Sustainable development goals: classifying European countries through self-organizing maps

Cristina Davino, Nicola D'Alesio

1. Introduction

Environmental sustainability, despite being the subject of different interpretations (Hueting & Reijnders, 1998; Goodland, 1995), involves the preservation of things and qualities valued in the environment (Sutton, 2004). To achieve this goal, the United Nations (Brundtland et al., 1997) included three goals about environmental sustainability among the proposed 17 Sustainable Development Goals (SDGs). The SDGs related to environmental sustainability are the following: number 13, which refers to climate change and its impacts; number 14, which refers to the conservation of water and marine resources; and number 15, which refers to the preservation of forests. Each of these goals is measured through a set of indicators. An important question is understanding what Europe has achieved in terms of environmental sustainability. In this paper, a mapping of the environmental sustainability within the European territory is proposed using Machine Learning techniques. In particular, Self-Organizing Maps (SOMs), an unsupervised clustering method in the framework of artificial neural networks, are exploited to identify and visualize European countries into a low-dimensional grid (Kohonen, 1982a, 1982b). The analysis considers the indicators related to the three SDGs of environmental sustainability (SDG 13, 14, and 15) and aims to identify groups of countries with similar characteristics through a dimensionality reduction, representing them in a two-dimensional map. The reference year was 2019, except for two indicators updated in 2018 and 2020. To ensure the stability of our results, we built several SOMs with different grids and chose the best one using accuracy measures and a Leave-One-Out procedure. The paper is divided as follows: Section 2 shows the concept of environmental sustainability and the different methods of measurement. In Section 3 there is a description of the data and methodology. Section 4 provides the presentation of the results. All the computations are realized using the R packages *kohonen* (Wehrens & Buydens, 2007), *aweSOM* (Julien et al., 2021), *factomineR* (Husson et al., 2016), and *Factoextra* (Kassambara & Mundt, 2017).

2. Literature review

Sustainability has a long and complex history. It was discussed at the end of the eighteenth century as a "derivation from the noun sustenance" (Jenkins & Schröder, 2013). A key point on sustainability is the perspective for the future: it is necessary to manage resources to guarantee them also for future generations (Hueting & Reijnders, 1998). Because of the difficulties to define sustainability, environmental sustainability has also been subject to different interpretations and discussions over time (Goodland, 1995). A proper definition is the following: "the ability to maintain things or qualities that are valued in the physical environment" (Sutton, 2004). This definition seems more appropriate as it allows us to include the sustenance of all facets of physical capital. The definition of environmental sustainability is crucial to provide policymakers with precise information on its development, but an important step of this process is also to understand how to measure it. Efforts to build indicators to measure environmental sustainability have led to the creation of several evaluation exercises. Among the best known there are the SDGs proposed by the United Nations which cover all fields of sustainability (economic, social, and environmental). They are not exempt from

Cristina Davino, University of Naples Federico II, Italy, cdavino@unina.it, 0000-0003-1154-4209
Nicola D'Alesio, University of Campania Luigi Vanvitelli, Italy, nicola.dalesio@unicampania.it

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Cristina Davino, Nicola D'Alesio, *Sustainable development goals: classifying European countries through self-organizing maps*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.17, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 95-100, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

criticism, as they are recent and, according to experts, must be integrated and updated constantly (Hak et al., 2016). Notwithstanding this, they provide an accurate framework of indicators to measure sustainability. In particular, SDGs n°13, 14, and 15 consider indicators aiming to measure environmental sustainability: climate change and its impacts (Climate Action - SDG 13), conservation and sustainable use of the oceans, seas, and marine resources and reduce marine pollution and water acidification (Life Below Water - SDG 14), protection, restoration, and sustainable use of terrestrial, inland and mountain ecosystems (Life on Land - SDG 15).

3. Data and methods

3.1 Data

Data for the three considered SDGs are available on the Eurostat website. We used 2019 as the base year (just two indicators of the SDG-15 are updated to 2018 and 2020). A subset of 14 indicators from the set of 21 indicators was used for the analysis because some of them are not available at the national level for each country and/or because they contained more than 80% of missing values. The units of analysis are represented by the 31 countries[1]. Table 1 shows the list of considered indicators, divided by SDGs, with the acronym used in results figures and tables and with some descriptive statistics[2]. The asterisk ("*") denotes indicators with negative polarity with respect to the concept of environmental sustainability. Missing data and outliers have not been treated because the algorithm of the SOMs can impute a value for the missing data and isolate the effect of the outliers in the extreme regions of the network. All the considered indicators have been standardized before applying the SOM algorithm.

3.2 Methods

Self-Organizing Maps (SOMs) are artificial neural networks that produce a low-dimensional representation of the input space, allowing a dimensionality reduction (Kohonen, 1982a, 1982b, 1990). They use a neighborhood function to preserve the topological properties of the input space. The SOM algorithm is divided into two phases: the competitive phase and the cooperative phase. In the competitive phase for each input vector, the neuron with the minimum distance from the input is selected and it represents the winner. Although several distance measures are available, the Euclidean distance is the most used (Miljković, 2017). The neurons within a grid interact with each other using a neighborhood function such as the Gaussian function. In the cooperative phase, on the other hand, the weights are modified as topologically related subsets on which similar weight updates are performed. During learning, not only the weight vector of the winning neuron is updated, but also those of its reticular neighbors and, therefore, that end up responding to similar inputs. This is achieved with the neighborhood function, which is centered on the winning neuron and decreases with the distance of the grid from the winning neuron. Once the units (the weights) have been initialized, the training phase starts. SOMs training is done through unsupervised learning that can be realized in a sequential formation (or online algorithm: a single statistical unit is inserted into the network at a time) or in batch modality (or batch algorithm: all statistical units are inserted into the network at once) (Matsushita & Nishio, 2020). In our case, it was preferred the online algorithm. We chose the Euclidean distance as a distance measure and the Gaussian function as a neighborhood function.

[1] Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, the Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, Iceland, Norway, Switzerland, and the United Kingdom.

[2] VC means variation coefficient.

Table 1: SDGs Indicators.

SDG – 13 Indicators	**Acronym**	**Min**	**Max**	**Mean**	**VC**	**Skewness**
Net Greenhouse gas emissions*	Net_GHG_Emission	34.80	156.20	80.43	0.34	0.49
Net Greenhouse gas emissions of the LULUCF sector*	Net_GHG_Land	-137.60	172.00	-33.28	2.19	0.97
Contribution to the international 100bn USD commitment on climate related expending	Contr_Intern_commitment	1.00	27.00	14.00	0.57	0.00
Population covered by the Covenant of Mayors for Climate	Population Covered	7.30	91.10	44.09	0.47	0.18
Share of renewable energy in final energy consumption by sector	Share_Ren_Energy	7.05	78.61	25.69	0.70	1.54
Average CO2 emissions per km from new passenger cars*	Average_CO2	59.90	133.00	119.81	0.12	-2.54
SDG – 14 Indicators	**Acronym**	**Min**	**Max**	**Mean**	**VC**	**Skewness**
Surface of Marine Protected Areas	Surface_Marine_Protected_Area	2.30	45.90	16.99	0.69	0.85
Bathing sites with excellent water quality by locality	Bathing_Sites	12.00	4894.00	648.66	1.66	2.41
Marine waters affected by eutrophication*	Waters_Eutrophicated	0.00	5856.00	616.68	2.45	2.61
SDG – 15 Indicators	**Acronym**	**Min**	**Max**	**Mean**	**VC**	**Skewness**
Share of forest area (**2018**)	Forest_Area	10.40	69.90	39.71	0.41	-0.04
Surface of the terrestrial protected areas (**2020**)	Protected_Area	13.20	51.50	27.27	0.38	0.39
Soil Sealing Index*	Soil_Sealing_Index	0.07	17.08	2.53	1.30	3.00
Biochemical oxygen in rivers*	Oxygen_In_Rivers	0.75	3.60	1.95	0.42	0.55
Phosphate in rivers*	Phospate_in_Rivers	0.01	0.22	0.06	0.96	1.31

The most widespread accuracy measures used in the SOM framework are the following:

- Quantization error: Average distance squared between the data points and the nodes in which they are inserted. The lower the value, the more accurate the network will be.
- Percentage of explained variance: it expresses the percentage of variance explained by the model. The higher the value, the more valid the model will be.
- Topographic error: measures how the topographic structure of data is preserved on the map. Assuming values between 0 and 1: 0 indicates an excellent topographic representation (all the best corresponding nodes and best seconds are close), and 1 is the maximum error (the best nodes and the best seconds are never close).
- Kaski-Lagus error: It is the sum of the average distance between the points and their best matching prototypes, and the average geodesic distance between the points and their second-best corresponding prototype. The smaller the error, the more accurate our network will be.

SOMs prove to be a useful and innovative tool for our study, being able to reduce dimensionality and provide a two-or three-dimensional representation of European countries in the different facets of environmental sustainability. There are many studies of the application of these networks in environmental contexts, also in Italy (Carboni et al., 2015).

4. Results

After the indicator selection described in Section 3.1, the analysis is carried out through the following steps: identification of the best SOM through the estimation of several SOMs and accuracy evaluation, clustering of countries, visualization, and interpretation of the results.

4.1 Identification of the best self-organizing map

It is well known that one of the main drawbacks of neural networks is the selection of the architecture. We decided to train several networks with different numbers of neurons and with a grid compatible with the sample size and to select the best SOM by comparing the accuracy measures. The results in Table 2 showed that SOMs with grids 3x5 and 5x4 have very similar performance.

Table 2 - SOMs trials: evaluation with accuracy measures

Grid	Quantization error	% Explained Variance	Topograhic Error	Kaski-Lagus error
3 x 3	4,07	64,28	0,16	5,26
3 x 4	3,28	71,21	0,16	4,92
3 x 5	2,38	79,1	0,06	4,53
4 x 3	2,93	74,24	0,16	5,25
4 x 4	2,97	73,96	0,03	4,15
4 x 5	2,76	75,75	0,06	3,78
5 x 4	2,5	78,04	0,06	3,91
5 x 5	2,54	77,69	0,06	3,47

The choice of the best network between these two SOMs was made taking into account the stability of the results in terms of sensitivity to the specific statistical units (countries). The two networks were trained using a leave-one-out procedure, i.e., they were estimated n-1 times by excluding one country each time. The aim is to assess how sensitive the results shown in Table 2 may be to the exclusion of even one country. Results are shown in Figure 1 where we plot the percentage of variability explained and the quantization error of the 3x5 (left-hand side) and 5 x 4 (right-hand side) networks trained excluding each time a country. We decided to use these two measures because the other two accuracy measures give the same information about the topographic qualities of a SOM. The red lines represent the values of the reference network (with all statistical units and shown in Table 2). Observing the two graphs, it results that the accuracy of the 3x5 SOM improves (quadrant in the bottom right part) by removing 5 statistical units, while the 5x4 SOM is much more unstable as it improves by removing more than half of the observations.

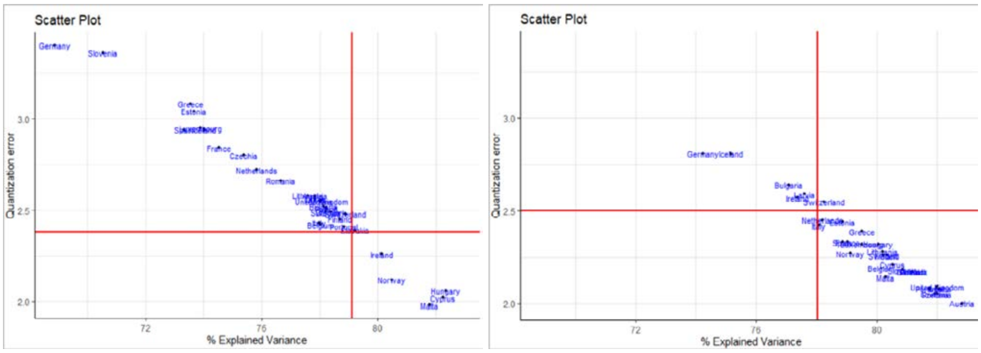

Figure 1 - Scatter Plot of the accuracy measures for the two SOMs (grid 3x5 – left; grid 5x4 - right)

Although of the two selected networks, the 3x5 network is more stable, it is necessary to find its optimal configuration by trying to figure out which of the five countries displayed in the bottom right-hand quadrant is appropriate to eliminate. The proposed procedure proceeds one step at a time starting from the elimination of the statistical unit that provides the most benefit (Hungary) to the one that provides the least benefit (Iceland). Table 6 shows the accuracy measures of these 3x5 SOMs and highlights that the best compromise is obtained just by eliminating Hungary because all the accuracy measures worsen if two or more countries are removed from the analysis.

Table 3 – Grids comparison

Countries	Quantization error	% Explained Variance	Topographic Error	Kaski-Lagus Error
Hungary	2,06	82,34	0,1	4,54
Cyprus	2,83	75,67	0,1	4,54
Malta	2,79	74,73	0,04	3,82
Norway	2,7	74,29	0,11	4,15
Iceland	2,32	77,57	0,08	3,73

4.2 Classification of countries

Once a stable SOM has been achieved, it is possible to identify the best partition of countries by applying a clustering procedure. The SOM built without Hungary is shown in Figure 2 where colors highlight the four groups identified using the Ward criterion.

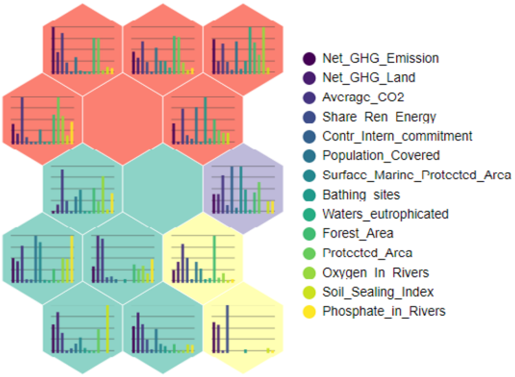

Figure 2 - Visualization of the SOM 3x5 and the partition in four groups

The characterization of the clusters is typically done by comparing, for each indicator, the group averages with the averages on the total sample. Due to lack of space, we report the result of this comparison and the countries belonging to each cluster directly below:

- Group 1, in blue, consisting of Lithuania, Romania, Belgium, the Czech Republic, the United Kingdom, Malta, the Netherlands, Denmark, and Ireland (mainly countries in the continental area), has high net emissions in land use (SDG-13), phosphate in rivers (SDG-14), and land cover index (SDG-15). It can be tagged as the group of "Countries far from achieving all SDGs".
- Group 2, in yellow, consisting of Estonia, Latvia, Finland, Sweden, Iceland, Norway, and Switzerland, (almost all countries in the northern area) has high renewable energy use in the energy sector (SDG-13) and forest areas (SDG-15). It can be tagged as the group of "Countries close to achieving SDG-13 and SDG-15".
- Group 3, in purple, consists of Germany and France and has high marine protected areas (SDG-14) and the highest values of climate change contributions (SDG-13). It can be tagged as the group of "Countries close to achieving SDG-13 and SDG-14".

– Group 4, in red, is composed of Italy, Spain, Portugal, Greece, Croatia, Cyprus, Austria, Slovenia, Bulgaria, Poland, Slovakia, and Luxembourg (these are mainly countries in the Mediterranean region). These countries have a high number of protected areas (SDG-15) but high net emissions (SDG-13). It can be tagged as the group of "Countries close to achieving SDG-15 but far from achieving SDG-13".

The previous classification separates countries closer to achieving a goal and those which are very far from some or all SDGs. This information could help policymakers in assessing what has been achieved so far, what policies need to be implemented to achieve, and which policies in the countries furthest from attainment have either not been implemented or have not been implemented appropriately. The main limitation of this paper is the typical black box effect of neural networks even if the SOMs provide at least a visualization of the grid. A possible future development could be a comparison with other techniques such as cluster analysis, although it will be necessary, in this case, to address the problem of missing data that SOMs are capable of handling. A further problem is the small sample size which has been faced proposing a study of the stability of the results through a leave-one-out procedure.

References

European Commission (2020). Sustainable development in the European Union — Overview of progress towards the SDGs in an EU context — 2020 edition.

G.H. Brundtland et al. (World Commission on Environment and Development, 1987) in *Our Common Future*, Oxford University Press, Oxford.

Goodland, R., (1995). *The concept of environmental sustainability.* Annual review of ecology and systematics, **26**(1), pp. 1-24.

Hák, T., Janoušková, S., Moldan, B. (2016). Sustainable Development Goals: A need for relevant indicators. Ecological indicators, 60, pp. 565-573.

Hueting, R., Reijnders, L. (1998) Sustainability is an objective concept. *Ecological economics*, **27** (2), pp. 139-148.

Husson, F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2016). Package 'factominer'. An R package, 96, 698.

Jenkins, I., Schröder, R. (2013). Sustainability in Tourism A Multidisciplinary Approach, *Springer Science & Business Media*, 1.

Boelaert J., Ollion, E., Sodoge, J. (2021). aweSOM: Interactive Self-Organizing Maps. R package version 1.2. https://CRAN.R-project.org/package=aweSOM

Kassambara, A., Mundt, F. (2017). Package 'factoextra'. Extract and visualize the results of multivariate data analyses, **76**(2).

Kohonen, T. (1982a). Clustering, Taxonomy, and Topological Maps of Patterns in *Sixth International Conference on Pattern Recognition*, Munich, Germany, Oct. 19-22, pp. 114-128.

Kohonen, T. (1982b). Self-organized formation of topologically correct feature maps. *Biol. Cybern*. 43, pp. 59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, **78**(9), pp. 1464-1480.

Matsushita, H., Nishio, Y. (2010). Batch-Learning Self-Organizing Map with Weighted Connections avoiding false-neighbor effects in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-6.

Miljković, D. (2017). Brief review of self-organizing maps in the *40th International Convention on Information and Communication Technology*, Electronics and Microelectronics (MIPRO). IEEE.

Sutton, P. (2004). A perspective on environmental sustainability. *Victorian Commissioner for Environmental Sustainability*, pp. 1-32.

Wehrens, R., Buydens, L.M. (2007). Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software*, **21**(5), pp. 1-19.