Exploratory GIS Analysis via Spatially Weighted Regression Trees

Analisi esplorativa di dati GIS mediante alberi di regressione pesati spazialmente

Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, and Roberta Siciliano

Abstract The challenging goal of the *Moving Towards Adaptive Governance in Complexity: Informing Nexus Security* (MAGIC) project is to quantitatively enlighten the nexus among energy, food, water and land use toward informed governance inside EU aimed at long term environmental feasibility, economic viability and social desirability. Within the framework of recursive partitioning algorithms by tree-based methods, this paper provides an application on a real Geographic Information System (GIS) dataset regarding the irrigation communities in Almeria, Spain. We propose to build an explorative regression tree – spatially weighted – aiming at classifying the specific consumption of water (per hectare) of the farming communities based on either water management areas and different mix of sources for irrigation water (surface, groundwater, waste water, desalination).

Abstract *Il progetto Moving Towards Adaptive Governance in Complexity: Informing Nexus Security (MAGIC) mira ad illustrare quantitativamente il nesso tra consumi di energia, cibo ed acqua e l'uso del suolo nell'Unione Europea, con lo scopo di sostenerne i processi decisionali orientati alla compatibilità ambientale, fattibilità economica e desiderabilità sociale di lungo periodo. Nel presente lavoro è descritta un'applicazione a dati reali, raccolti in un sistema informativo geografico (GIS) relativo alle comunità irrigue dell'Almeria, inquadrata nell'ottica dei metodi*

Carmela Iorio
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: carmela.iorio@unina.it

Giuseppe Pandolfo
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: giuseppe.pandolfo@unina.it

Michele Staiano
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: michele.staiano@unina.it

Roberta Siciliano
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: roberta@unina.it

di partizione ricorsiva su alberi. La proposta consiste nella costruzione di un albero di regressione esplorativo opportunamente pesato, per classificare il consumo specifico di acqua (per ettaro di superficie irrigua) nelle diverse comunit agricole della regione in base alle aree di appartenenza ed al differente mix di fonti.

Key words: Regression trees, Supervised statistical learning, Spatial methods

1 Motivation

The MAGIC project aims at suitably tackling the nexus among energy, food and water to assess the sustainability as a complex predicate. It features a novel perspective rooted in bioeconomics toward the accounting of technical and environmental resources required to assure the living standards of our societies. In order to inform and steer the processes of decision and policy making inside UE, quantitative contents of narratives about the various theme related to the project – water, energy, food accounts along with labour and land use entanglement, their nexus and scenarios for innovations in the current state of affairs – are to be produced by means of a rigorous and transparent approach to official data, specific domain knowledge and agreed models. Therefore, statistical learning is envisioned as a key tool to lift data to information and up to knowledge, as needed to cope with the complex predicament of sustainability.

Among a set of pilot case studies approached during the first year of the project, some real datasets were exploited as test bench for the fruitful interaction of domain experts and statisticians.

2 A short introduction to exploratory regression trees

Classification And Regression Trees (CART) developed by [1] is a milestone in the evolution and spreading of the tree-based methodology. Recursive partitioning tree procedures adopt a supervised approach: the response variable (of numerical or categorical type) drives the learning process on the basis of a set of predictors (of numerical or categorical type). For this reason CART methodology can be viewed as precursor of supervised statistical learning introduced by [10] and outlined by [5] . Tree-based methods have been proposed for both prediction and exploratory purposes. In the exploratory context, binary segmentation can be understood as a recursive partitioning of objects into two subgroups due to some splitting variables derived from available predictors such to obtain internally homogeneous and externally heterogeneous subgroups with respect to a target or response variable [6]. The final result is a binary tree visualizing the dependence relationship between the response variable and the predictors.

In regression trees, the splitting criterion at each non terminal node can be to maximize the decrease of impurity of the response variable within the two sub-nodes, where the impurity is a measure of variation for numerical responses. Nodes are declared to be terminal on the basis of a stopping rule. Terminal nodes include disjoint and homogeneous subgroups of objects, defining a partition of the starting group of objects at the root node with respect to the response variable. The tree with only one split at the root node is called stump. It describes the best partition of the objects into just two subgroups such to minimize the internal variation of the target variable within the two nodes. The quality of any tree can be measured by an overall impurity measure of the tree. In regression trees, this is calculated by the sum of the variation measures of the target variable within its terminal nodes. Typically, the overall impurity of any tree is compared with respect to the impurity at the root node: the ratio of the two quantities provides a relative cost reduction measure named deviance. By definition of splitting criterion, the deviance reduces as the number of terminal nodes or tree size increases. For a more detailed literature study one can refer to [7, 8, 9, 3, 2] and to the references therein.

3 A Real Problem

The analysis has been performed on a data set collected in 2013 and related to the all the irrigation communities in Almeria which are grouped geographically by 18 water management areas and technically by 38 distinct water sources patterns. Raw GIS data matrix consists of 376 instances, designated as irrigation communities. As the entire population is surveyed, the analysis is only exploratory, not confirmatory nor predictive. Exploratory trees belong to data mining methods where also visualization helps the analyst to better understand the phenomena [4]. We built an exploratory regression tree spatially weighted aiming at classifying the specific consumption of water per hectare of the irrigation communities base on either water management area (coded as in Table 1), and sources of water used (surface, groundwater, waste water, desalination), grouped in 8 different profiles described in Table 2. An expanded regression tree with the purest terminal nodes, accordingly to an initial stopping rule, has 48 leaves. Since this structure is rather complex to be interpreted, a simpler structure of the regression tree can be identified. One can fix

Table 1: Water Management Area (WMA) codes

WMA	**Code**	**WMA**	**Code**	**WMA**	**Code**
Alpujarra	a	Campo de Tabernas	g	Medio Almanzora	m
Alto Almanzora	b	Comarca de Guadix	h	Medio Andarax	n
Alto Andarax	c	El Saltador	i	Nacimiento	o
Bajo Almanzora	d	Higueral de Tjola	j	Poniente	p
Bajo Andarax	e	Los Guiraos	k	Riegos de Pulp	q
Campo de Njar	f	Los Vlez	l	Z.R. Cuevas de Almanzora	r

Table 2: Profiles codes

Profile	**Code**
80-99% surface, remaining groundwater	1
surface water only	2
groundwater only	3
60-80% surface, remaining groundwater	4
40-60% surface, remaining groundwater	5
10-39% surface, remaining groundwater	6
1-10% surface, remaining groundwater	7
remaining profiles (reused and desalted)	8

a certain threshold value for the overall impurity, then select among the sub-trees not exceeding the fixed threshold value the one with minimum number of leaves. Specifically, Fig. 1 shows that the deviance reduction is marginal for the trees with more than 9 leaves. Consequently, we chose the regression tree with 9 terminal nodes as final tree. We named it "intermediate regression tree" and its representation is displayed in Fig. 2. The labels set in Fig. 2 at any internal node indicate the splitting variables with those categories inducing the objects to the left sub-node; at the leaves of the tree are reported the mean values of the target variable (cubic meters of water consumed in one year per hectar of farming surface). Fig. 2 points out that the mean response values increase from the left-hand side to the right-hand side. This clearly conveys to the domain expert the information that the irrigation communities belonging to the leftmost leaves consume less water per hectare than those belonging to the rightmost terminal nodes. By watching the mean fitted values of the rightmost terminal leaves it can state that the WMA belonging to both Campo de Nijar and Poniente are the more water intensive consumption areas. Specifically, the irrigation communities belonging to Poniente has the highest mean value of water use per hectare. The simplest result of analysis reduces the tree to a stump with the only splitting rule due to WMA. Thus, the stump partitions the irrigation communities into two regions of predictor space. The resulting partition singles out the most water intensive management areas. On the right-hand side, there are the management areas with lower values of water consumed per hectare (mean value equals to 3377); on the left-hand side, there are the remaining management areas, Campo de Njar and Poniente, with higher values of water consumed per hectare (mean value equals to 6199). Thus, the stump reveals that the most important factor associated with water consumption per hectare is due the water management areas. Fig. 3 displays a geographic visualization of the stump (the map was produced elaborating the R output by open source GIS software Q-GIS and includes a satellite imagery title obtained by BING).

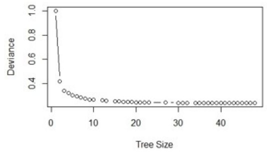

Fig. 1: The overall deviance reduction of the regression tree with increasing the number of leaves

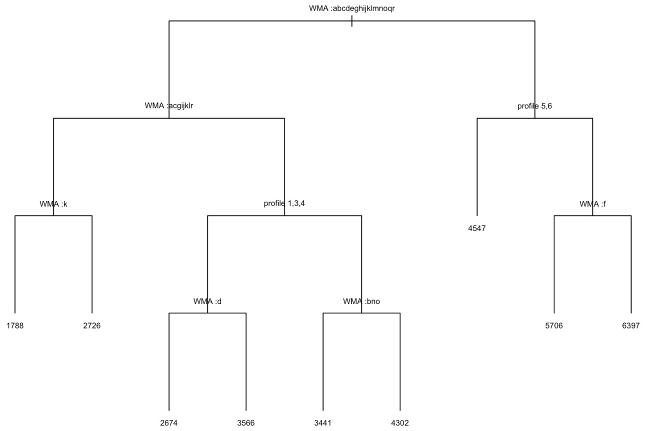

Fig. 2: The intermediate regression tree for Almeria data set.

4 Final remarks

In the framework of MAGIC project, this paper was designed to deal with a real problem of statistical analysis. We analyzed the water consumption by irrigation communities in Almeria. The regression tree is a device simple to be presented and understood; nevertheless, it required a careful tailoring of the standard approach (being the observations to be spatially weighted) and a trade off between easiness and richness of the representation should be agreed in order to choose the number of leaves. The "intermediate regression tree" enables to highlight some key features in the data set. Indeed, the leaves resulted ordered from left to right hand side of the tree with increasing levels of average water consumption per hectare: this classification is a good starting point for deepening the analysis. The water consumption per hectare in some management areas is quite more spread than in some others; this clearly depend on the types of crops and cropping methods along with the irrigation means and water sources, so grafting current data to other information could offer a

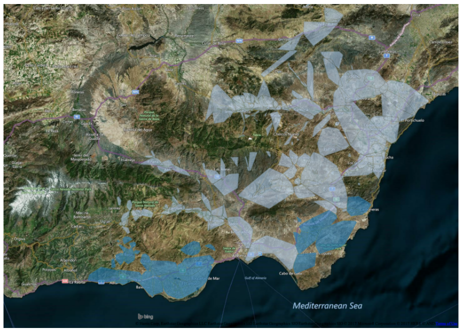

Fig. 3: GIS visualization of the STUMP (the blue areas mark the WMAs with higher consumption of water per hectare Campo de Nijar and Poniente compared to the gray ones).

better insight. The role of profiles of sources varies in different sets of water management areas, so we could technically say that the two predictors interact: this is the clue for a further investigation. A richer GIS approach to the problem could be beneficial for the analysis, so collecting more layers of data and comparing spatially any pattern could strenghten the knowledge discovery process.

Acknowledgements MAGIC project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 689669.
The present work reflects only the authors' view and the Funding Agency can not be held responsible for any use that may be made of the information it contains.

References

1. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: Classification and regression trees. CRC press (1984).
2. D'Ambrosio, A., Aria, M., Iorio, C., Siciliano, R.: Regression trees for multivalued numerical response variables. Expert Systems with Applications **69**, 21–28 (2017)
3. D'Ambrosio, A., Heiser, W.J.: A recursive partitioning method for the prediction of preference rankings based upon kemeny distances. Psychometrika **81**(3), 774–794 (2016)
4. Fayyad, U.M., Wierse, A., Grinstein, G.G.: Information visualization in data mining and knowledge discovery. Morgan Kaufmann (2002).
5. Friedman, J., Hastie, T., & Tibshirani, R.: The elements of statistical learning. New York: Springer series in statistics (2001).
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media (2009)
7. Mola, F., Siciliano, R.: A two-stage predictive splitting algorithm in binary segmentation. In: Computational statistics, pp. 179-184. Springer (1992)
8. Mola, F., Siciliano, R.: A fast splitting procedure for classification trees. Statistics and Computing **7**(3), 209–216 (1997)
9. Siciliano, R., Mola, F.: Multivariate data analysis and modeling through classification and regression trees. Computational Statistics & Data Analysis **32**(3), 285–301 (2000)
10. Vapnik, V.N.: The nature of statistical learning theory (1995).