Book of Short Papers SIS 2018

Editors: Antonino Abbruzzo - Eugenio Brentari

Marcello Chiodi - Davide Piacentino

PUBLISHED BY PEARSON

WWW.PEARSON.COM

First printing, November 2018, ISBN-9788891910233

SIS 2018 49TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY
PALERMO 20-22 JUNE

Contents

Evaluating sustainability through an input-state-output framework: the case of the Italian provinces

Valutazione delle sostenibilità attraverso un sistema input-state-output: analisi sulle province Italiane

Achille Lemmi[1], Laura Neri[2], Federico M. Pulselli[3]

Abstract In line with the recommendation of monitoring local context, in this paper we propose to investigate regional (NUTS2) and provincial (NUTS3) economic systems in a schematic and usable way: three different indicators are used to take into account resource use (input), societal organization (state) and to quantify the outputs of the system (output). A fuzzy cluster analysis is applied to the input-state-output indicator framework, that, as a whole, represents the interconnection of the three aspects of sustainability, namely environmental, social and economic. This framework is a useful and comprehensive tool for investigating and monitoring local context economic systems.

Abstract *In linea con le direttive di monitorare sistemi come le economie nazionali in un contesto locale, questo lavoro, propone l'analisi dei sistemi economici regionali (NUTS2) e provinciali (NUTS3) in modo semplice schematico e fruibile: tre diversi indicatori sono utilizzati per tenere conto dell'uso delle risorse (input), organizzazione della società (stato) e per quantificare l'output del sistema (output). Un'analisi cluster sfocata viene applicata agli indicatori del framework input-state-output, che, nel suo complesso, rappresentano l'interconnessione dei tre aspetti della sostenibilità, in particolare ambientale, sociale* ed economica.

Key words: Sustainability, input-state-output, fuzzy cluster analysis

[1] Achille Lemmi, ASESD Tuscan Universities Research Centre "Camilo Dagum", lemmiachille@virgilio.it

[2] Laura Neri, Department of Economics & Statistics, University of Siena, laura.neri@unisi.it

[3] Federico M. Pulselli, Ecodynamics Group, Department of Earth, Environmental and Physical Sciences, University of Siena, federico.pulselli@unisi.it

1 Introduction

In recent years there has been an increasing interest in the measurement of collective phenomena at the local level. The EU Committee of the Regions (2014) strongly rec ommend local authorities to define their own "2020 vision", based on a territorial di mension, overcoming the present top-down approach of country targets fitting all reg ions irrespectively. Italy is subdivided into 20 regions (NUTS2) representing the first -level of administrative divisions; the country is further subdivided into 107 province s (NUTS3). Though progressive measures are trying to eliminate this intermediate ad ministrative level, provinces still play an important role in planning, coordination an d cooperation at local level in connection with municipalities and other local bodies. According to OECD Regional Well-being (2016, http://www.oecd.org/cfe/regional-p olicy/hows-life-country-facts-italy.pdf) "Italy has the largest regional disparities amo ng the OECD countries in safety, with the Aosta Valley ranking in the top 1% and Si cily in the bottom 10% of the OECD regions. Important regional differences are foun d also in jobs, environment, community, civic engagement, income and access to ser vices". In line with the recommendation of monitoring local context, in this paper we propose to investigate regional (NUTS2) and provincial (NUTS3) economic systems in a schematic and usable way: three indicators are used to take into account resource use, societal organization and to quantify the outputs of the system. This framework i s consistent with an input-state-output scheme (I-S-O, Pulselli et al., 2015), represent ing the ordered triad environment–society–economy. A three-storey pyramid represe nts the mutual relationships among the three dimensions of sustainability, rotating th e pyramid clockwise, the succession of the stages is oriented from left to right, consi stently with the I-S-O framework

2 Data and Methods

In this framework different combinations of indicators can be used to account for the input-state-output. The study here presented is referred to provincial areas, so the preliminary challenge to face is the data availability. Then, the aim of the research is to produce an "objective" classification of the Italian provinces in terms of the three aspects of sustainability. Such classification should be useful for designing and delivering policy responses to economic, environmental and social needs at the local level.

2.1 The indicators

The input indicator should be representative of what a system extracts/obtains, directly or indirectly, from the environment. Referring to provincial areas, which are sub-regional systems, poor datasets are systematically produced, especially in the environmental field. Therefore, no encompassing methods, like emergy evaluation (as in Pulselli et al., 2015) or ecological footprint, or other

environmental accounting methods can apply. In this case we used an aggregation of energy consumption measures, collected from two institutional databases (Terna: National electric network; DGERM: Ministry of Economic Development). In particular, we selected electricity consumption and sale of a set of fuel types, for all the provinces of Italy. In order to aggregate these measures, we calculated the equivalent in terms of CO_2 emission to show both the use of resources (electricity and fuels) and the environmental pressure (emissions) on the other. The result is an estimation of gross CO_2 emission due to almost all the items composing the energy sector. In order to monitor the environmental pressure of human activities on each provincial territory and compare provinces, the amount of CO_2 per unit area is computed. This choice helps determine the contribution of human actions (that imply energy consumption) to climate change independently of the number of inhabitants in each area. To encompass the characteristics of the state, a synthetic indicator describing a form of societal organization should be used. Considering the critical importance of reducing unemployment, in order to drive toward inclusive society, an indicator related to the labour market seems to be appropriate. Again, availability and reliability of data at NUTS3 level is the critical issue to face: it has been retained that the most reliable official statistics should be the Labour Force Survey, so the unemployment rate has been chosen as key state indicator.Gross Domestic Product (GDP) is the principal aggregate for measuring economic development/growth of a country/region. In this analysis we maintain this logic, considering the GDP per inhabitants in purchasing parity power as output indicator of the I-S-O system.

2.2 Methods

Our goal is to produce an "objective" classification of the Italian provinces in terms of the three aspects of sustainability, namely environmental, social and economic, according to the chosen triad: CO_2 per unit area, unemployment rate and GDP per inhabitants in purchasing parity power.
Data clustering is recognized as a statistical technique for classifying data elements into different groups (known as clusters) in such a way that the elements within a group possess high similarity while they differ from the elements in a different group. By using the triad of indicators, the clustering algorithm starts from an initial partition of the provinces into a fixed number of groups, where each group is initially randomly chosen. The grouping is then updated: based on the distance between every single observation and the reference objects of each group, every observation is reallocated to the closest group aiming to produce a classification that is reasonably "objective" and "stable". The classification obtained by using the crisp cluster analysis suffers for both a poor homogeneity within group and a lacking separation between the groups. For this reason, we explored the clustering procedure by using a soft clustering known as Fuzzy Cluster Analysis, a very important clustering technique based on fuzzy logic. In case of soft clustering techniques, fuzzy sets are used to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering.

Objects on the boundaries between several clusters are not forced to fully belong to one of the cluster, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The most popular algorithm, the fuzzy c-means (developed by Dunn in 1973 and improved by Bezdek in 1981), aims at minimizing an objective function –the weighted within-groups sum of square- whose (main) parameters are the membership degrees and the parameters determining the localisation as well as the shape of the clusters. Objects are assigned to clusters according to membership degrees in [0,1]: 0 is where the data point is at the farthest possible point from a cluster's center and 1 is where the data point is the closest to the center. Each of the c clusters is represented by a cluster center. These centers are chosen randomly in the beginning, then each data vector is assigned to the nearest prototype according to a suitable similarity measure and each center is replaced by the centre of gravity of those data assigned to it. The alternating assignment of data to the nearest center and the update of the cluster centres is repeated until the algorithm converges, i.e., no more changes happen.

Although the extension from crisp to fuzzy clustering seems to be an obvious concept, it turns out that to actually obtain membership degrees between zero and one, it is necessary to introduce a so-called fuzzifier in fuzzy clustering. Usually, the fuzzifier is simply used to control how much clusters are allowed to overlap. This fuzzifier function creates an area of crisp membership values around a prototype while outside of these areas of crisp membership values, fuzzy values are assigned. The analysis has been performed by using the fuzzy clustering with polynomial fuzzifier (Frank Klawonn and Frank Hoppner, 2003).

A problem that frequently occurs in real data analysis is the presence of one or more observation presenting anomalous values, i.e. outliers. Such a subset, that may be referred to as noise, tends to disrupt clustering algorithms making difficult to detect the cluster structure of the remaining domain points. According to the adopted approach the first k standard clusters are homogeneous, whereas the noise cluster, serving as a "garbage collector", contains the outliers and is usually not formed by objects with homogeneous.

All data have been standardized before performing the cluster analysis. The analysis has been conducted by using R and specifically the R package *fclust* (Giordani, Ferraro, 2018). The package provides the cluster solution, cluster validity index and plots and also the visualization of fuzzy clustering results.

The analysis has been performed by using the fuzzy clustering with polynomial fuzzifier with noise cluster. For assessing cluster validity, some have been evaluated: here, just the so called partition coefficient (PC) is presented. Given that the closer to unity the PC index the "crisper" the clustering is and that value close to $1/n_c$ (where n_c is the number of clusters) indicates that there is no clustering tendency, a PC=0.92, indicates a significant cluster structure.

As regard to the visual inspection of fuzzy clustering results -VIFCR- (Klawonn et al., 2003) is a scatter plot where, for each object, the coordinates u_1 and u_2 denote, respectively, the highest and the second highest membership degrees. All points lie within the triangle with vertices (0,0), (0.5,0.5) and (1,0). In the ideal case of (almost) crisp membership degrees all points are near the vertex (1,0). This graph has been

evaluated for different partition and the partition with three cluster, plus the noise one, containing just one province (Milan), seems to be the best one (Fig.1).
The algorithm applied to reach the fuzzy clustering solution, assigns an objects to clusters only if the corresponding member function degree is greater than 0.5. In this way the closest hardest partition can be identified by assigning an object to the cluster according to the maximal membership function (>0.5) and the characteristics of each cluster can be identified. The cardinalities and the average membership function of the closest hardest cluster are reported in Table 1, as well as the average of each indicator considered in the clustering procedure. Specifically, in Table 2, the list of provinces belonging to each cluster, according to the maximal membership function, are reported; it is worth to point out that cluster 4, the noise one, includes just Milan.

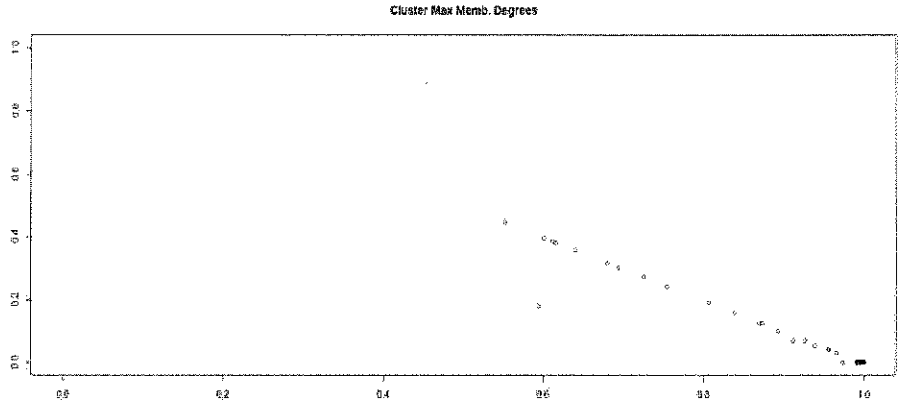

Figure 1: Scatter plot: for each observation, the coordinates u_1 and u_2 denote, respectively, the highest and the second highest membership degrees.

Table 1: Size, average membership function and average values for the indicators by cluster

Cluster	*size*	*m.f.*	*CO_2_area*	*Unempl_rate*	*pps_ab*
1	27	0.96	2272.88	7.75	30866.67
2	44	0.94	912.13	10.31	24854.34
3	35	0.97	804.13	19.78	17046.71
4	1	1	12755.85	7.68	44493.13

In order to have a vision of the distribution of each indicator within each cluster, the boxplots can be observed (Figure 2).
PCA plot (Figure 3) is a very useful tools to visualize the data: this has nothing to do with the type of clustering algorithm or the accuracy of the algorithm used, however it is a useful representation to recognize the utility of the fuzzy clustering, given that the clusters, are clearly separated but the borderline units provinces in this study, are very close.

Table 2: The list of provinces belonging to the four clusters according to the maximal membership function

Cluster	*Provinces*
1	Aosta,Bergamo,Bologna,Bolzano,Brescia,Como,Cremona,Firenze, Forlì-Cesena, Genova,Lecco,Livorno,Mantova, Modena,Padova, Parma,Prato,Ravenna,Reggio-Emilia,Roma,Trento,Treviso,Trieste,Varese,Venezia,erona,Vicenza
2	Alessandria,Ancona,Arezzo,Ascoli,Piceno,Asti,Belluno,Biella,Chieti,Cuneo, Ferrara,Frosinone,Gorizia,Grosseto,Imperia,L'Aquila,La Spezia, Latina,Lodi, Lucca,Macerata,Massa Carrara,Novara,Nuoro,Pavia,Perugia,Pesaro Urbino, Pescara,Piacenza,Pisa,Pistoia,Pordenone,Potenza,Rieti,Rimini,Rovigo,Savona, Siena,Sondrio,Teramo,Terni,TorinoUdine,Verbano-Cusio-Ossola,Vercelli
3	Agrigento,Avellino,Bari,Benevento,Brindisi,Cagliari,Caltanissetta,Campobasso, Carbonia-I.,Caserta,Catania,Catanzaro,Cosenza, Crotone,Enna,Foggia,Isernia, Lecce,Matera,Medio,Campidano,Messina,Napoli,Ogliastra,Olbia-T.,Oristano, Palermo,Ragusa,Reggio C., Salerno,Sassari,Siracusa,Taranto,Trapani,Vibo Valentia,Viterbo
4	Milano

3. An attempt of taxonomy according to the triad

Cluster composition (Table 2) and average values of indicators (Tab. 1) suggest a hig h heterogeneity among clusters emphasizing the existence of economic disparities. In Cluster 1 there are 27 provinces: Rome, plus 26 provinces located in the North or Central Italy. Comparing Cluster 1 with Cluster 2 and 3, we can observe the highest a verage level of CO_2 per area, the highest GDP per capita and the lowest unemployme nt rate. Observing Fig. 2, we can also realize that the unemployment rate is quite ho mogeneous within the cluster while the GDP per capita and the level of CO_2 per area present the largest variability if compared with the variability of Cluster 2 and 3. Thi s result is justified considering the presence of very peculiar provinces in this cluster, like Bolzano (belonging to an autonomous region) with the highest GDP per capita (40000 Euro) (out of Milan), and the lowest unemployment rate among all the Italian provinces, and the lowest level of CO_2 per area within Cluster 1. and Livorno, the pr ovince presenting the lower membership function for this cluster (0.64), maybe due t o its GDP which is closer to the average GDP per capita computed for Cluster 2. Cluster 2 is the most heterogeneous cluster as regard to the geography of provinces. I t is composed mainly by provinces located in the North and Central Italy plus provin ces of Lazio (out of Rome and Viterbo) and Abruzzo plus Nuoro and Potenza. All th e provinces in Cluster2, out of Torino, have small medium dimension. This cluster is characterized by high variability in GDP, ranging from 17900 for Nuoro to 29900 fo r Siena and low variability in CO_2. Torino presents the lowest membership function (0.55) among all the provinces, being a borderline unit between Cluster 2 and 1. Clus ter 3 is composed by whole Calabria, Campania, Molise, Puglia, Sicilia and Sardegn a (out of Nuoro), plus Viterbo (Lazio). This cluster is very similar to the one, labelle d as "Minimal system with high social risks" by Bertin (2012) in his classification co ncerning the welfare state systems of the Italian regions. Cluster 3 is characterized by the lowest average level of CO_2 emission per area, although such level is biased

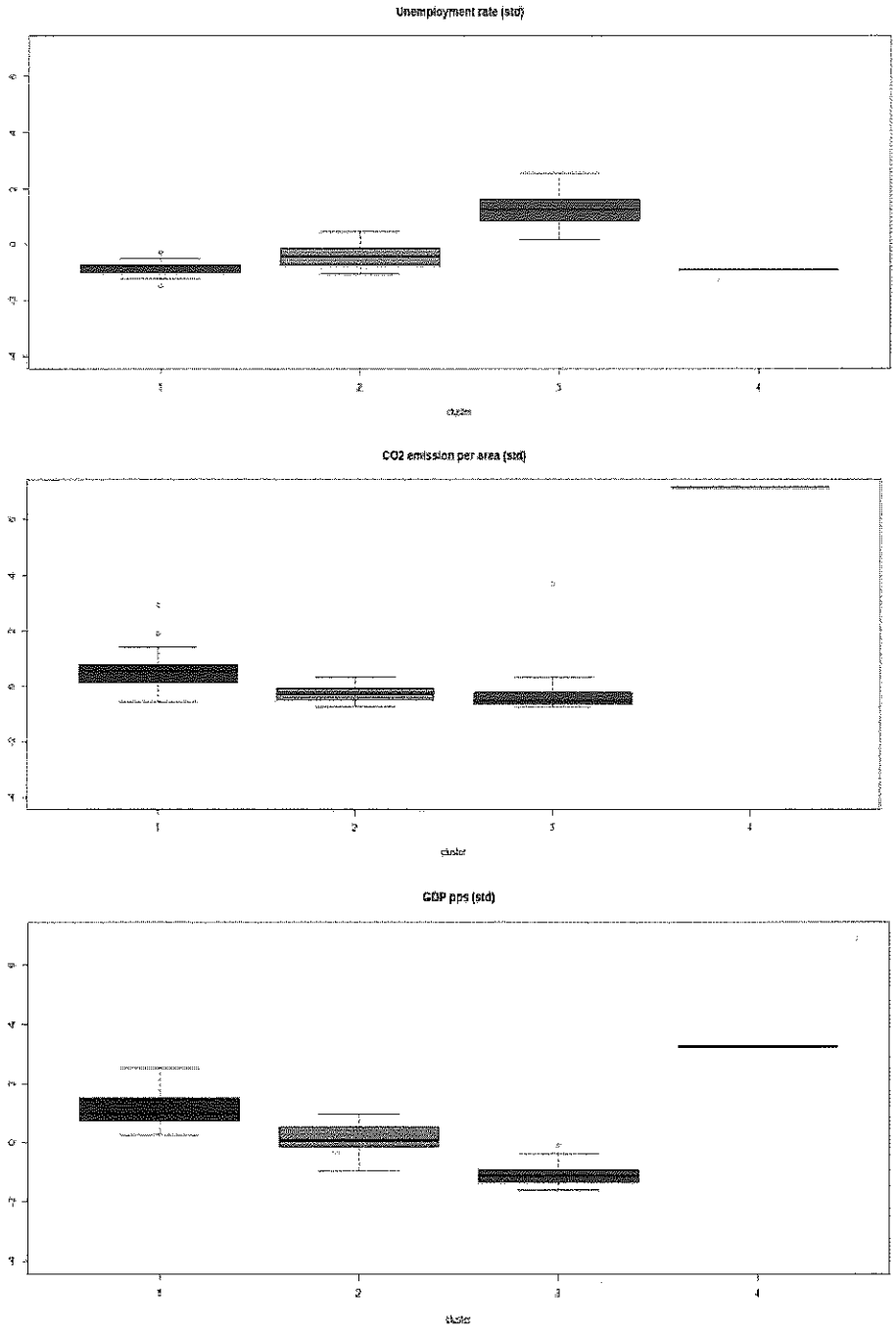

Figure 2: Boxplot-cluster structure

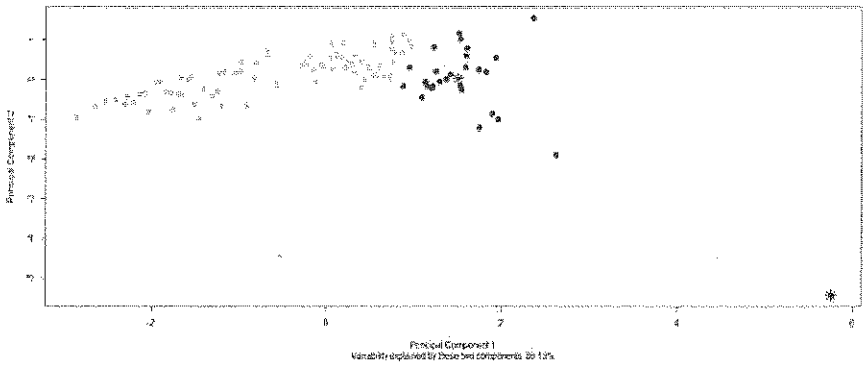

Figure 3: Principal component plot of the cluster structure

due to the membership of Napoli, presenting the highest level of CO_2 among all Itali an provinces (out of Milan). As regard to the unemployment rate and GDP, the clust er presents an evident disparity with respect to the others. As regard to GDP per capi ta, the level of Cagliari (23400), the highest value within cluster 3, is lower than the minimum value in Cluster 1 (25500) registered for Livorno. It is interesting to observ e that Napoli is a borderline unit between cluster 3 and 2.

Finally, Cluster 4, the one containing just Milan, is for sure a true outlier with respect to the level of CO2 emission per area (more than ten times the average values of all the other provinces i.e.12755.85 vs 1218.29) and GDP per capita which in Milan is nearly double than the average values of all the other provinces (44493 vs 23812).

The analysis conducted confirm the well-known dualism, resulted in a North-South divide in GDP per capita and in labour-market performance, adding a new element: the Southern Italian provinces are homogeneous with respect to the considered characteristics whilst the Northern and Central provinces are not homogeneous even if they belong to the same region. This result suggests that local policies can be better aligned and tailored to specific local opportunities and challenges. Moreover, the clustering structure obtained can be an useful tool to adopt, considering that provinces belonging to the same cluster could share ways to generate better outcomes for environment, jobs and the economy, trying to reach provinces of their cluster.

References

1. Bertin, G.: Crises and change processes of the welfare systems, Italian Sociological Review, vol. 2, n. 1, pp. 1–13 (2012)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Ac. Pub., Norwell, MA, (1981)
3. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3, 3, 32-57 (1974)
4. Klawonn, F., Höppner, F.: What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier. In: Advances in intelligent data analysis, pp. 254-264. LNCS (2003)
5. Giordani, P., Ferraro, M.B.: fclust: Fuzzy Clustering, R package version 1.1.2, http://cran.r-project.org/package=fclust.
6. Pulselli, F.M., Coscieme, L., Neri, L., Regoli, A., Sutton, P., Lemmi, A., Bastianoni, S.: The World Economy in a cube: a more rational structural representation of Sustainability. Glob. Environ. Change 35, pp. 41–51 (2015).