ON AN ANALYSIS ON SOME INDICATORS OF WELL-BEING IN ITALY

Gabriella Schoier, Massimiliano Giacalone

1. Introduction

In the last decades the progress made in science and technology has contributed to the evidence of an evolving world.

The new perspectives in knowledge discovery in databases upon economic and social data apply data mining mechanisms that monitor models and patterns, compare them, detect changes, and describe these changes. Having this in mind, some data mining researchers have developed methods and techniques to study the evolution of different phenomena (see eg. Aggarwal, 2005).

In particular, as regards the problem of explaining the economic evolution of the well-being of people and households, some macro-economic statistics such as GDP seem to not give a detailed picture of the living conditions of common people (see, e.g. Maggino, 2017; Chelli *et al.*, 2016).

For this reason the Italian National Institute of Statistics (ISTAT) declined a multidimensional approach at a detailed territorial level, that is, at the provincial level (NUTS3) using a wide spectrum of indicators grouped into domains. (see, e.g., Mazziotta and Pareto, 2017) related to Health, Education, Work and Life balance, Economic well-being, Social relationships, Politics and institutions, Security, [1]Landscape and cultural heritage, Environment, Innovation research and creativity, and Quality of services. These indicators can help in describing the territories because they can spot situations concerning different places. Different authors work on the Sustainable and Equitable Well-Being at local level[2] using multidimensional techniques (see e.g. Monte *et al.*, 2022, Giacalone *et al.*, 2022).

The Organization for Economic Co-operation and Development (OECD) states, "the OECD Framework for Measuring Well-Being and Progress is built around three distinct domains: material conditions, quality of life and sustainability, each with their relevant dimensions" (OECD, 2019).

[1]

[2] See: https://www.istat.it/it/files//2019/05/Nota-metodologica.pdf; https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-wellbeing/bes-at-local-level; www.besdelleprovince.it.

In this paper, we address the problem of monitoring the well-being in Italy. This can help decision makers of different areas make better economic and political decisions.

The indicators, one for each of the equitable and sustainable well-being domain, have been used to measure it. They were selected from the ISTAT database (see ISTAT(c), 2022) on the base of objective criteria.

The aim of this research is to understand how equitable and sustainable well-being in the territories (provinces) influence different variables.

The new idea regards the application of decision trees to some indicators, one for each of the equitable and sustainable well-being domain for Italian provinces.

We consider both classification and regression trees; two different dependent variables have been chosen:

-classification trees dependent variable Macroregion according to NUTS 1: North West, North East, Center, Islands and South of Italy,

- regression trees dependent variable Life Expectancy at Birth .

Our objective is to see how different indicators (one for each domain) move to analyze and then monitor well-being in Italy in particular regarding the chosen dependent variables.

2. Equitable and sustainable well-being at territorial level

The economic evolution regarding the well-being of people and households can be analyzed by using the equitable and sustainable well-being (BES) a multidimensional approach, that identifies 12 well-being domains[3]; for each of them, a set of indicators is given (at NUTS2 level).

BES is becoming a more and more important tool to evaluate the progress of society from an economic, social, and environmental point of view. Consequently, the Italian Economic and Financial Document has included some BES selected indicators since 2017[4]. The interest in BES has been growing over time, especially for Italian provinces and cities (NUTS3) (see Taralli, 2013), so in the 2018, ISTAT issued for the first time a system of BES indicators at the NUTS3 level.

The BES domains at the local level are the same as those at national level, with an exception made for the subjective well-being domain because of the lack of subjective indicators at the local level. The 11 domains and the chosen indicators are listed in Table 1 they belong to the 2022 (10/3/2022) version of the database (see ISTAT(c), 2022).

[3] https://www.istat.it/it/files//2013/03/bes_2013.pdf.

[4] See https://www.gazzettaufficiale.it/eli/id/2017/11/15/17A07695/sg.

Table 1 – *Domain, name and description of considered indicators.*

Domain	Name	Description
Health	Life_Exp	Life expectancy at birth
Education	Neet	People not in education, employment, or training (Neet)
Work and life balance	Unempl	Non-partecipation rate
Economic well-being	Loans	Rate of bad debts of bank loans to families
Social relationships	Acc_Sc	Accessible schools
Politics and institutions	Women	Women and political representation at municipality level
Security	Crimes	Number of other crimes reported (theft of any kind and robberies at home) on total population per 10,000 inhabitants
Landscape and cultural heritage	Rural	Spread of rural tourism facilities
Environment	Waste	Separate collection of municipal waste
Innovation, research and creativity	Cult_Emp	Cultural employment (% of total employment)
Quality of services	Elect	Irregularities in electric power distribution

In 2022, a set of indicators consisting of 70 measures has been published; each domain is not formed by the same number of indicators, and almost half of the indicators do not give values before 2008. Three domains (Social relations, Landscape and cultural heritage, and Innovation research and creativity) do not have data from before the 2008. Furthermore, the data related to the Social relations domain are still missing up to 2014. Some indicators present values only in well-defined years, that is, Voter turnout in European elections and Voter turnout in regional elections (Politics and institutions domain).

The criteria for the choice of each indicator inside the dimension is done as in Monte *et al.*, 2022, i.e. the correlations within each domain, the variability indexes (i.e. coefficient of variation and the quartile difference for standardized data) and the adequacy of the indicator to the analysis to be carried out.

3. The methodology: decisional trees

As it is well known decision trees are a part of hierarchical classification or segmentation techniques. These techniques have the purpose of "sorting" statistical units into the various classes of a dependent variable on the basis of the values of one or more explanatory variables.

In 1984 Breiman and others introduced an innovative segmentation technique. This technique is called Classification and RegressionTrees (CART). It is presented as a recursive and binary partition methodology.

The hierarchical segmentation process used in the construction of decisional trees consists in divide the statistical units in a finite number of disjoint subgroups in order to guarantee an internal homogeneity higher than that of the initial dataset and a high heterogeneity between the subgroups.

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name says, it uses a tree-like model of decisions. Though it is a commonly used tool in data mining for deriving a strategy to reach a particular goal, it is also widely used in machine learning. They are easy to interpret and make for visualizations, they make possible to reproduce work, they can handle both numerical and categorical data, they perform well on large dataset and are extremely fast.This type of models are used for both classification and regression (see eg. Gareth *et al.*, 2021).

The performance of a tree can be further increased by pruning in particular by using CART (see Breiman *et. al.*, 1984) example of indices used for pruning are Gini and Entropy for classification trees and RSS (residual sum of squares) for regression trees.

$$G = 1 - \sum_j f_j^2 \text{ Gini index} \quad (1)$$

$$H = -\sum_j f_j \ log \ log \ f_j \text{ Shannon index} \quad (2)$$

where f_j relative frequency.

It involves removing the branches that make use of features having low importance. In so doing the complexity of tree is reduced and the power by reducing overfitting is increased. The simplest method of pruning starts at leaves and removes each node with most popular class in that leaf, this change is kept if it does not deteriorate accuracy. It is also called reduced error pruning. More sophisticated pruning methods can be used such as cost complexity pruning where a learning parameter (alpha) is used to weigh whether nodes can be removed based on the size of the sub-tree. In order to obtain the sequence of trees of decreasing dimension, one defines, for every tree $T \leq T_{max}$ a cost complexity function $R_\alpha(T)$

$$R_\alpha(T) = \hat{R}(T) + \alpha\left|\tilde{T}\right| \quad (3)$$

where $\hat{R}(T)$ estimate of rate of wrong classification, $\left|\tilde{T}\right|$ the numbers of leaves.

4. The application

This section describes the data[5] used and the results of applying decision trees. As previously anticipated, we use the ISTAT database "*Misure del benessere dei territori. Tavole di dati (2022)*" [6]. We apply one indicator for each domain[7]. This is because of the following considerations:

- the different number of indicators by domain would lead to an initial distortion, resulting in different weights for each domain;
- there are cases in which the choice of only one comparable indicator is the only possible, because of the presence of numerous missing data in the table in relation to some domains;
- a similar approach is used by Ciommi *et al.* (2017), in which the domains of the territorial BES are described by a single indicator given the limited availability of homogeneous data.

4.1. Classification trees

As regards classification trees we have considered as dependent variable Macroregion according to NUTS 1: North West, North East, Center, Islands and South of Italy.

We have applied different models with and without pruning. At the end on the base of a compromise between complexity and rate of error we have chosen a classification tree with split and prune Gini index.

As we can see on the base of the model the rate of error is very low for Islands, North West and South and higher for North East and Center (Table 2).

Table 2 □ *Confusion matrix based on the model.*

Macroregion	Center	Islands	North East	North West	South	Error rate
Center	16	1	0	0	5	0.2727
Islands	0	12	0	0	1	0.0769
North East	3	0	14	5	0	0.3636
North West	1	0	1	23	0	0.0800
South	0	1	0	0	23	0.0417

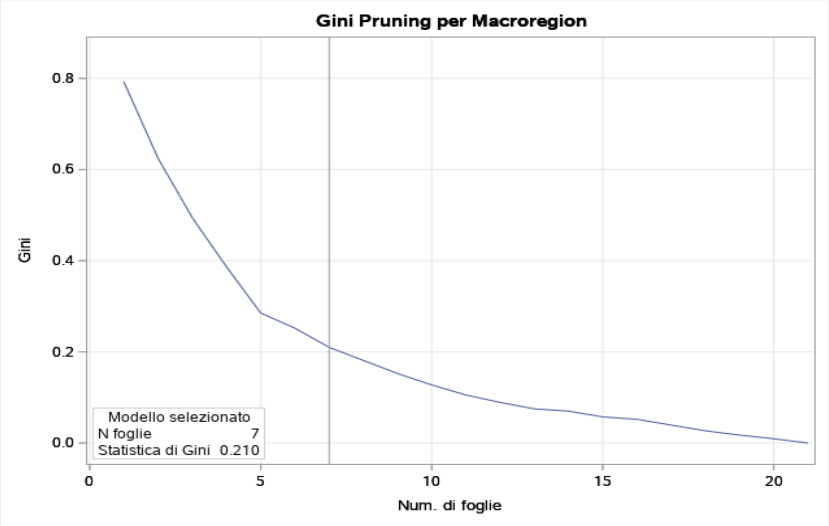
[5] We have considered 106 provinces as the data for Olbia Tempio Pausania, Ogliastra, Medio Campidano, South Sardegna, Carbonia Iglesias are not available.

[6] The 11 domains and the chosen indicators are listed in Table 1. The data are available at the url https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level

[7] We have performed the analysis using SAS language.

In Figure 1 the selected model using the Gini's index for the pruning is reported.

Figure 1 – *The selected model using the Gini's index for pruning*.



As one can see the number of chosen leaf according to the Gini index are 7 ; the value of the Gini index is 0.210 while the entropy of Shannon is higher that is 0.547.

The importance of the variables for the construction of the classification trees is reported in Table 3

Table 3 – *Variable importance.*

Variable	Realative importance	Importance	Count
Unempl2021	1.0000	5.3405	2
Life_Exp2020	0.6957	3.7155	1
Women2021	0.6323	3.3769	1
Elect2020	0.3948	2.1082	1
Neet2021	0.3512	1.8755	1

In the next figure the obtained classification tree is produced. The leaves (terminal nodes) are: node3, node 6, node 7, node 8, node A, node B, node C.

Figure 2 – *Classification tree.*

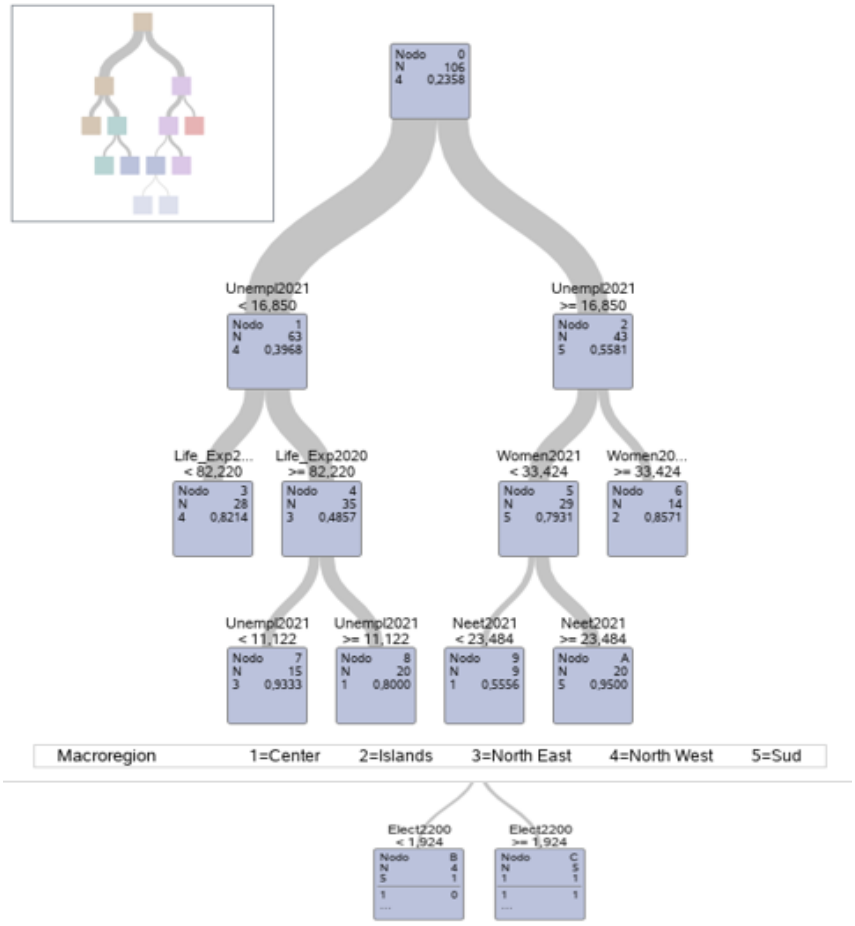


Figure note:.Node 9 is divided in nodeB and node C.

There are 106 units in the root node (node 0). These units are divided into 63 units with Unempl2021 <16.850 (node 1) and 43 units for node 2 with Unempl2021≥ 16.850 (node 2). Node 1 is assigned to class 4 (North East) while node 2 to class 5 (South).

The variable Unempl2021 and the division point 16.850 are chosen to minimize the impurity of the root node measured by the Gini index.

The units of node 1 have been divided into 28 units for which Life_Exp2020 <82,220 (node 3) assigned to class 4 (North West) and 35 for which Life_Exp2020 ≥82,220 (node 4) assigned to clas 3 (North East).

The units of node 2 have been divided into 29 units for which Women2021 <33.424 (node 5) and 14 units for which Women2021 ≥33.424 (node 6).

The classification tree provides simple rules for predicting Macroregion. For example, a unit for which it is expected that Unempl2021≥ 16.850 the percentage of unemployed is greater than 16.850 and the percentage of Women and political representation at municipality level is greater than 33,424 is assigned to the Islands.

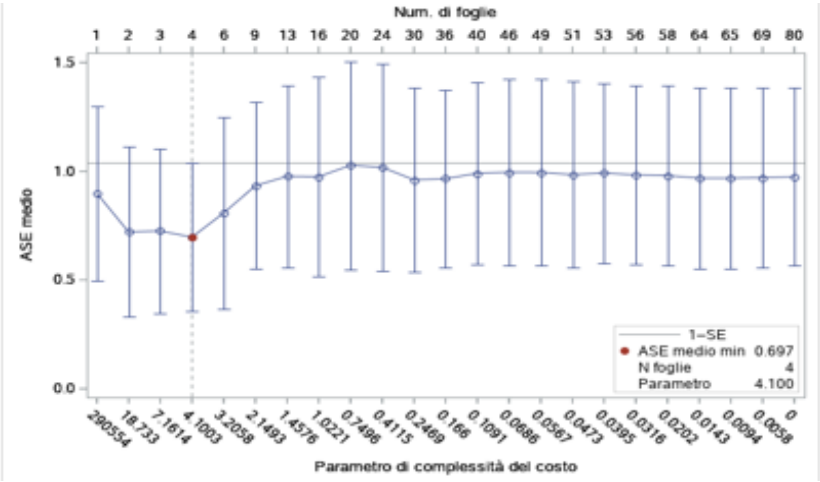
4.2. Regression trees

In order to apply regression trees we have chosen as dependent variable Life Expectancy at Birth. We try to predict the dependent variable on the base of the other variables.

We have applied different models with and without pruning. At the end on the base of a compromise between complexity and rate of error we have choosen a regresssion tree with split RSS and prune cost complexity.

We have carried out the pruning of the tree to avoid overfitting of the model on the data and to find a compromise between simplicity and discriminatory power.As regards the pruning we have preferred cost complexity which is an algorithm based on a trade off between the complexity (size) of the tree and the error rate to prevent overfitting. The final tree has a depth equal to 5 leaves and a parameter of cost complexity of 4.100.The ASE (Average Square Error for Regression) is given by the ratio between RSS (residual sum of squares) and the number of units of the node.In our case the ASE medium minimum is equal to 0.697.

Figure 3 – *The selected model using pruning costcomplexity.*

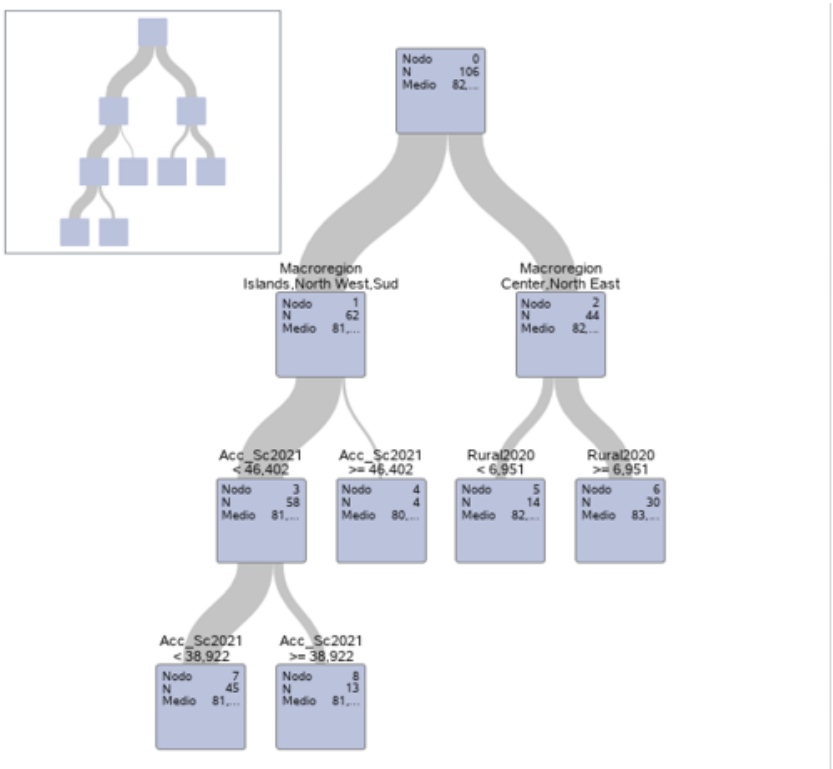


Considering variable importance from Table 3 we can see which variables are the most important: Macroregion, Accessible schools and Spread of rural tourism facilities. The presence of the Macroregion variable shows us once again that there is actually a basic difference between the various Italian macro areas.

Table 4 – *Variable importance.*

Variable	Realative importance	Importance	Count
Macroregion	1.0000	5.3903	1
Acc_Sc2021	0.7484	4.0340	2
Rural2020	0.3846	2.0730	1

Figure 4 – *Regression trees with pruning costcomplexity*



There are 106 units in the root node (node 0). These units are divided into 62 units which belong to Islands, North West and South (node 1) and 44 units with Macroregion Center and North East (node 2). Node 1 is assigned to people with Life Expectancy at birth of approximately 81 years while node 2 to people with Life Expectancy at birth of approximately 82 years.

The units of node 1 have been divided into 28 units for which Acc_Sc2021 <46.402 (node 3) with Life Expectancy at birth of approximately 81 years and 4 for Acc_Sc2021 ≥46.402 (node 4) with Life Expectancy at birth of approximately 80 years.

The units of node 2 have been divided into 14 units for which Rural2020 <6.951 (node 5) and 30 units for which Rural2020 ≥6.951 (node 6).

The regression tree provides simple rules for predicting Life Expectancy at birth. For example, a unit for which it is expected that lives in the North East with a value for Rural2020>6.951 is expected to live in mean approximately 83 years.

5. Conclusions

The purpose of this paper is to see how, on the basis of the data relating to sustainable equitable well-being defined through some indicators, there is the possibility of predicting the trend of variables of interest. In order to obtain this result decision trees have been used; as regards classification trees the choosen dependent variable has been Macroregion while for applying regression trees the dependent variable has been Life Expectancy at Birth.

Future developments regards the applications of decisional trees in conjunction with random forest. Random forest can to improve the forecasting of regression trees as it is, one of the most popular machine learning prediction algorithms. It can be considered an elaboration of regression trees by averaging the predictions of a large number of randomly subsampled regression trees so to obtain a more stable solution.

References

AGGARWAL C. C. 2005. On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, pp. 587-600

BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J. 1984. *Classification and Regression Trees*. New York: Taylor & Francis.

CIOMMI M., GIGLIARANO C., EMILI A., TARALLI S., CHELLI F.M. 2017. A New Class of Composite Indicators for Measuring Well-Being at The Local Level:

An Application to the Equitable and Sustainable Well-Being (BES) of the Italian Provinces, *Ecological Indicators*, Vol. 76, pp. 281-296

CHELLI F., CIOMMI M., EMILI A., GIGLIARANO C., TARALLI S. 2016. Assessing the Equitable and Sustainable Well-being (BES) of the Italian Provinces. *International Journal of Uncertainty, Fuzziness and Knowledge- Based Systems*, Vol. 24, No. suppl. 1, pp. 39-62.

GARETH J., WITTEN D., HASTIE T., TIBSHIRANI R. 2021. *Introduzione all'apprendimento statistico*, in Salini S., Gaito S., Boracchi P., Ambrogi F., Manzi G., Biganzoli E. (Eds), Piccin Editore.

GIACALONE M., MATTERA R., NISSI, E. 2022. Well-Being Analysis of Italian Provinces with Spatial Principal Components, *Socio-Economic Planning Sciences*, Vol. 84, pp.101377.

MAGGINO F., 2017. Developing indicators and managing the complexity. In F. Maggino (Ed.), Complexity in society: From indicators construction to their synthesis, Social indicators research series Cham: Springer, Vol. 70, pp. 87–114.

MAZZIOTTA M., PARETO A. 2017. Synthesis of indicators: The composite indicators approach. In F. Maggino (Ed.) Complexity in society: From indicators construction to their synthesis, Social indicators research series Cham: Springer, Vol. 70, pp. 159-191.

MONTE A., SCHOIER G. 2022. A Multivariate Statistical Analysis of Equitable and Sustainable Well -Being Over Time, *Social Indicators Research*, Vol. 161, No. 2-3, pp. 735-750.

OECD 2019. Measuring Well-being and Progress: Well-being Research, Paris: OECD Publishing.

TARALLI S., 2013. Indicatori di Benessere Equo e Sostenibile delle Province: Informazioni Statistiche a Supporto Del Policy-Cycle e della Valutazione a Livello Locale, *RIV Rivista Italiana di Valutazione*, Vol. 55, pp. 171-187.

SUMMARY

The starting point of this paper has been the consideration that the use of GDP as an indicator of the well-being nowadays is not sufficient to describe the economic situation of a country in terms of sustainable well-being in Italian Benessere Equo Sostenibile (BES). The Italian National Institute of Statistics (ISTAT) consider a multidimensional approach to measure equitable and sustainable well-being. Following this approach we have chosen a certain number of indicators, on the basis of their features, *i.e.* Health, Education, Work and life balance, Economic well-being, Social relationships, Politics and institutions, Security, Landscape and cultural heritage, Environment. These indicators can help in describing the territories in particular in our case the provinces (NUTS3).

To have a description of some aspects of the economic situation multivariate analysis have been applied. The new idea regards the application of decision trees.

We consider both classification and regression trees; two different variables have been choosen. As regards classification trees the dependent variable is Macroregion while for regression trees dependent variable is Life Expectancy at Birth.

Gabriella SCHOIER, Department of Economics, Business, Mathematics and Statistics, University of Trieste, Italy, gabriella.schoier@deams.units.it

Massimiliano GIACALONE, Department of Economics, University of Campania "Luigi Vanvitelli", Italy, massimiliano.giacalone@unicampania.it