growth objectives.

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

While green energy production, like any other technology, has various advantages and disadvantages, the main adoption issues are higher costs, sometimes, and social and technological inertia.

A way to contend with the said adoption issues is to provide various stimuli for green energy production, stimuli sourced by the state or some other entity. While many types of stimuli or counter-stimuli were applied in various countries to upset the green energy/traditional energy production ratio (e.g., carbon taxes, green stamps, co-generation, etc.) their efficacy and their efficiency are not entirely known. This paper is not intended to extensively list all stimuli or counter-stimuli applied by various states and entities and neither to compare some of these stimuli, but rather to extract valuable insights from large datasets and perform comprehensive analytics using statistics and ANN on datasets from official sources (such as Eurostat). For analyses, both traditional statistical tools (visual analysis and various types of correlation analysis) and machine learning analytics (unsupervised clustering with SOM and k-means) were applied.

The contribution of our study consists in proposing a data processing methodology and finding valuable insights, analyzing several energy indicators of the EU-SM related to environmental issues and green energy transition process.

The paper is further structured as follows: in the second section, a brief literature survey is presented, the third and fourth sections focus on the datasets' description and proposed methodology for data processing, whereas the fifth and sixth sections provide the results using statistics and SOM-ANN. The seventh section consists of an interesting discussion synthesizing the results, while the eighth section presents the main conclusion of this study.

II. LITERATURE REVIEW

The consumption of energy from renewable sources has been steadily rising and the share of renewable energy consumption in final energy consumption has also risen by more than double over the past decade [1]. This trend leads to numerous studies analyzing the integration of larger volumes of Renewable Energy Sources (RES) [4]–[6]. The main issues regarding RES integration consist in the necessity for a more flexible generation capacity to correct the fluctuations of wind and photovoltaic generation [7] and additional grid capacity to transmit power from generation to demand areas [8].

The EU-SM are applying various short- and long-term policies, either directly related to rising green energy production [1] or organizing the green energy generation and transport [2]. Most of them offer consistent incentives such as feed-in tariff [9] and green certificates [10]. Apart from that, substantial investment in grid reinforcement (transmission overhead lines and power substations) is already taking place to cope with large volumes of RES concentrated in specific areas [11].

There is a constant preoccupation in the recent scientific literature for energy and green energy as determinant factors for economic growth [3], [12]–[14]. In addition, the focus is also on load and its flexibility [15] to cope with RES fluctuations. Therefore, due to the progress of technologies, especially sensors and smart metering systems [16], more attention is given to Demand Response (DR) programs and services [17], [18], DR enabling technologies, and opportunities created for service providers or aggregators to access the energy markets and provide more ancillary services [19], [20].

Data for energy generation and, in particular, green energy generation is available from multiple sources, such as [21]–[23]. The availability of data is essential for research purposes, as it allows researchers to perform simulations, analyses, and studies, find solutions, and enhance the progress, policies and strategies towards green energy transition [24]. Unfortunately, most of the data providers are from developed countries, whereas the developing countries are more reluctant to disclose data.

Multiple recent studies used, for various purposes, both traditional and/or new statistical methods for analyzing the available data [1], [25]–[27]. Clustering consumers with SOM [28] and from a smart city's perspective [29] or other methods [30]–[32] is a usual approach to find out patterns and understand changes in consumers' behavior.

In this paper, we propose a data processing methodology to better analyze and understand energy indicators calculated mainly for EU-SM or ex-EU-SM. Relevant statistics and SOM are applied to emphasize the correlations and interdependencies among the nine energy indicators and underline the tendencies, previous and actual progress of each country, and its effects to accomplish more green energy. Over time, only a few studies were performed with this scope. Therefore, we can mention studies concerning energy efficiency and carbon dioxide emissions [33]–[35], energy security [36], [37], energy, transport, and environment indicators provided by Eurostat [38].

Another study proposed and reviewed several energy indicators for sustainable development, offering an analytical tool for evaluating energy generation and patterns at the country level, aiming to monitor the energy policy implementation and their effect from different perspectives: such as economic, social, and environmental perspectives [39]. Furthermore, energy indicators for sustainable development and a guideline methodology are provided in [40].

III. DATA SERIES DESCRIPTION

In our study, we collected data from Eurostat regarding the following nine energy indicators, as listed in Table 1 – the variables used in the study, with their acronyms, description, and availability data intervals are presented.

The data was extracted from Eurostat website (https://ec.europa.eu/eurostat/web/environment/environmental-protection) and DG Taxation and Customs Union, in March 2017, consisting of data regarding the above indicators for each country in EU, an Eurozone indicator, for the 19 states - Euro area and a global indicator of all 28 SM, EU-28, starting with

TABLE 1. Variable used in the study, with their acronyms and brief description.

Variable	Description of variable	Availability period
Tei	Total environmental investments - Environmental protection expenditure of the public sector by type % of GDP	2002 – 2012
Selt	Share of environmental and labor taxes in total tax revenues from taxes and social contributions - % Environmental taxes	2002 – 2014
Pec	Primary energy consumption - Million TOE	2002 – 2015
Itre	The implicit tax rate on energy - Energy taxes in Euro per TOE	2002 – 2014
Ggei	Greenhouse gas emissions intensity of energy consumption Index (2000 = 100)	2002 – 2014
Etr	Environmental tax revenues - % of GDP	1995 – 2015
Epe	Environmental protection expenditure of the public sector by type - % of GDP	2002 – 2012
Enp	Energy productivity - Euro per KGOE	2005 – 2015
End	Energy dependence % All products	2000 – 2015

TABLE 2. Dataset with descriptive statistics.

	selt	*tei*	*pec*	*itre*	*ggei*	*etr*	*epe*	*enp*	*end*
nbr.val	312.00	262.00	312.00	312.00	312.00	312.00	260.00	240.00	312.00
nbr.null	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nbr.na	0.00	50.00	0.00	0.00	0.00	0.00	52.00	72.00	0.00
Min	4.33	0.00	0.80	36.23	77.00	1.57	0.02	1.60	-11.90
Max	11.12	0.96	1722.20	420.54	122.90	4.13	1.92	10.50	104.10
range	6.79	0.96	1721.40	384.31	45.90	2.56	1.90	8.90	116.00
sum	2290.13	44.93	41395.70	49627.47	29854.10	801.89	152.01	1456.90	17534.10
median	7.11	0.13	32.35	156.53	95.70	2.51	0.55	6.55	53.80
mean	7.34	0.17	132.68	159.06	95.69	2.57	0.58	6.07	56.20
SE.mean	0.09	0.01	18.51	3.60	0.39	0.03	0.02	0.14	1.30
CI.mean.0.95	0.18	0.02	36.42	7.08	0.78	0.06	0.04	0.28	2.56
var	2.72	0.02	106916.24	4038.99	48.64	0.28	0.11	4.94	528.57
std.dev	1.65	0.14	326.98	63.55	6.97	0.53	0.33	2.22	22.99
coef.var	0.22	0.84	2.46	0.40	0.07	0.21	0.56	0.37	0.41

1995 and until 2015. Table 2 shows the descriptive statistics of the energy indicators.

Because not all 28 states were EU members in the given interval and some of the data is unavailable, we eliminated a part of the interval or some of the countries. Thus, we performed some exploratory analysis on the data to find the best interval and countries for which the comparable data is consistent. More descriptive statistics are shown in Table 3.

Therefore, the dataset narrowed to 23 countries and years between 2002 and 2014. The final list of countries used in the analysis is Austria, Belgium, Bulgaria, Croatia, Estonia, Finland, France, Germany, Hungary, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom. We also took into account EU-28 as an average indicator.

IV. PROPOSED METHODOLOGY

In this study, we propose to perform a correlation analysis to identify dependencies between the studied variables. We must go through several stages in performing the proposed statistical analysis on the chosen data. Following loosely the methodology proposed by [44], the main stages of the process are data pre-processing, algorithm testing, and validation, presentation of results.

We will conduct our analysis in R using RStudio version 1.1.423 (https://www.rstudio.com/) an open-source integrated development environment for our R version 3.4.3 (https://www.r-project.org/).

Our data were collected as independent Excel and *csv* files for each of the topics presented. The data was preformatted eliminating irrelevant information. We had to deal with missing values that had to be marked in an R recognizable format. First, we made a visual analysis of the data, and because we looked for patterns or associations between variables, we plot the data on charts with multiple axes. We present the results obtained by generating 3D charts. By plotting the data on different charts, we observe that there are some associations between points on specific axes as in Fig. 1 – only the variables which have relevant correlations were chosen and are shown in the figure. We notice that there is a clear segmentation of data regarding the relation between Enp, Selt, and Tei:

TABLE 3. Dataset with descriptive statistics.

Var	*selt*	*tei*	*pec*	*itre*	*ggei*	*etr*	*epe*	*enp*	*end*
Min.	4.33	0	0.8	36.23	77	1.57	0.02	1.6	-11.9
1st Qu.	6.098	0.07	8.275	111.73	91.47	2.208	0.34	4.1	39.48
Median	7.105	0.13	32.35	156.53	95.7	2.505	0.55	6.55	53.8
Mean	7.34	0.1715	132.679	159.06	95.69	2.57	0.5847	6.07	56.2
3rd Qu.	8.59	0.22	100	209.9	99.25	2.868	0.73	7.9	73.6
Max.	11.12	0.96	1722.2	420.54	122.9	4.13	1.92	10.5	104.1
NA's	#N/A	50	#N/A	#N/A	#N/A	#N/A	52	72	#N/A

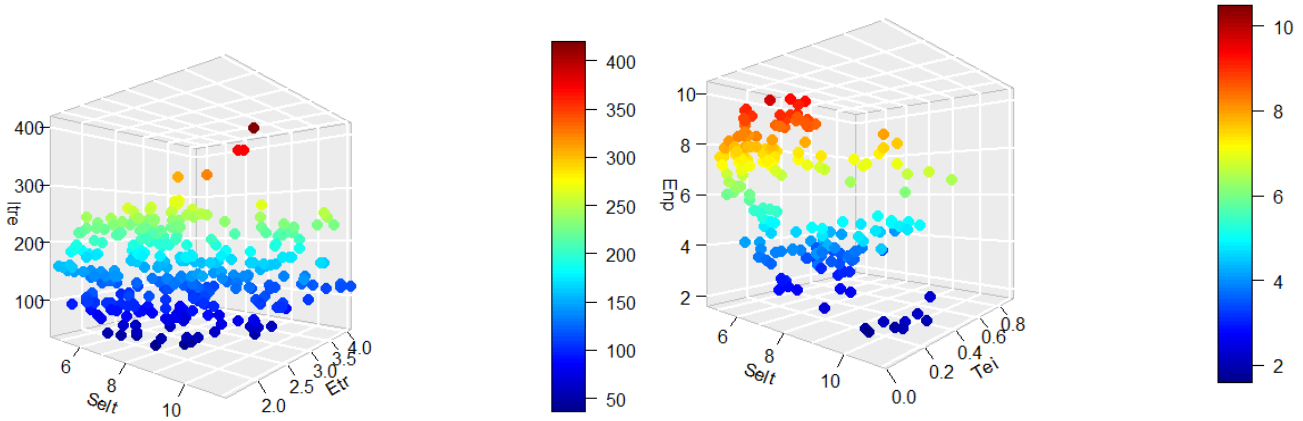


FIGURE 1. 3D Scatter Chart for Itre, Selt, and Etr/Enp, Selt, and Tei.

These observations suggest that there are some relations between our proposed variables. To better determine these relations, we performed some correlation tests.

Correlation analysis studies the relationship or associations between two or more variables. Two variables are correlated when a movement of one variable is accompanied by the movement of the other one. When performing a correlation analysis, the following stages are followed:

- Determining the existence of a relation;
- Test significance;
- Establishing the cause-and-effect relation [45].

In correlation analysis, we have dependent and independent variables and we're trying to determine that changes in independent variables determine or not changes in the dependent variables.

There are different methods for performing correlation analysis:

- Pearson correlation – which measures a linear dependence between variables is the most commonly used method [46];
- Kendall correlation – rank-based correlation;
- Spearman correlation – rank-based correlation.

Pearson correlation, which is a parametric correlation test, depending on the distribution of data, measures the linear dependence between two variables. Kendall and Spearman are nonparametric rank-based correlations [47].

If we have two vectors x and y of length n, and m_x and m_y are the means of those vectors, then the Pearson correlation formula is:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}} \tag{1}$$

The significance level of correlation p-level is determined by:

Using the correlation coefficient table for the degree of freedom (df), given n is the number of observations in vectors x and y:

$$df = n - 2 \tag{2}$$

Calculating t value and determining p-value using t distribution table for the corresponding degree of freedom:

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \tag{3}$$

We consider the correlation between x and y vectors to be significant if the p-value is $< 5\%$.

Spearman correlation formula computes the correlation between the rank of variables x and y [48]:

$$rho = \frac{\sum (x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum (x' - m_{x'})^2 \sum (y' - m_{y'})^2}} \tag{4}$$

where x' and y' are given by:

$$x' = rank\,(x)\,, \quad y' = rank\,(y) \tag{5}$$

Kendall correlation measures the correspondence between the ranking of variables x and y. The pairs of x with y observations are ordered and if x is correlated with y, they would have the same relative rank order. For each yi, we define the number of concordant pairs nc, as count the number of yj > yi, number of discordant pairs nd as number of yj < yi, where the size of x and y is n.

Therefore, Kendall correlation distance is defined as [49]:

$$tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{6}$$

Correlation analysis is performed usually between two variables. In our case, we have multiple variables per country with values for each year in the interval. In this case, we'll use a correlation matrix to investigate the correlation between our variables.

A correlation matrix R:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix} \tag{7}$$

is a matrix composed of symmetric arrays in which each of the values in the arrays is of the form r_{jk} and represents the correlation coefficient between x_j and x_k.

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} \tag{8}$$

The correlation matrix R is calculated as:

$$R = \frac{1}{n}X_S' X_S \tag{9}$$

$$X_S = CXD^{-1} \tag{10}$$

where C is a centering matrix:

$$C = I_n - n^{-1}1_n 1_n' \tag{11}$$

And D a diagonal scaling matrix:

$$D = diag(s_1, \ldots, s_p) \tag{12}$$

In this case, the X_s matrix has the form [50]:

$$X_S = \begin{pmatrix} (x_{11} - \bar{x}_1)/s_1 & (x_{12} - \bar{x}_2)/s_2 & \cdots & (x_{1p} - \bar{x}_p)/s_p \\ (x_{21} - \bar{x}_1)/s_1 & (x_{22} - \bar{x}_2)/s_2 & \cdots & (x_{2p} - \bar{x}_p)/s_p \\ (x_{31} - \bar{x}_1)/s_1 & (x_{32} - \bar{x}_2)/s_2 & \cdots & (x_{3p} - \bar{x}_p)/s_p \\ \vdots & \vdots & \ddots & \vdots \\ (x_{n1} - \bar{x}_1)/s_1 & (x_{n2} - \bar{x}_2)/s_2 & \cdots & (x_{np} - \bar{x}_p)/s_p \end{pmatrix} \tag{13}$$

Thus, for all $j \in \{1, \ldots, p\}$ in the assumption $s_j^2 > 0$, and $r_{jk} = 1$ when $j = k$, we have:

$$Cor\left(x_j, x_k\right) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} = \begin{cases} 1 & if\ j = k \\ r_{jk} & if\ j \neq k \end{cases} \tag{14}$$

In Fig. 2, the data processing flow is presented. First, the datasets on energy indicators are imported mainly from CSV files into R. Then, several preprocessing steps are required such as country and interval selection, elimination of missing values, outliers, or inconsistent values. As for data processing, we propose classical statistics and ANN methods to reduce dimensionality, show interdependencies, and group the countries with similarities. Lastly, we presented the results considering both the entire interval (12-year time span) and the early/late year of the interval to better understand the efforts, shifted trends, tendencies, and strategical influence towards green energy transition.

V. RESULTS OBTAINED FROM CORRELATION ANALYSIS

Before we proceed with the correlation analysis, we perform some preliminary tests to verify the assumptions needed in the described methods. The first assessment we have to make is the normality of the data. This is important because if the data does not have a normal distribution, then a nonparametric correlation test should be applied.

Normal distribution of data can be checked by performing visual tests or to be more precise using significance tests.

Visual normality tests performed on histogram plots, on a few of our variables could be seen in the density plots in Fig. 3. What we need for normality is the distribution to be bell-shaped.

Another visual test could be performed by using Q-Q plots (quantile-quantile plot) that draws the correlation between the data and normal distribution, adding a 45-degree reference line. In Fig. 4, we show the Q-Q plots for *Selt* and *Etr* variables and notice that most of the points fall within the reference line interval.

Because the visual inspection is not quite exact and usually unreliable, for more precision, we perform a significance test. The most used normality test is Shapiro-Wilk. The results of applying the Shapiro-Wilk test to our variables are:

- for the *Selt* variable, we obtained W = 0.97141, p-value = 7.41e-06.
- for the *Etr* variable, we obtained W = 0.97621, p-value = 4.834e-05.

Shapiro-Wilk test has a null hypothesis that the data is normally distributed and an alternative hypothesis that the data is not normally distributed. As we can see from our results, we can assume the normality of the data because the *p-value* for each of the variables is greater than the 0.05 significance level, which means the null hypothesis is validated.

Validating the assumptions of data normality required to perform the analysis, we can proceed with the correlation test. The results of the correlation test are coefficients with a value between −1 and 1 which indicate the following relation

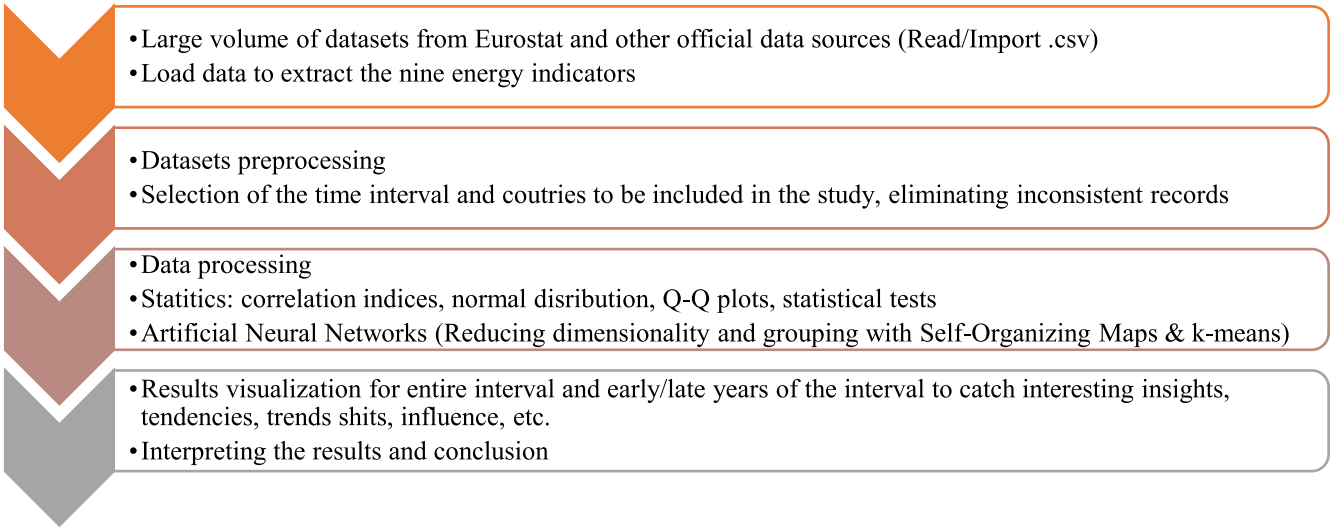


FIGURE 2. Data processing flow.

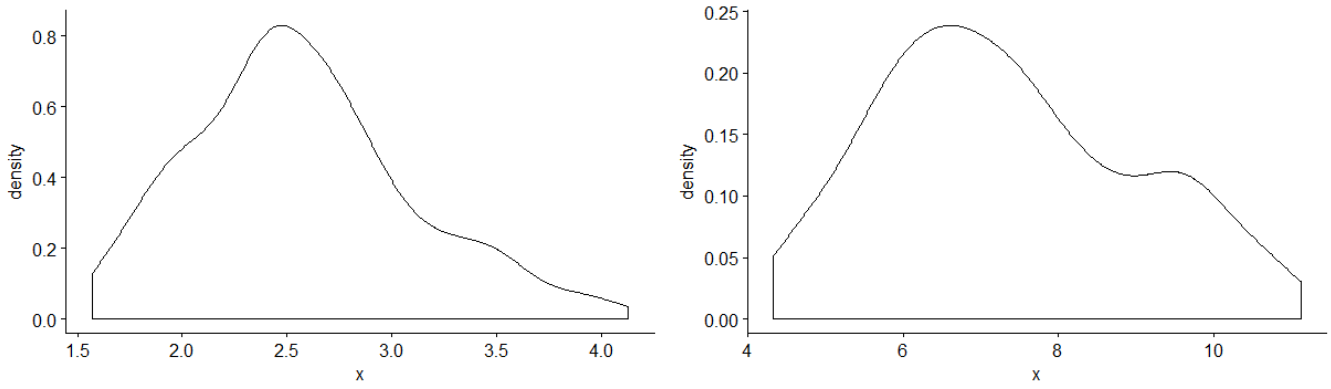


FIGURE 3. Density plots for (a) selt and (b) etr variables.

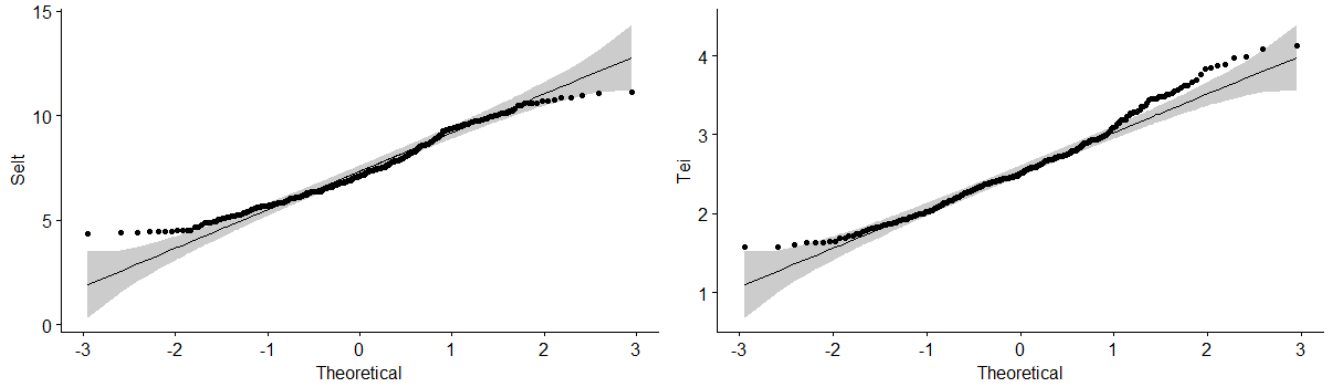


FIGURE 4. Q-Q plots for (a) selt and (b) tei.

between variables: −1 or negative indicates a negative correlation – when the first variable increases, the other decrease; 0 indicates there is no association between variables; and positive or close to 1 indicates a positive correlation – both variables increase.

We perform a Pearson's correlation test. In the results provided by the cor.test for the Pearson's correlation test, we have the following parameters:

- t is the t-test statistic value ($t = 21.861$);
- df is the degrees of freedom ($df = 310$);

TABLE 4. Number of observations used in analysis of each pair of variables.

	selt	*tei*	*pec*	*itre*	*ggei*	*etr*	*epe*	*enp*	*end*
selt	312	262	312	312	312	312	260	240	312
tei	262	262	262	262	262	262	260	193	262
pec	312	262	312	312	312	312	260	240	312
itre	312	262	312	312	312	312	260	240	312
ggei	312	262	312	312	312	312	260	240	312
etr	312	262	312	312	312	312	260	240	312
epe	260	260	260	260	260	260	260	193	260
enp	240	193	240	240	240	240	193	240	240
end	312	262	312	312	312	312	260	240	312

TABLE 5. Correlation matrix with significance levels (*p-value*).

	selt	*tei*	*pec*	*itre*	*ggei*	*etr*	*epe*	*enp*	*end*
selt	1.00	0.26	-0.22	-0.14	0.25	0.78	0.24	-0.40	-0.13
tei	0.26	1.00	-0.10	-0.12	0.19	0.14	0.51	-0.33	0.03
pec	-0.22	-0.10	1.00	0.30	-0.05	-0.12	0.05	0.24	-0.06
itre	-0.14	-0.12	0.30	1.00	-0.20	0.21	0.35	0.82	0.17
ggei	0.25	0.19	-0.05	-0.20	1.00	0.16	0.09	-0.20	0.11
etr	0.78	0.14	-0.12	0.21	0.16	1.00	0.26	-0.04	-0.04
epe	0.24	0.51	0.05	0.35	0.09	0.26	1.00	0.22	0.28
enp	-0.40	-0.33	0.24	0.82	-0.20	-0.04	0.22	1.00	0.43
end	-0.13	0.03	-0.06	0.17	0.11	-0.04	0.28	0.43	1.00

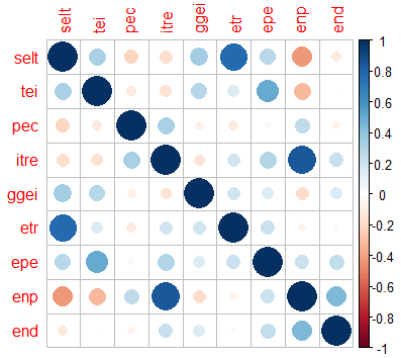


FIGURE 5. The heatmap corresponding to the Person correlation matrix as in Table 7.

- *p-value* is the significance level of the *t-test* (*p-value* < 2.2e-16);
- 95 percent confidence interval is the confidence interval of the correlation coefficient at 95% (conf.int = [0.7309913, 0.8190254]);
- the sample estimate is the correlation coefficient (Cor.coeff = 0.7788149).

Because the *p-value* is less than the significance level of α =0.05, this means that our *selt* and *etr* variables are correlated with a correlation coefficient of 0.7788149 and a *p-value* < 2.2e-16.

We now continue by analyzing the correlation between all our variables in the final 24 countries dataset with time series between 2002 and 2014. In the previous steps, we marked the missing data from our dataset. We observed that the data still contains missing values in some cases. There are several ways to deal with missing data [51]. Casewise deletion handles this matter by removing all cases that contain NA values anywhere. Another strategy might be completing the data with the mean of other values or with randomly chosen samples of other values. These strategies would preserve the mean of the variable, or both mean and variance. (Leeper) In our case, those strategies would have altered the statistical properties of our data to a greater extent, so we treat this by *R* case-wise deletion.

Table 4 shows the (*n*) elements outputted by the analysis, which represent the matrix of the number of observations taken into consideration for each pair of variables:

The correlation matrix used to analyze the correlation between our multiple variables [23] along with other descriptive results are presented in the following tables. In our case, we used a Pearson method for performing correlations. The resulting data after performing the Pearson matrix correlation are presented in Table 5 (also shown as heatmap in Fig. 5).

In Fig. 6, we see a plot of correlation results in which on the diagonal we have the distribution of each variable, the scatter plot with fitted lines of the bivariate below the diagonal, and the correlation coefficients with significance level on the top of diagonal. The significance level is symbolized as stars and is associated with a *p-value*.

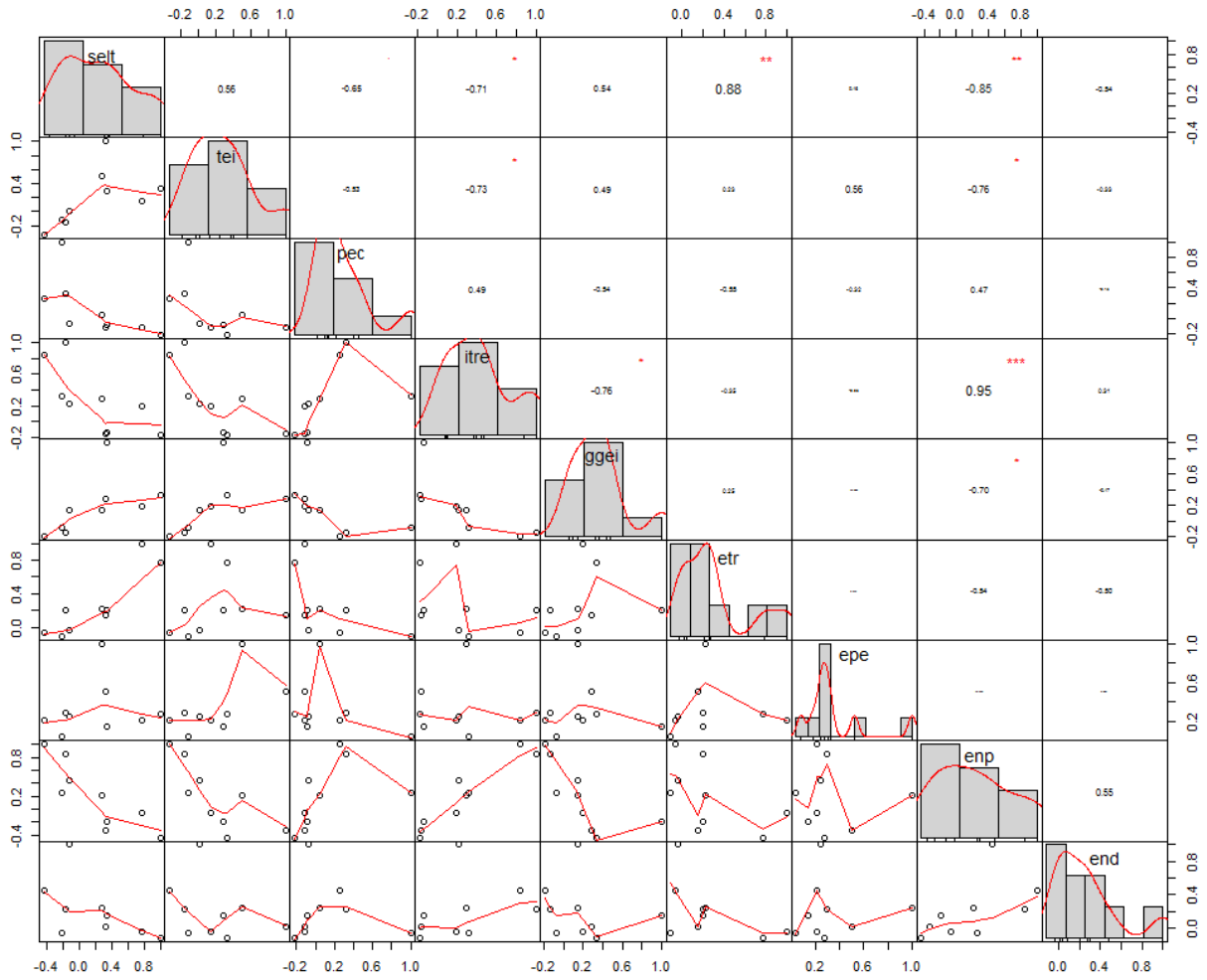


FIGURE 6. Correlation matrix plot.

We obtained the correlation coefficients and the significance level for all combinations of column pairs in the analyzed data frame. We are trying now to identify the strong correlations between our variables. First, we rearrange the order of the correlation results for better visibility (Fig. 7).

Table 6 presents the *p-values*, the significance levels for each of the correlations calculated:

The next step would be to have a visual representation of the correlation results. This could be done in several ways like correlogram, scatter plot, and heatmap.

The visual representation of the correlation results is presented in Fig. 8. The correlogram diagram in the left figure presents the top part of the correlations in colors according to negative and positive values – red for negative values and blue for positive ones, and the strength of correlation by the size of points. The right figure is a heatmap representation of correlations in which the positivity or negativity of coefficients are also represented by red and blue and strength by the intensity of color tones.

The heatmap representations help us to quickly identify the nature of correlations between variables and patterns. For more detailed values and comparations, we would then use the correlogram and values from Fig. 8.

We observe we could split the interactions into 3 categories:

- positive correlations – in our case this is the largest category, we have more positive correlations with different levels of correlation, a few strong ones, and more of lower ones. The strongest correlations are between *enp - itre* with a value of 0.95, *etr - selt* with a value of 0.88, *epe - tei* with a value of 0.56, *selt - tei* with a value of 0.56, *selt - ggei* with a value of 0.54. There is also a correlation between *ggei – tei*, *itre - pec* with a value of 0.49, *enp – pec* with a value of 0.47;
- negative correlations – we have fewer negative correlations, but these are stronger;
- the strongest inverse correlated variables are *selt - enp* with a value of −0.85, *enp – tei*, *itre - ggei* with a value of −0.76, *itre - tei* with a value of −0.73, *selt - itre* with a value of −0.71, *enp - ggei* with a value of −0.7. Furthermore, the strong anticorrelated are *pec - selt* with the value of −0.65, *pec – etr* with a value of −0.55, *enp*

```
       tei    epe  ggei  selt   etr   end  pec itre enp
tei      1
epe   0.56      1
ggei  0.49 -0.046     1
selt  0.56   0.16  0.54     1
etr   0.23  0.026  0.25  0.88     1
end  -0.33  0.042 -0.17 -0.54  -0.5     1
pec  -0.53  -0.32 -0.54 -0.65 -0.55 -0.13    1
itre -0.73  -0.05 -0.76 -0.71 -0.35  0.31 0.49    1
enp  -0.76 -0.083  -0.7 -0.85 -0.54  0.55 0.47 0.95   1

$p
       tei  epe  ggei   selt  etr  end  pec  itre enp
tei      0
epe   0.12    0
ggei  0.18 0.91     0
selt  0.12 0.69  0.13      0
etr   0.55 0.95  0.51 0.0017    0
end   0.39 0.92  0.66   0.13 0.17    0
pec   0.14  0.4  0.13  0.059 0.13 0.74    0
itre 0.025  0.9 0.018  0.031 0.36 0.42 0.18     0
enp  0.017 0.83 0.035 0.0039 0.13 0.12  0.2 1e-04    0

$sym
     tei epe ggei selt etr end pec itre enp
tei  1
epe  .   1
ggei .       1
selt .       .    1
etr                +    1
end  .             .    .   1
pec  .   .   .     ,    .       1
itre ,       ,     ,    .   .   .   1
enp  ,       ,     +    .   .   .   *    1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

FIGURE 7. Bottom correlation with p-value and significance levels.

TABLE 6. *p-values* corresponding to the significance levels of correlations.

	selt	*tei*	*pec*	*itre*	*ggei*	*etr*	*epe*	*enp*	*end*
selt		0.0000	0.0001	0.0109	0.0000	0.0000	0.0001	0.0000	0.0188
tei	0.0000		0.0970	0.0593	0.0019	0.0259	0.0000	0.0000	0.6035
pec	0.0001	0.0970		0.0000	0.3793	0.0354	0.4610	0.0001	0.2639
itre	0.0109	0.0593	0.0000		0.0004	0.0002	0.0000	0.0000	0.0031
ggei	0.0000	0.0019	0.3793	0.0004		0.0045	0.1410	0.0017	0.0495
etr	0.0000	0.0259	0.0354	0.0002	0.0045		0.0000	0.5499	0.4721
epe	0.0001	0.0000	0.4610	0.0000	0.1410	0.0000		0.0026	0.0000
enp	0.0000	0.0000	0.0001	0.0000	0.0017	0.5499	0.0026		0.0000
end	0.0188	0.6035	0.2639	0.0031	0.0495	0.4721	0.0000	0.0000	

- *etr*, *pec* - *ggei* with the value of −0.54, *pec* - *tei* with the value of −0.53;
- no correlation – we can conclude that there is no correlation between those pairs of variables. These pairs with little or no correlation are: *pec* - *enp*, *pec* – *itre*, *end* - *itre*, *end* - *epe*, *end* - *ggei*, *enp* – *epe*, *itre* - *epe*, *itre* - *etr*, *tei* - *etr*, *tei* - *selt*, *tei* - *ggei*, *epe* - *etr*, *epe* - *selt*, *epe* - *ggei*, *ggei* - *etr*, *ggei* - *selt*. They all have a correlation index with absolute values below 0.3.

The next step would be to determine the most significant coefficients and to mark and eliminate the insignificant ones according to the p-value significance level. Below, in Fig. 9, we present the initial representation of the top correlogram for our data and in the right zone the final representation of significant correlations. For the right matrix, in the top, we have the strength of correlation represented by the dimension and color of squares, and in the bottom symmetric part the actual values also with the same color code.

For our analyzed data frame with variables for all years between 2002 – 2014, we obtained the following results:

- we have a strong positive correlation between *itre* and *enp* variables with a significant 0.84 correlation coefficient;

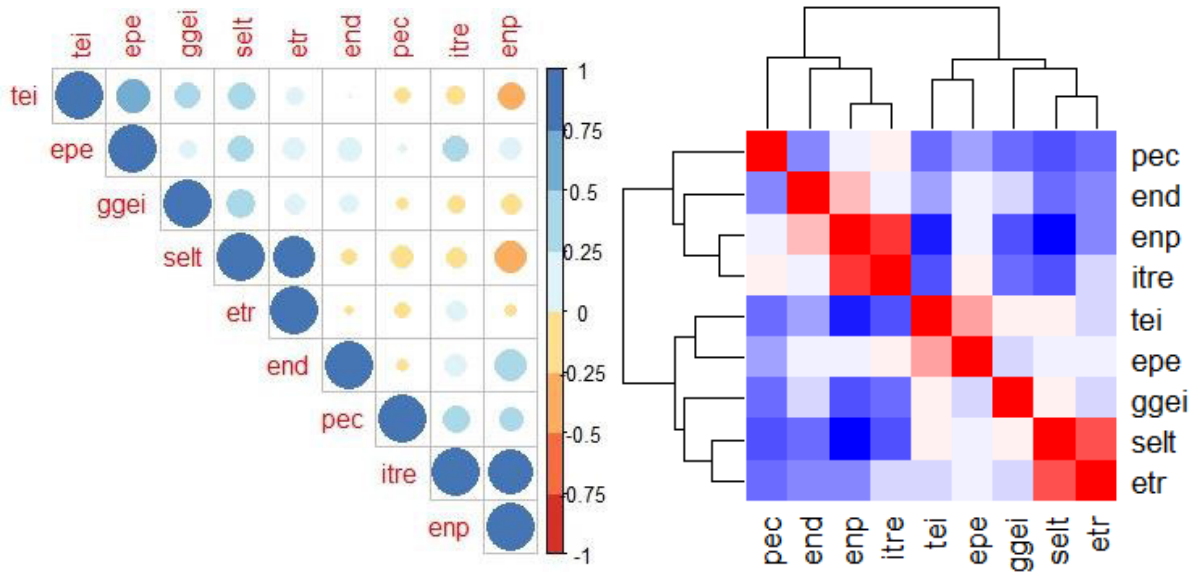


FIGURE 8. Visual representation of the correlation matrix as (a) correlogram and (b) heatmap.

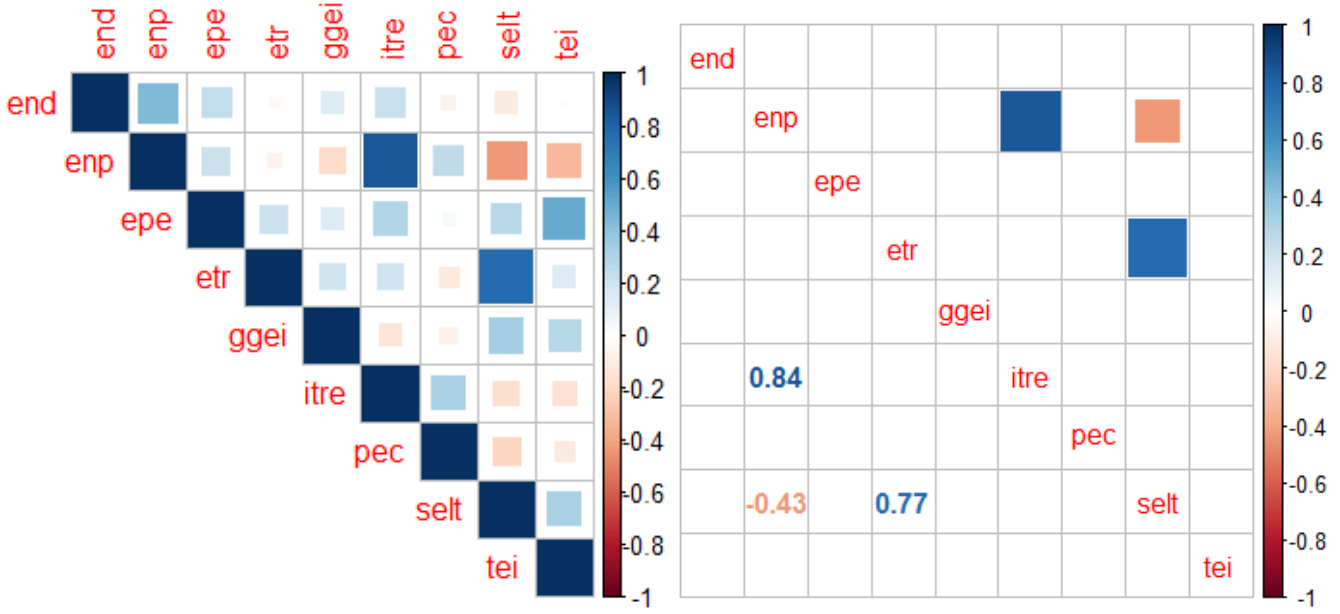


FIGURE 9. Correlation matrix representation with significance for years 2002 – 2014 interval.

- we have a strong positive correlation between *selt* and *etr* variables with a significant 0.77 correlation coefficient;
- we also have a negative correlation between *enp* and *selt* with a significant – 0.43 negative correlation coefficient.

Thus, taking into consideration the entire analyzed interval 2002 - 2014 and all variables, we can conclude that there is a strong positive correlation between the implicit tax rate on energy – the energy tax per TOE and energy productivity - as the amount per KGOE. Furthermore, there is a strong positive correlation between the shares of environmental and labor taxes in total tax revenues from taxes and social contributions - % of environmental taxes and environmental tax revenues - % of GDP.

There is also an inverse correlation between energy productivity - amount per KGOE and shares of environmental and labor taxes in total tax revenues from taxes and social contributions - % of environmental taxes.

Besides, we analyzed if this tendency is the same for each period of the interval or if we obtain a different result for the distinctive interval, or if the correlation between variables has a dynamic over the years. According to this, we decided to

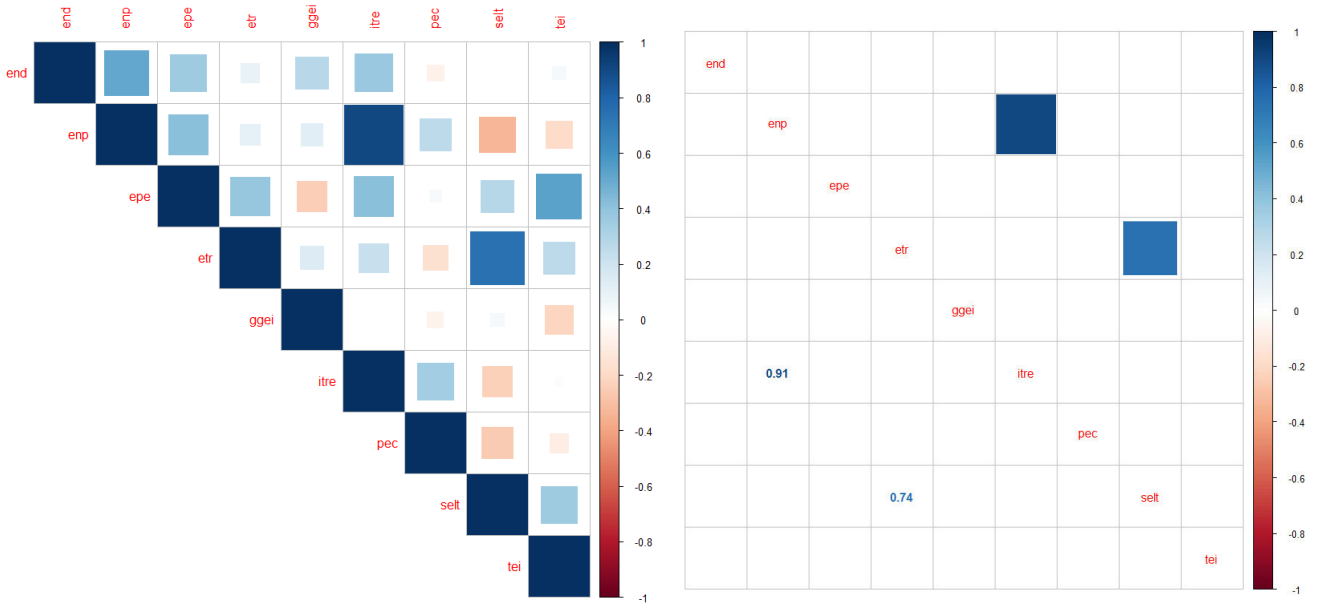


FIGURE 10. Correlation matrix representation with significance for the year 2005 – at the beginning of the interval.

choose a year at the beginning of the interval and one at the end of the interval to calculate the correlation matrix of our studied variables.

First, we made a correlation analysis for the start of the interval. For this, we choose a year from the beginning and the first year for all studied countries and the minimum of missing data was 2005. Thus, in Fig. 10, the visual representation of the top correlogram and the correlation matrix is shown.

For the year 2005, from the beginning of all data intervals we obtained the following results:

- we have a strong positive correlation between *itre* and *enp* variables with a significant 0.91 correlation coefficient – the correlation is even stronger at the beginning interval between these variables;
- we have a strong positive correlation between *selt* and *etr* variables with a significant 0.74 correlation coefficient – the correlation remains in the same parameters;
- we don't have any significant negative correlation between variables at the beginning of the interval.

Thus, at the start of the data frame in our study, the strong positive correlation between the implicit tax rate on energy – the energy tax per TOE and energy productivity - as the amount per KGOE is even more evident. The positive correlation between the shares of environmental and labor taxes in total tax revenues from taxes and social contributions - % of environmental tax and environmental tax revenues - % of GDP remains in the same limit as for the entire interval.

The inverse correlation between energy productivity and the share of environmental and labor taxes in total tax revenues from taxes and social contributions is not present at the beginning interval.

We also made the correlation analysis for the end of the interval, and the year with all necessary data that fitted the necessary criteria was 2011. The visual representation of the obtained results – the top correlogram and significance matrix are presented in Fig. 11.

For the year 2011 interval, we obtained the following results:

- we have a strong positive correlation between *itre* and *enp* variables with a significant 0.82 correlation coefficient – the correlation is in concordance with the result on the entire interval;
- the correlation between *selt* and *etr* variables is not significant in the analyzed year;
- we have several significant negative correlations between variables in this year of which the strongest are between:
  - variables *enp* and *tei* with a −0.46 negative correlation coefficient;
  - variables *selt* and *enp* with a −0.44 negative correlation coefficient;
  - variables *ggei* and *enp* with a −0.33 negative correlation coefficient;
  - variables *itre* and *tei* with a −0.31 negative correlation coefficient.

Confronting the findings for this year, we obtain some interesting results that differ significantly from the one obtained at the start of the interval and differ from the ones in the entire interval.

The strong positive correlation between the implicit tax rate on energy – the energy tax per TOE and energy produc-

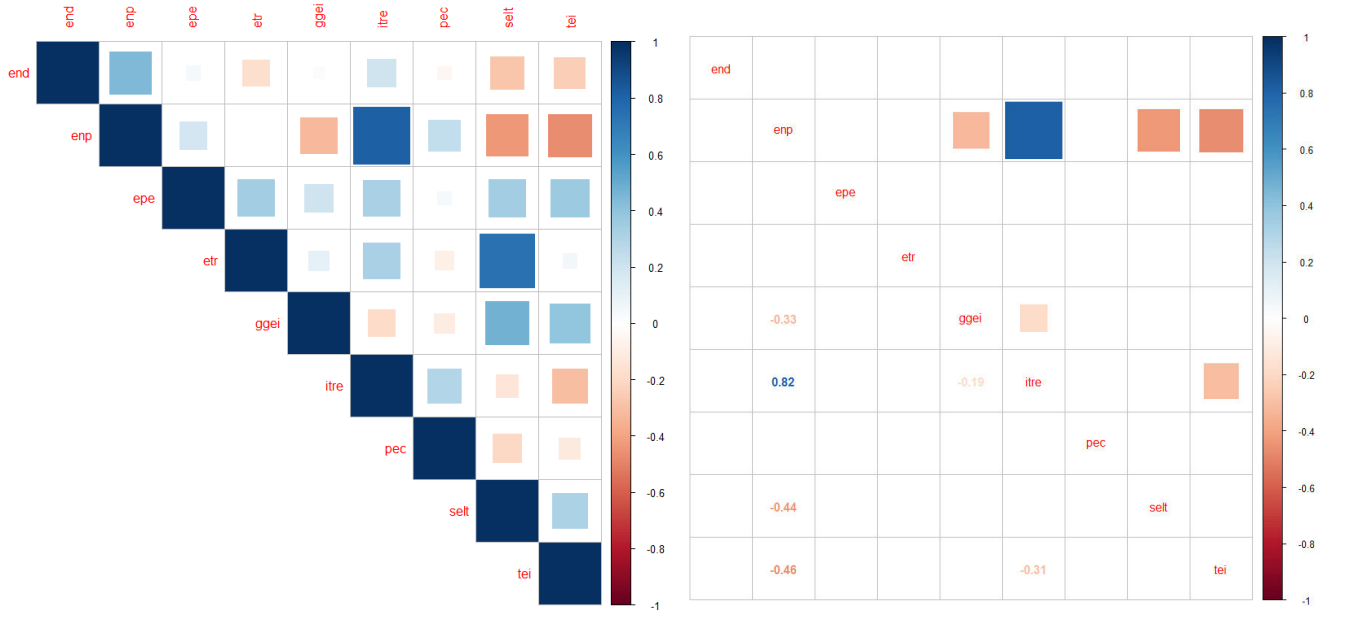


FIGURE 11. Correlation matrix representation with significance for the year 2011 – at the end of the interval.

tivity - as the amount per KGOE remains in the same interval as at the start of the interval a little more raised than for the entire interval. The positive correlation between the shares of environmental and labor taxes in total tax revenues from taxes and social contributions - % of environmental tax and environmental tax revenues - % of GDP which is present at the beginning of interval is not present at all at the end of the interval.

The inverse correlation between the energy productivity and shares of environmental and labor taxes in total tax revenues from taxes and social contributions, which is not present in the beginning interval is visible in the end one in the same limits as for the entire interval.

Besides these correlations analyzed in the entire interval, we also have some new inversed correlated variables in the end interval.

There is a more significant inverse correlation (-0.46) between energy productivity - amount per KGOE and total environmental investments - Environmental protection expenditure of the public sector by type % of GDP.

There is a smaller inverse correlation between greenhouse gas emission intensity of energy consumption Index ($2000 = 100$) and energy productivity - amount per KGOE (anticorrelation index of -0.33).

There is also an inverse correlation between the implicit tax rate on energy - energy taxes in Euro per TOE and the total environmental investment - environmental protection expenditure of the public sector by type % of GDP.

We can conclude from this that at the begging of the interval, the measures taken by states to reduce environmental pollution were visible and as the economy progresses and environmental measures are taken, we have a less pollutant economy.

VI. CLUSTERING THE ENERGY INDICATORS

Machine learning is one of the novel approaches used for analyzing energy consumption and energy performance [27], [25], and, in particular, green energy-related issues [26]. A particular flavor of machine learning is the unsupervised clustering using Artificial Neural Networks (ANN).

In our study, we analyzed a method for applying unsupervised dimensionality reduction to data. For this, we explored Self-Organizing Maps (SOM). SOMs are different from usual neural networks as they use competitive learning rather than error-correction learning (like backpropagation and gradient descent). SOM uses a neighborhood function for preserving the topological properties of the input space [53].

We converted our data frame to a matrix, eliminated the null data, and then centered and scaled all variables. We then created a grid and tried several sizes and hexagonal and circular topologies. We determined the best size in our case (8 by 8, for our case, with a hexagonal topology), and then trained the SOM specifying iterations, learning rate, and neighborhood.

We used some specific visualizations for determining the quality of our trained SOM and the relation between variables. To decide if the number of training iterations is enough, a plateau of the mean distance to the closest unit in the training progress chart must be obtained. We needed to check the counts' plot to determine the quality of the map. Ideally, we should have a minimum of 5 samples per node and a minimal number of empty nodes, as shown in Fig. 12.

The code plot (Fig. 13) presents the node weight vectors made of normalized values of the original variables used to generate SOM. Using this chart, we can detect patterns in the data distribution.

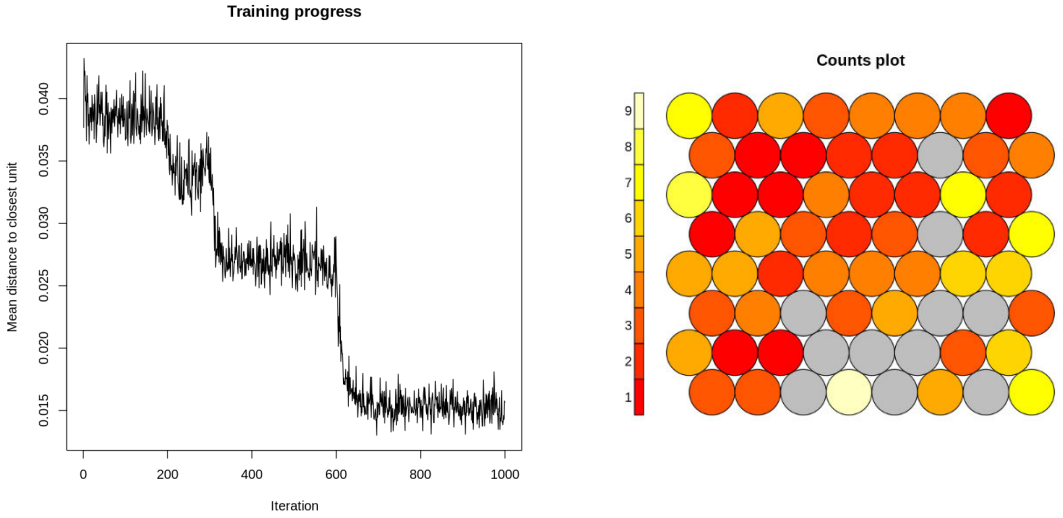


FIGURE 12. Training Progress and Counts Plot.

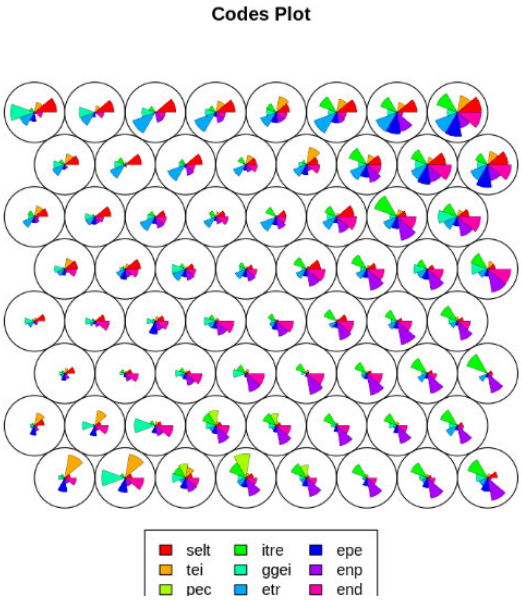


FIGURE 13. Codes/Weight vectors.

One of the most important visualizations for SOMs is the heatmaps for each variable. For high dimensional data, this is where we detect relations between variables. We can see a direct relation between ***itre*** and ***enp*** variables. There is also an inverse relationship between ***selt*** and ***end*** variables or between ***tei*** and ***end***. A clear inverse pattern we can also see between ***selt*** and ***enp*** variables (Fig. 14).

We proceeded with determining the group of samples with similar metrics. For this, we performed clustering on SOM nodes [54]. We then try to determine the optimum number of clusters. For this, we used several methods such as the elbow method, average silhouette method, Hubert index, and D index from the NbClust R package (significant peak in the second plot, corresponding to a significant increase of values in the first one) [55]. (as in Fig. 15):

Using Hierarchical clustering, we analyzed the maps obtained. Ideally, we're looking to determine clusters contiguous on the map surface (Fig. 16) [56]:

Using the k-means method, we created six clusters, and plotting the results we obtained the cluster map shown below – Fig. 17:

By looking at the cluster map (Fig. 19), we can see the main characteristics for each of the 6 clusters. Each has a specific combination of dominant variables from the input data. Fig. 18 is a table of samples, each sample associated with the corresponding cluster:

Therefore, SOM proved to be a useful tool for creating segmentation profiles on data that are intuitive and visually recognizable [57]. They help reduce high dimensional data to 2D maps, thus making it easy to detect specific patterns.

VII. DISCUSSION

This paper aims to provide valuable insights using classic statistical analysis methods vs. newer generation ANN methods (e.g., SOM) in establishing, based on existing statistical data, what stimuli are most relevant when attempting to motivate energy investors in switching towards green energy production.

By implementing the classic statistical analysis methods on Eurostat data, we gained the following knowledge:

- We found a strong positive correlation between *the implicit tax rate on energy* and *energy productivity*. Our interpretation of this finding is that higher tax rates force energy producers to increase productivity, maybe to be able to generate energy still cheap enough to be competitive on the energy market.
- We also found a strong positive correlation between *the shares of environmental and labor taxes in total*

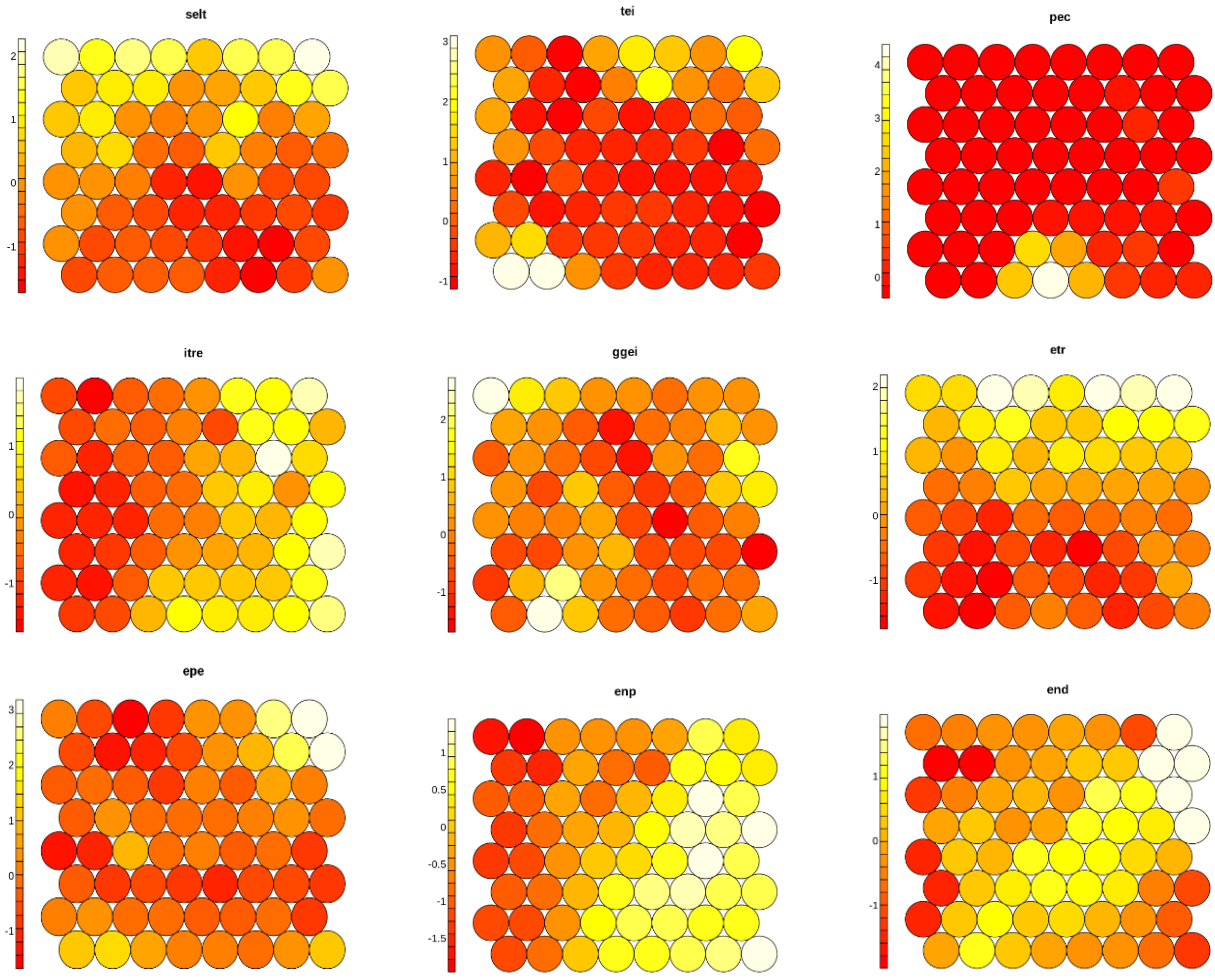


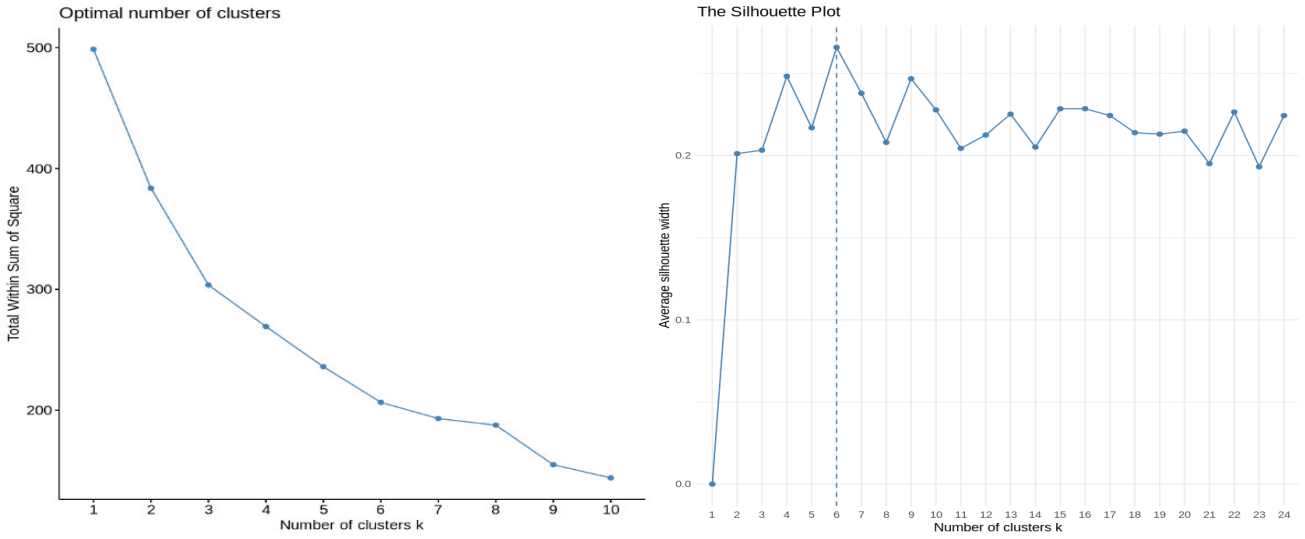
FIGURE 14. SOM Heatmaps for each input variable.

tax revenues from taxes and social contributions and *environmental tax revenues*. While this is one of the strongest correlations we found, it's also expected, so we gained no new knowledge from this.

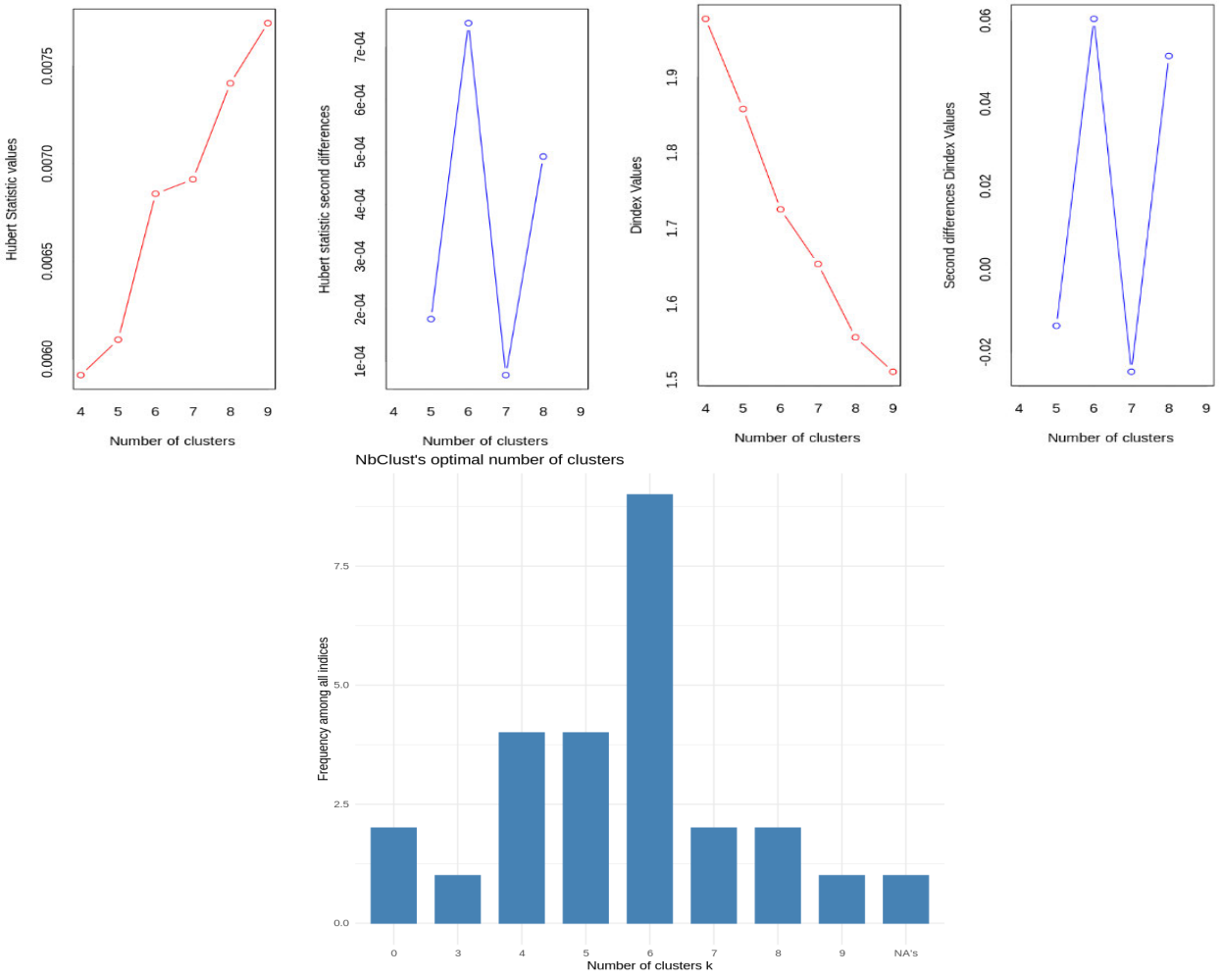
- There is also an inverse correlation between *energy productivity* and *shares of environmental and labor taxes in total tax revenues from taxes and social contributions*. This is an interesting finding, as we interpreted in this way: higher energy productivity reduces the environmental and labor taxes, and low environmental and labor taxes are usually assimilated into a more green-oriented energy production system. Thus, we gained the knowledge that low environmental and labor taxes can indicate both a green-oriented energy production system or a high productivity one.

By implementing machine learning: unsupervised clustering with ANN, we extracted the following advantages:

- We were able to circumvent some of the shortcomings of the classic statistical analysis methods such as the need for weighting, transformation, and ranking, by leveraging an unsupervised data clustering technique known as Self-Organized Maps to extract the inherent patterns present in the analyzed data.
- In contrast with traditional statistical methods, SOM does not use arbitrary rankings or data transformations and instead constructs a data-driven representation of the underlying similarity between the respective environmental performance of countries.
- The correlations we found with classical statistical methods were also confirmed by SOM.
- With SOM, we were also able to cluster the EU-SM into six clusters with significant differences regarding the energy indicators.
- The optimal number of clusters was determined with the elbow method, ensuring this way that a maximal amount of information from the analyzed data was taken into consideration when running the cluster analysis.
- The first cluster contains Italy, Luxembourg, Malta, and Portugal (in the early years of the analyzed period).



(a)



(b)

FIGURE 15. (a) Identification of the optimal number of clusters using the elbow method and the average silhouette method. (b) Identification of the optimal number of clusters using the Hubert index and the D index from the NbClust R package.

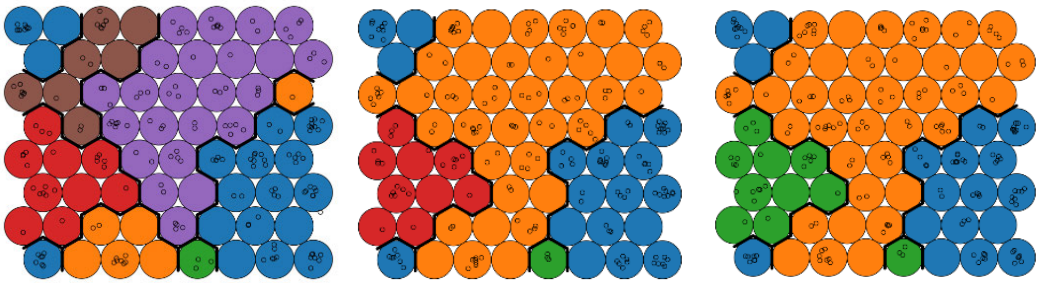


FIGURE 16. Various attempts to find clusters that are contiguous on the map surface.

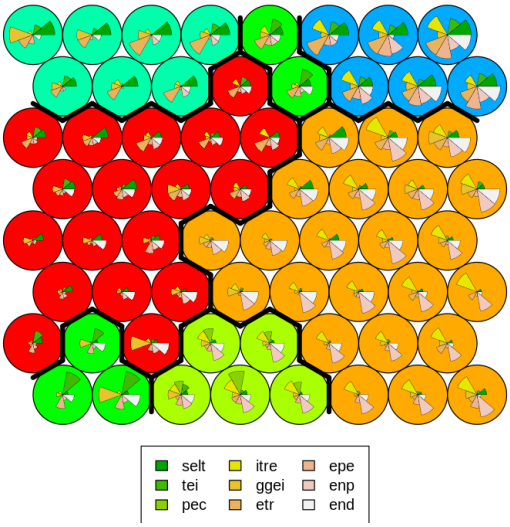


FIGURE 17. SOM Cluster Map.

- The second cluster contains Bulgaria, Croatia, Estonia (in the later years of the analyzed period), and Poland (in the early years of the analyzed period).
- The third cluster contains Estonia (in the early years of the analyzed period), Finland, Hungary (in the later years of the analyzed period), Latvia, Lithuania (in the first two years of the analyzed period), Poland (in the later years of the analyzed period), Romania and Slovakia.
- The fourth cluster contains Austria, Belgium, Lithuania (in the last year of the analyzed period), Portugal (in the later years of the analyzed period), and Spain.
- The fifth cluster contains France, Germany, Sweden, and the United Kingdom. As an interesting fact, the average of EU-28 was also located in the fifth cluster, confirming the fact that the three largest industrialized countries in EU-28 are the ones that establish the economic trends of the EU.
- The sixth cluster contains Bulgaria (in the last year of the analyzed period), Hungary (in the early years of the analyzed period), Lithuania (in the middle part of the analyzed period), Malta, Netherlands, and Slovenia.

Bulgaria, Estonia, Hungary, Lithuania, Poland, and Portugal shifted trends during the analyzed period (Lithuania twice). Further analysis may correlate these shifts with some political and/or policy changes in these countries.

	selt	tei	pec	itre	ggei	etr	epe	enp	end	cluster
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
94	8.90	0.12	4.3	101.03793	91.1	2.46	0.68	4.3	59.9	1
95	8.59	0.13	4.4	105.53465	84.2	2.44	0.73	4.3	56.4	1
96	7.85	0.29	8.0	79.19362	96.7	2.29	0.48	3.0	56.8	1
97	5.97	0.41	7.8	80.55885	99.6	1.80	0.75	3.3	62.0	4
98	5.82	0.55	8.0	89.12200	92.9	1.75	0.89	3.4	61.2	4
99	5.34	0.51	8.2	98.15694	91.9	1.63	0.85	3.5	57.8	4
100	6.68	0.80	7.8	111.62353	91.1	2.02	1.20	3.3	49.9	4
101	6.46	0.96	6.1	103.35287	122.9	1.83	1.36	4.1	81.8	4
102	6.20	0.56	5.9	105.30465	110.7	1.69	0.94	4.2	81.7	4
103	6.09	0.50	5.9	106.52576	109.7	1.64	0.90	4.3	80.3	4
104	6.04	0.12	5.7	113.52793	110.4	1.64	0.56	4.8	78.3	1
105	7.84	0.20	4.8	192.30587	108.0	3.00	0.72	7.4	97.4	2
106	7.36	0.18	4.7	193.96970	108.0	2.67	0.63	7.9	98.2	2
107	7.10	0.21	4.6	204.39059	104.4	2.59	0.52	8.7	96.7	2
108	7.05	0.24	4.6	209.70669	103.5	2.62	0.51	8.6	97.5	2
109	6.58	0.29	4.3	210.49365	105.2	2.57	0.68	8.7	97.5	2
110	6.38	0.19	4.6	205.10285	104.2	2.40	0.55	8.6	97.1	2
111	6.36	0.19	4.5	221.27972	104.5	2.38	0.53	8.9	97.3	2
112	6.15	0.19	4.4	231.75028	104.6	2.37	0.56	9.1	97.5	2
113	5.65	0.19	4.3	224.85973	101.9	2.17	0.57	9.8	97.1	2
114	9.74	0.38	1.0	162.71510	93.5	3.08	1.42	6.1	100.0	6
115	9.99	0.34	0.9	175.36085	100.9	3.19	1.54	6.7	100.0	6
116	10.85	0.45	1.0	250.77934	96.9	3.57	1.66	6.5	100.0	6

FIGURE 18. Sample input data by cluster.

VIII. CONCLUSION

Various statistical analysis methods were applied to several economic indicators extracted from Eurostat datasets that lead to interesting insights revealed in the previous discussion section. Besides, machine learning unsupervised algorithms, such as Self-Organizing Maps, while being able to confirm the results already obtained with the classic correlation methods, are also capable of extracting new insights for the available data, such as identifying similarities and trends with visual analysis of energy-related data that can be applied for a consistent search of patterns or associations between variables.

While no correlation analysis technique can be used alone for processing energy-related data, various combinations of them can lead to valuable insights. Thus, SOM are a useful tool for reducing high dimensional data to 2D maps, thus making it easy to detect specific patterns. Hence, we were able to group the EU-SM into six clusters with strong similarities regarding the energy indicators. As such, we concluded that EU-SM, while having common principles towards green energy adoption, do not have a common policy too, as there

are significant wide differences between the green energy adoption stimuli usage. It was not exactly a surprise the fact that the countries contained by the six clusters with different green energy policies are also correlated by other geographical and economic factors. Our conclusion matches the initial premise: no single statistical analysis method can extract all information related to renewable energy production from a batch of datasets and a combination of both traditional and machine learning techniques should be the preferred approach for a research study in this field.

While green energy production is yet in its developing stages and requires various stimuli, the relevance of each such stimulus has to be analyzed using various methods to obtain a consistent conclusion. Moreover, the proposed analyses should be yearly repeated as they can catch the changes and represent significant indicators for policymakers.

REFERENCES

[1] A. Mehedintu, M. Sterpu, and G. Soava, "Estimation and forecasts for the share of renewable energy consumption in final energy consumption by 2020 in the European union," *Sustainability*, vol. 10, no. 5, p. 1515, May 2018.

[2] J. Lowitzsch, C. E. Hoicka, and F. J. van Tulder, "Renewable energy communities under the 2019 European clean energy package–governance model for the energy clusters of the future?" *Renew. Sustain. Energy Rev.*, vol. 122, Apr. 2020, Art. no. 109489.

[3] J. V. Andrei, M. Mieila, and M. Panait, "The impact and determinants of the energy paradigm on economic growth in European union," *PLoS ONE*, vol. 12, no. 3, Mar. 2017, Art. no. e0173282.

[4] J. M. Morales, A. J. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Integrating Renewables in Electricity Markets*. New York, NY, USA: Springer, 2014.

[5] *International Renewable Energy Agency—IRENA, FUTURE OF SOLAR PHOTOVOLTAIC Deployment, Investment, Technology, Grid Integration and Socio-Economic Aspects*, Int. Renew. Energy Agency, Abu Dhabi, United Arab Emirates, 2019.

[6] J. M. Carrasco, L. G. Franquelo, J. T. Bialasiewicz, E. Galvan, R. C. PortilloGuisado, M. A. M. Prats, J. I. Leon, and N. Moreno-Alfonso, "Power-electronic systems for the grid integration of renewable energy sources: A survey," *IEEE Trans. Ind. Electron.*, vol. 53, no. 4, pp. 1002–1016, Jun. 2006, doi: 10.1109/TIE.2006.878356.

[7] S.-V. Oprea, A. Bára, and G. Majstroviá, "Aspects referring wind energy integration from the power system point of view in the region of southeast Europe. Study case of romania," *Energies*, vol. 11, no. 1, p. 251, Jan. 2018, doi: 10.3390/en11010251.

[8] T. Adefarati and R. C. Bansal, "Integration of renewable distributed generators into the distribution system: A review," *IET Renew. Power Gener.*, vol. 10, no. 7, pp. 873–884, Aug. 2016, doi: 10.1049/iet-rpg.2015.0378.

[9] M. T. Costa-Campi and E. Trujillo-Baute, "Retail price effects of feed-in tariff regulation," *Energy Econ.*, vol. 51, pp. 157–165, Sep. 2015, doi: 10.1016/j.eneco.2015.06.002.

[10] S.-V. Oprea and A. Bára, "Analyses of wind and photovoltaic energy integration from the promoting scheme point of view: Study case of romania," *Energies*, vol. 10, no. 12, p. 2101, Dec. 2017, doi: 10.3390/en10122101.

[11] M. Fársch, S. Hagspiel, C. Jägemann, S. Nagl, D. Lindenberger, and E. Tröster, "The role of grid extensions in a cost-efficient transformation of the European electricity system until 2050," *Appl. Energy*, vol. 104, pp. 642–652, Apr. 2013, doi: 10.1016/j.apenergy.2012.11.050.

[12] D. Dusmanescu, J. Andrei, and J. Subic, "Scenario for implementation of renewable energy sources in romania," *Procedia Econ. Finance*, vol. 8, pp. 300–305, Dec. 2014.

[13] G. H. Popescu, J. V. Andrei, E. Nica, M. Mieilä, and M. Panait, "Analysis on the impact of investments, energy use and domestic material consumption in changing the Romanian economic paradigm," *Technol. Econ. Develop. Econ.*, vol. 25, no. 1, pp. 59–81, Jan. 2019.

[14] G. H. Popescu, M. Mieila, E. Nica, and J. V. Andrei, "The emergence of the effects and determinants of the energy paradigm changes on European union economy," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 768–774, Jan. 2018.

[15] K. Klein, S. Herkel, H.-M. Henning, and C. Felsmann, "Load shifting using the heating and cooling system of an office building: Quantitative potential evaluation for different flexibility and storage options," *Appl. Energy*, vol. 203, pp. 917–937, Oct. 2017, doi: 10.1016/j.apenergy.2017.06.073.

[16] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019.

[17] T. Müller and D. Möst, "Demand response potential: Available when needed?" *Energy Policy*, vol. 115, pp. 181–198, Apr. 2018, doi: 10.1016/j.enpol.2017.12.025.

[18] R. D'hulst, W. Labeeuw, B. Beusen, S. Claessens, G. Deconinck, and K. Vanthournout, "Demand response flexibility and flexibility potential of residential smart appliances: Experiences from large pilot test in belgium," *Appl. Energy*, vol. 155, pp. 79–90, Oct. 2015, doi: 10.1016/j.apenergy.2015.05.101.

[19] E. A. M. Klaassen, R. J. F. van Gerwen, J. Frunt, and J. G. Slootweg, "A methodology to assess demand response benefits from a system perspective: A Dutch case study," *Utilities Policy*, vol. 44, pp. 25–37, Feb. 2017, doi: 10.1016/j.jup.2016.11.001.

[20] G. Reynders, R. Amaral Lopes, A. Marszal-Pomianowska, D. Aelenei, J. Martins, and D. Saelens, "Energy flexible buildings: An evaluation of definitions and quantification methodologies applied to thermal storage," *Energy Buildings*, vol. 166, pp. 372–390, May 2018, doi: 10.1016/j.enbuild.2018.02.040.

[21] Eurostat. 2020. *Share of Renewable Energy in Gross Final Energy Consumption*. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/t2020_31/default/table?lang=en

[22] United Nations. (2020). *Energy Statistics*. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/t2020_31/default/table?lang=en

[23] O. National Statistics and U. Kingdom. (2018). *Low Carbon and Renewable Energy Economy, UK: 2018*. [Online]. Available: https://www.ons.gov.uk/economy/environmentalaccounts/bulletins/finalestimates/2018.

[24] S. V. Oprea, A. Bára, and G. Ifrim, "Flattening the electricity consumption peak and reducing the electricity payment for residential consumers in the context of smart grid by means of shifting optimization algorithm," *Comput. Ind. Eng.*, vol. 122, pp. 125–139, Aug. 2018, doi: 10.1016/j.cie.2018.05.053.

[25] A. Mosavi, M. Salimi, S. Faizollahzadeh Ardabili, T. Rabczuk, S. Shamshirband, and A. Varkonyi-Koczy, "State of the art of machine learning models in energy systems, a systematic review," *Energies*, vol. 12, no. 7, p. 1301, Apr. 2019.

[26] K. S. Perera, Z. Aung, and W. L. Woon, "Machine learning techniques for supporting renewable energy generation and integration: A survey," in *Proc. Int. Workshop Data Anal. Renew. Energy Integr.* Cham, Switzerland: Springer, 2014, pp. 81–96.

[27] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. Roper, "Machine learning for estimation of building energy consumption and performance: A review," *Visualizat. Eng.*, vol. 6, no. 1, p. 58, Dec. 2018.

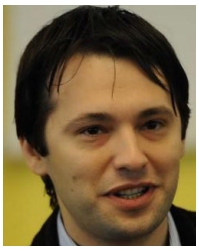
[28] S.-V. Oprea, A. Bára, B. G. Tudorică, and G. Dobriéa, "Sustainable development with smart meter data analytics using NoSQL and self-organizing maps," *Sustainability*, vol. 12, no. 8, p. 3442, Apr. 2020, doi: 10.3390/su12083442.

[29] K. Kourtit, P. Nijkamp, and D. Arribas, "Smart cities in perspective —A comparative European study by means of self-organizing maps," *Innov., Eur. J. Social Sci. Res.*, vol. 25, no. 2, pp. 229–246, Jun. 2012, doi: 10.1080/13511610.2012.660330.

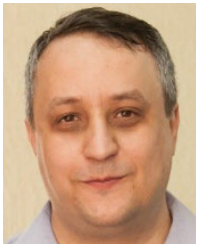
[30] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016, doi: 10.1109/TSG.2016.2548565.

[31] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S. E. Lee, C. Sekhar, and K. W. Tham, "K-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy Buildings*, vol. 146, pp. 27–37, Jul. 2017, doi: 10.1016/j.enbuild.2017.03.071.

[32] D. Hsu, "Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data," *Appl. Energy*, vol. 160, pp. 153–163, Dec. 2015, doi: 10.1016/j.apenergy.2015.08.126.
[33] *Energy Efficiency Indicators?: Fundamentals on Statistics*, IEA, Paris, France, 2014.
[34] *Energy Efficiency Indicators?: Essentials for Policy Making*, IEA, Paris, France, 2014.
[35] *Energy Efficiency Indicators Highlights*, IEA, Paris, France,2017, doi: 10.1017/CBO9781107415324.004.
[36] B. Kruyt, D. P. van Vuuren, H. J. M. de Vries, and H. Groenenberg, "Indicators for energy security," *Energy Policy*, vol. 37, no. 6, pp. 2166–2181, Jun. 2009, doi: 10.1016/j.enpol.2009.02.006.
[37] A. Löschel, U. Moslener, and D. T. G. Rübbelke, "Indicators of energy security in industrialised countries," *Energy Policy*, vol. 38, no. 4, pp. 1665–1671, Apr. 2010, doi: 10.1016/j.enpol.2009.03.061.
[38] *Eurostat, Energy, Transport and Environment Indicators*, Eurostat, Luxembourg City, Luxembourg, 2014.
[39] *Energy Indicators for Sustainable Development: Guidelines and Methodologies*, Environ. Policy Law, Int. At. Energy Agency, Vienna, Austria, 2005. [Online]. Available: https://www.iaea.org/publications/7201/energy-indicators-for-sustainable-development-guidelines-and-methodologies
[40] I. Vera and L. Langlois, "Energy indicators for sustainable development," *Energy*, vol. 32, no. 6, pp. 875–882, Jun. 2007, doi: 10.1016/j.energy.2006.08.006.
[41] Eurostat. (2020). *Total Environmental Investments*. [Online]. Available: https://ec.europa.eu/eurostat/web/environment/environmental-protection
[42] Eurostat. (2020). *Environmental Tax Revenues*. [Online]. Available: https://10.1787/eco_surveys-deu-2012-graph32-en
[43] Eurostat. (2020). *Environmental Protection Expenditure of the Public Sector by Type*. [Online]. Available: https://data.europa.eu/euodp/en/data/dataset/KamWIqgewNrqGH6wPnO10A.
[44] S. L. Cox, A. J. Lopez, A. C. Watson, N. W. Grue, and J. E. Leisch, "Renewable energy data, analysis, and decisions: A guide for practitioners," Nat. Renew. Energy, Golden, Colorado, Tech. Rep. NREL/TP-6A20-68913, 2018.
[45] *Correlation Analysis*. Accessed: Sep. 21, 2020. [Online]. Available: https://businessjargons.com/correlation-analysis.html
[46] I. Hut. *Correlation Tests, Correlation Matrix, and Corresponding Visualization Methods in R*. Accessed: Oct. 25, 2020. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html
[47] *Correlation Test Between Two Variables in R*. Accessed: Aug. 17, 2020. [Online]. Available: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
[48] (2020). *Correlation_and_Dependence*. Accessed: Nov. 29, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Correlation_and_dependence
[49] N. E. Helwig. (2017). *Data, Covariance and Correlation Matrix*. Accessed: Nov. 29, 2020. [Online]. Available: http://users.stat.umn.edu/~helwig/notes/datamat-Notes.pdf
[50] Statistical Tools for High-Throughput Data Analysis. (2020). *Correlation Matrix?: A Quick Start Guide to Analyze, Format and Visualize a Correlation Matrix Using R Software*. Accessed: Nov. 29, 2020. [Online]. Available: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software
[51] T. J. Leeper. (2017). *Missing Data Handling*. Accessed: Sep. 2, 2021. [Online]. Available: https://thomasleeper.com/Rcourse/Tutorials/NAhandling.html
[52] B. V. G. Ciaburro, *Neural Networks with R—Smart Models Using CNN, RNN, Deep Learning, and Artificial Intelligence Principles*. Birmingham, U.K.: Packt, 2019.
[53] A. Ralhan. (Feb. 2018). *Self Organizing Maps*. [Online]. Available: https://towardsdatascience.com/self-organizing-maps-ff5853a118d4
[54] R-Bloggerscom. *Self-Organising Maps for Customer Segmentation Using R*. Accessed: Nov. 26, 2020. [Online]. Available: https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/
[55] M. Oldach. (2019). *10 Tips for Choosing the Optimal Number of Clusters*. Accessed: Sep. 2, 2021. [Online]. Available: https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92
[56] L. Shane. (2014). *Self-Organising Maps for Customer Segmentation using R*. Accessed: Sep. 2, 2021. [Online]. Available: https://www.shanelynn.ie/self-organising-maps-for-customer-segmentation-using-r/
[57] R. RStudio. (May 2019). *Self-Organizing Maps*. [Online]. Available: https://rpubs.com/AlgoritmaAcademy/som



CRISTIAN BUCUR received the master's degree in informatics and the master's degree in economic systems management from the Faculty of Letters and Sciences, specialization Mathematics-Informatics, in 2007 and 2009, respectively, and the Ph.D. degree in web mining in the field of economic informatics from the Bucharest University of Economic Studies, in 2012. He graduated with a postdoctoral internship in deep learning, in 2015. He is currently a Lecturer with the Petroleum-Gas University of Ploiesti. He has 12 years of experience as a developer, an architect, a consultant, and a project manager in the development of cutting-edge analytical applications based on online technologies.



BOGDAN GEORGE TUDORICĂ graduated from the Mathematics—Informatics Department, Faculty of Letters and Sciences, Petroleum-Gas University of Ploiesti, in 1998. He received the master's degree in databases used as business support from the Faculty of Cybernetics, Statistics, and Economic Informatics, Bucharest University of Economic Studies, in 2009, and the Ph.D. degree in economic informatics, in 2015.

His Ph.D. thesis titled Solutions for Organizing Large Volumes of Data.

SIMONA-VASILICA OPREA received the M.Sc. degree through the Infrastructure Management Program from Yokohama National University, Japan, in 2007, the Ph.D. degree in power system engineering from the Bucharest Polytechnic University, in 2009, and the Ph.D. degree in economic informatics from the Bucharest University of Economic Studies, in 2017. She is currently a Lecturer. She also teaches databases, database management systems, and software packages with the Faculty of Economic Cybernetics, Statistics, and Informatics, Bucharest University of Economic Studies.

DUMITRU NANCU is currently a Ph.D. Lecturer with the Faculty of Economic Sciences, University Ovidius, Constanta, and the General Director of the National Credit Guarantee Fund for Small and Medium Size Enterprises (FNGCIMM-S.A.-IFN). He has expertise in the field of business development strategies, business management in small and medium-sized businesses, business development strategies, being at the same time author or coauthor of numerous scientific articles and books.

DOREL MIHAIL DUŞMĂNESCU graduated from the Faculty of Mechanical and Electrical Engineering, Petroleum-Gas University of Ploiesti, in 1989. He received the Ph.D. degree in technical sciences, specialty automation (currently systems engineering), in 2001, the master's degree in management, in 2009, and the Ph.D. degree in economic sciences, economics specialty, in 2013.