MULTIVARIATE REGRESSION TREE TO INVESTIGATE THE ITALIAN MORTALITY RATES

Giulia Contu [1], Luca Frigau[1], Marco Ortu [1] and Sara Pau [2]

[1] Department of Business and Economics, University of Cagliari,(e-mail: giulia.contu@unica.it, frigau@unica.it, marco.ortu@unica.it)

[2] Department of Business and Economics, University of Sassari,(e-mail: spau@uniss.it)

ABSTRACT: Multivariate Regression Tree is a tree where univariate response variable has been substituted by a multivariate response variable. It has been proposed to investigate complex ecological data. We apply the Multivariate Regression Tree to investigate social and economical issues in order to comprehend if this method can be generalized and used in different research fields. We apply the Multivariate Regression Tree to identify the causes of death in Italian Counties in 2019. The first results evidence the capacity of Multivariate Regression Tree to define nodes characterized for specific causes of death and to classify together geographical areas with similar impact levels of the variables.

KEYWORDS: Multivariate regression tree; semi-supervised clustering; causes of deaths

1 Introduction

Tree-based methods define a wide set of methodologies finalized to partition the features' space in different areas to realize classification and regression analysis (De'ath & Fabricius, 2000). The aim is to obtain a subset more homogeneous compared to the initial set. A tree can be *univariate* or *multivariate*. The adjective "multivariate" is used to define two approaches. The first is related to the use of more than one attribute in the partition of the observations. The second is characterized by the introduction of the model of one outcome variable composed of more than one level.
In this paper, we focus on the second approach and on its possible use in the economical field. Specifically, we focus on the Multivariate regression tree (MRT) proposed by (De'Ath, 2002). It is *a natural extension of univariate regression trees, with the univariate response of the latter being replaced by a multivariate response* (De'Ath, 2002, p. 1106). The method has been proposed to investigate, describe and predict the relationship between the multi-species

data and the environmental characteristics. It is structured to analyze the community data without making assumptions about the form of relationship between the species and their environment. The nodes identified with MRT are characterized by the presence of a reduced number of species and a habitat with specific environmental characteristics. To our knowledge, this method has been used only to analyze complex ecological data. We attempt to use it to investigate a social, medical and economical issues. More in detail, we apply MRT to comprehend which aspects can impact on the number of deaths in a specific area. We focus on the data of Italian Counties in 2019.

Three sections, besides the introduction, complete this study. Firstly, the MRT methodology has been presented. Secondly, the results have been proposed and, finally, some concluding remarks are highlighted.

2 Methodology

MRT transforms the univariate tree into a multivariate including in the model a multivariate response and redefining the impurity of the node. De'Ath, 2002 has proposed two different measures of impurity. In the first case, MRT operates using an impurity measure called *sums of squared distances* (SSD) and minimizing the SSD of sites from the centroids of the nodes to which they belong. The sum of squares multivariate tree (SS-MRT) has been calculated through the following formula: $\sum_{ij}(x_{ij} - \bar{x_j})$, where x_{ij} is the species data for site i and species j and $\bar{x_j}$ is the mean. The measure can also be calculated considering the median value. In the second case, MRT is built using a dissimilarity matrix and considering the dissimilarities as a distance measure. The nodes are defined as minimizing the intersite sums of squared distances within the clusters. The impurity measure is defined as: $\sum_{i>kk} d_{ik}^2$, where d_{ik}^2 identifies the squared dissimilarities between sites i and k. The MRT built using the first impurity measure can be considered a form of the multivariate regression. Instead, the MRT built using distance measures can be considered a method of constrained clustering, because it allows obtaining clusters that are similar with respect to a measure of species dissimilarity. In both cases, it allows identifying nodes that are characterized for the presence of a reduced number of species and a habitat with specific environmental characteristics.

In this paper, we focus on SS-MRT. We define two different models: the response variable is defined by the number of deaths distinct for disease and gender, the covariates are related to the characteristics of the counties as the percentage of degrees, and the number of specialists, as explained in Table 1

Table 1. *Variables names and description. The AMR (Adjusted Mortality Rate) acronyms suffix denotes the target variable ($\mathcal{Y}$), all other variables are considered as predictors ($\mathcal{X}$).*

Variable name	**Extended label**
AMRm	Adjusted mortality rate from diseases (males)
AMRf	Adjusted mortality rate from system diseases (females)
Doctors rate	Rate of doctors enrolled in professional register
Graduates	Percentage of graduates over population
Eployment rate	Employment rate 15-64 M+F
Pop	Population
Aging index	(Pop65+)/(Pop0-14)*100
% specialists	Percentage of active doctors per indicated specialization in the health system per 10,000 inhabitants
VA	Value added per person (current prices)

3 Conclusion

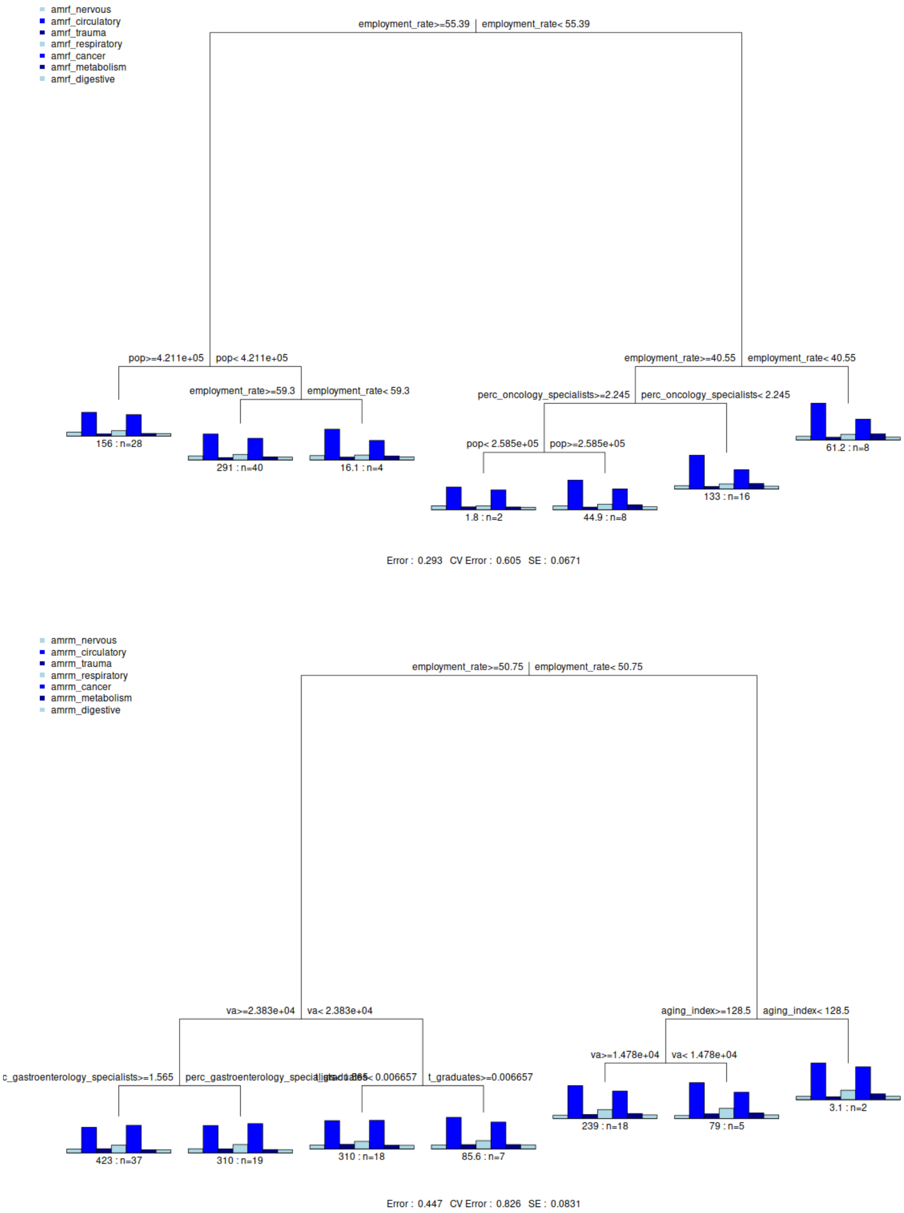
We use the MRT to investigate the elements that can impact the number of death in Italian counties. From a methodological point of view, our study highlights the importance of using advanced statistical methods to analyze the complex dataset and interpret the findings to obtain meaningful insights. From a managerial perspective, our results highlight which aspect can reduce the mortality rates and support the healthcare policy in allocation decisions.

References

De'Ath, Glenn. 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, **83**(4), 1105–1117.

De'ath, Glenn, & Fabricius, Katharina E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**(11), 3178–3192.

Figure 1. *The causes of death of Italian women and men, 2019*


amrf_nervous
amrf_circulatory
amrf_trauma
amrf_respiratory
amrf_cancer
amrf_metabolism
amrf_digestive
employment_rate>=55.39
employment_rate< 55.39
pop>=4.211e+05
pop< 4.211e+05
employment_rate>=59.3
employment_rate< 59.3
employment_rate>=40.55
employment_rate< 40.55
perc_oncology_specialists>=2.245
perc_oncology_specialists< 2.245
pop< 2.585e+05
pop>=2.585e+05
156 : n=28
291 : n=40
16.1 : n=4
1.8 : n=2
44.9 : n=8
133 : n=16
61.2 : n=8
Error : 0.293 CV Error : 0.605 SE : 0.0671
amrm_nervous
amrm_circulatory
amrm_trauma
amrm_respiratory
amrm_cancer
amrm_metabolism
amrm_digestive
employment_rate>=50.75
employment_rate< 50.75
va>=2.383e+04
va< 2.383e+04
aging_index>=128.5
aging_index< 128.5
c_gastroenterology_specialists>=1.565
perc_gastroenterology_specialists< 1.565
t_graduates< 0.006657
t_graduates>=0.006657
va>=1.478e+04
va< 1.478e+04
423 : n=37
310 : n=19
310 : n=18
85.6 : n=7
239 : n=18
79 : n=5
3.1 : n=2
Error : 0.447 CV Error : 0.826 SE : 0.0831