ADVANCES IN FOREST FIRE RESEARCH

2022

Edited by
DOMINGOS XAVIER VIEGAS
LUÍS MÁRIO RIBEIRO

Analyzing the EU forestry sector to seek new market opportunities using Minimum Spanning Tree based clustering analysis

Jongmin Han[1*]; Abílio Pereira Pacheco[2]; José Coelho Rodrigues[3]

[1,2,3] INESC TEC and Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 378, 4200-465, Porto, Portugal
[2]ForestWISE - Collaborative Laboratory for Integrated Forest and Fire Management, Quinta de Prados, 5000-801 Vila Real, Portugal
{jongmin.han, app, jpcr}@fe.up.pt

**Corresponding author*

Keywords

Clustering analysis; MST based clustering; selection of variables; EU 27 classification; European forest

Abstract

To enhance the economic viability and address the labour shortage in the forestry industry, alternative solutions using robotization and automation are emerging. However, due to technological barriers and lack of solid business models, successful commercialization in the forestry sector is yet to be challenging. As an initial market analysis for developing a business model for new forestry machineries, this study was conducted to reveal clusters of EU countries to seek the potential market opportunities outside of Portugal. To identify similar market conditions and restrictions, EU countries were clustered using Minimum Spanning Tree (MST) based clustering analysis clustering algorithm and selection of variables while considering the geographic, economic, and social conditions of each country.

1. Introduction

Robotization and automation of forestry machinery are emerging as an alternative to address significant challenges in the forestry industry in terms of economic viability, safety, labour shortages, and environmental performance (Lideskog, Karlberg, and Bergsten 2015; Couceiro et al. 2019)). However, such a technological shift has not yet had significant commercialization in the forestry sector except for logging due to technological challenges and a lack of solid business models (Teece 2010; Lideskog, Karlberg, and Bergsten 2015). A well-developed business model is essential for increasing the value of technology and opportunities into economic outputs through customers and markets leading to the success of a business (Chesbrough 2002; 2010). Thoroughly analyzing the main components of a business model is an essential process to create a solid framework for developing the business model (Couceiro et al. 2019). Especially, determining the value proposition and market segments are crucial in measuring and predicting the potential of new products and services.

In most developed countries, the forestry sector is highly internationalized, concerned with its sustainability, diversity and increasing complexity, responding to the growing competition (Melo, Cunha, and Ferreira 2017). Although the forestry industry in Portugal is rich in opportunities, mainly due to the abandonment of rural and deprived areas and the lack of investment in maintenance and conservation, the potential has not yet been fully exploited (Melo, Cunha, and Ferreira 2017). Different countries around the world have distinct levels of economic, social, and environmental issues in forestry due to several factors that each country faces, e.g. geographic characteristics, economic capabilities. The possibility of such issues which may directly lead to potential risks is influenced by the situations that a given country has. For example, damages from natural disasters may vary depending on a country's risk management capabilities. According to literature (Toya and Skidmore 2007), countries with higher income and educational attainment, greater openness, more complete financial systems, and smaller government are more likely to experience fewer losses. Thus, there is a need to investigate the geographic, economic, and social features to draw implications in a variety of scenarios, including potential market opportunities. In this sense, the use of different indicators can be a solution for analyzing the current situation of the forest sector at a national level.

As an initial phase for building a business model for new forestry machines, this study was conducted to reveal clusters of EU countries to seek the potential market opportunities outside of Portugal. To identify similar market conditions and opportunities, EU countries were clustered using a hierarchical clustering algorithm, an unsupervised learning branch from the Artificial Intelligence.

2. Methodology

2.1. Data collection

The main statistical data source used for the analysis was from major international organizations such as the Food and Agricultural Organization of the United Nations (FAO). The FAO statistical databases (FAOSTAT) platform is freely accessible and contains the largest statistical database on agriculture, fishery, and forestry in the world, with approximately 20,000 indicators covering 245 countries and territories, with around 2,000,000 users each year (FAOSTAT, 2022). The FAOSTAT database is used widely in peer-reviewed literature, serving as the basis for many Agriculture, Forestry, and Other Land Use (AFOLU)-related analyses (Tubiello et al. 2013). FAOSTAT is useful to make the cross-country comparisons because it has the user-friendly interface and the possibility to track the different items that we need in one place. To reflect the economic, social, and environmental issues in the forestry sector, twenty-seven indicators for 2015 (year) were selected related to forest production, land use, and social capital in FAOSTAT. The selected indicators had to respect the following criteria: (i) be fact based; (ii) be based on available data for twenty-seven EU countries; (iii) be easily interpreted. In creating the final list of indicators, new ad hoc indicators were created. One of the ad hoc indicators was, for example, "The production quantity of Pulpwood (m^3) per Forest Land (1000 ha)"

2.2. Data pre-processing

As the variables were reported in different units, to determine distances between the EU countries, it was necessary to standardize them and eliminate dependence. According to (Arabie et al. 1996), when the number of variables is large, there is a possibility that variables that do not contribute to cluster classification exist, and these variables may interfere with finding a cluster structure. Finding structure in a high-dimensional variable space with a small dataset is a general problem in both classification and cluster analysis. This may be due to the presence of several "noisy" noninformative variables and/or redundant features from strongly correlated or more generally strongly dependent variables that may produce, for instance, multicollinearity (Fraiman, Justel, and Svarc 2008). The basic methodology for the detection of redundant variables is correlation analysis under unsupervised dataset that there are no output variables to predict. The most used measures are Pearson correlation coefficient for a linear correlation, and Spearman and Kendall correlation coefficient for a nonlinear correlation (Marshall 1996; Chok 2010). According to previous research, the use of Pearson correlation coefficient is not robust in forms of associations other than linear or normal associations (Kowalski 1972; Speed 2011).

2.3. Clustering analysis based on Minimum Spanning Tree (MST)

Hierarchical cluster analysis as a multivariate statistical tool has also been widely used to group the data by simultaneously clustering objects and variables (Newman 2004). Hierarchical clustering has the advantage that it does not require a knowledge on the number or size of groups to look for beforehand. However, it does not tell us how many groups should be used to get the best division (Jain, Murty, and Flynn 1999; Xu and II 2005; Fortunato 2010). Another problem is the lack of the ability to detect clusters which are not defined by regular geometric curves (Grygorash, Zhou, and Jorgensen 2006; Zhong, Miao, and Fränti 2011; Wu et al. 2013; Tewarie et al. 2015) Detecting clusters with irregular boundaries has become a research interest in recent years. In particular, the clustering algorithm using the Minimum Spanning Tree (MST) has recently attracted a lot of attention (Gower and Ross 1969). MST is a tree that minimizes the total weight or lengths of the edges of the tree under a weighted, undirected graph.

3. Results and Discussion

As a result of the Shapiro-Wilk test for normality, twenty-nine indicators in 2010 and twenty-six indicators in 2015 did not follow a normal distribution. With large enough sample sizes (> 30 or 40), the violation of the

normality assumption should not cause major problems (Ghasemi and Zahediasl 2012). However, the dataset which this research dealt with is small sample size and its normality is violated. Thus, for variable reduction, the Kendall correlation measure was employed. It is known to be more robust and slightly more efficient than the Spearman correlation coefficient (Croux and Dehon 2010).

Following the above-mentioned methodology, the first step was the elimination of some higher correlated dependent and statistically significant variables (significance level = 5%) for the selection of appropriate features. The three indicators, such as "Share of Net value added of forestry in total net value added", "Proportion of naturally regenerating forest in forest area", "Proportion of land under meadows and pastures in agricultural area", consequently had been eliminated by correlation analysis. Finally, the analysis enabled the grouping of twenty-seven EU countries into clusters using thirty-seven different variables. When analyzing the similarity without SDG indicators between EU countries, twenty-nine indicators were used except eight SDG indicators.

Before correlation analysis, normalization (Min-Max scaling) was used to make variables comparable to each other on equal grounds. The variables were shift and rescaled so that they end up ranging between 0 and 1. The method did not change the shape of the distribution of data while adjusting the values. If the shape of the distribution is changed, it introduces bias into analysis because the distribution of a lot of collected data is not Gaussian. Applying Prim's algorithm for shaping a complete graph based on Manhattan distance among monitored features into MST. Finally, a network community detection technique was applied to identify the clusters of EU countries.

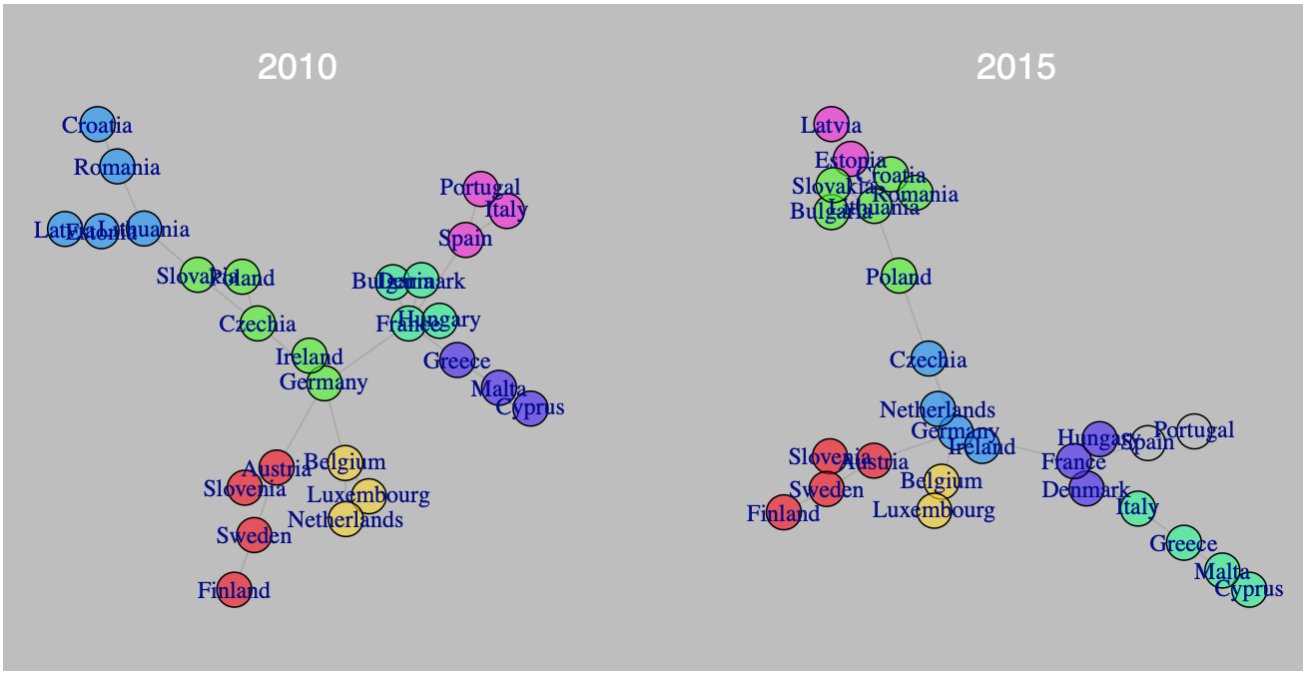

Figure 1- ***Network communities of EU countries with MST based on the Manhattan distance, and each cluster is identified by a special color***

Our analysis of the different indicators related to forestry through clustering analysis based on MST was performed on the EU27 countries for two time periods 2010 and 2015. Fig. 1 shows that there is a difference in the network community composition depending on different periods. There were six clusters created, representing the seven groups of EU-27 countries in 2010, while eight clusters were created in 2015.

Table1. Sorting countries into clusters using dataset 2010

1	2	3	4	5	6	7
Austria	Belgium	Czechia	Bulgaria	Croatia	Cyprus	Italy
Finland	Luxembourg	Germany	Denmark	Estonia	Greece	Portugal
Slovenia	Netherlands	Ireland	France	Latvia	Malta	Spain
Sweden		Poland	Hungary	Lithuania		
		Slovakia		Romania		

The results of year 2010 are presented by two clusters of five members, two of four members and three of three members (Table 2). On the other hand, the results of year 2015 are presented by one cluster of three members, one of five members, two of three members and three of three members (Table 2). The group consisting of Austria, Finland, Slovenia, and Sweden did not change during the two periods in 2010 and 2015, and Portugal and Spain, which were in the same group in 2010, were analyzed as belonging to the same group in 2015 as well. On the other hand, in Italy and Luxembourg, the group to which they belonged also changed over time.

Table 2. Sorting countries into clusters using dataset 2015

1	2	3	4	5	6	7	8
Austria	Belgium	Bulgaria	Cyprus	Czechia	Denmark	Latvia	Portugal
Finland	Luxembourg	Croatia	Greece	Germany	France	Estonia	Spain
Slovenia		Lithuania	Malta	Ireland	Hungary		
Sweden		Poland	Italy	Netherlands			
		Romania					
		Slovakia					

4. References

Arabie, P, L J Hubert, G De Soete, and Glenn W Milligan. 1996. "Clustering and Classification," 341–75. https://doi.org/10.1142/9789812832153_0010.

Chesbrough, Henry. 2002. "The Role of the Business Model in Capturing Value from Innovation: Evidence from Xerox Corporation's Technology Spin-off Companies." *Industrial and Corporate Change* 11 (3): 529–55. https://doi.org/10.1093/icc/11.3.529.

———. 2010. "Business Model Innovation: Opportunities and Barriers." *Long Range Planning* 43 (2–3): 354–63. https://doi.org/10.1016/j.lrp.2009.07.010.

Chok, Nian Shong. 2010. "Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data."

Couceiro, Micael S., David Portugal, João F. Ferreira, and Rui P. Rocha. 2019. "SEMFIRE: Towards a New Generation of Forestry Maintenance Multi-Robot Systems." *2019 IEEE/SICE International Symposium on System Integration (SII)* 00: 270–76. https://doi.org/10.1109/sii.2019.8700403.

Croux, Christophe, and Catherine Dehon. 2010. "Influence Functions of the Spearman and Kendall Correlation Measures." *Statistical Methods & Applications* 19 (4): 497–515. https://doi.org/10.1007/s10260-010-0142-z.

FAOSTAT. 2022. Database on Agriculture. Food and Agriculture Organization of the United Nations. Rome, Italy. Food and Agriculture Organization of the United Nations. 2022. https://www.fao.org/faostat/en/#home.

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486 (3–5): 75–174. https://doi.org/10.1016/j.physrep.2009.11.002.

Fraiman, Ricardo, Ana Justel, and Marcela Svarc. 2008. "Selection of Variables for Cluster Analysis and Classification Rules." *Journal of the American Statistical Association* 103 (483): 1294–1303. https://doi.org/10.1198/016214508000000544.

Ghasemi, Asghar, and Saleh Zahediasl. 2012. "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians." *International Journal of Endocrinology and Metabolism* 10 (2): 486–89. https://doi.org/10.5812/ijem.3505.

Gower, J. C., and G. J. S. Ross. 1969. "Minimum Spanning Trees and Single Linkage Cluster Analysis." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 18 (1): 54–64. https://doi.org/10.2307/2346439.

Grygorash, Oleksandr, Yan Zhou, and Zach Jorgensen. 2006. "Minimum Spanning Tree Based Clustering Algorithms." *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, 73–81. https://doi.org/10.1109/ictai.2006.83.

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. "Data Clustering: A Review." *ACM Computing Surveys (CSUR)* 31 (3): 264–323. https://doi.org/10.1145/331499.331504.

Kowalski, Charles J. 1972. "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 21 (1): 1–12. https://doi.org/10.2307/2346598.

Lideskog, Håkan, Magnus Karlberg, and Urban Bergsten. 2015. "Development of a Research Vehicle Platform to Improve Productivity and Value-Extraction in Forestry." *Procedia CIRP* 38: 68–73. https://doi.org/10.1016/j.procir.2015.07.014.

Marshall, Albert W. 1996. "Copulas, Marginals, and Joint Distributions." *Lecture Notes-Monograph Series* 28.

Melo, A., J. Cunha, and P. Ferreira. 2017. "Business Model for Forest Management." *Procedia Manufacturing* 13: 940–47. https://doi.org/10.1016/j.promfg.2017.09.164.

Newman, M. E. J. 2004. "Detecting Community Structure in Networks." *The European Physical Journal B* 38 (2): 321–30. https://doi.org/10.1140/epjb/e2004-00124-y.

Speed, Terry. 2011. "A Correlation for the 21st Century." *Science* 334 (6062): 1502–3. https://doi.org/10.1126/science.1215894.

Teece, David J. 2010. "Business Models, Business Strategy and Innovation." *Long Range Planning* 43 (2–3): 172–94. https://doi.org/10.1016/j.lrp.2009.07.003.

Tewarie, P., E. van Dellen, A. Hillebrand, and C.J. Stam. 2015. "The Minimum Spanning Tree: An Unbiased Method for Brain Network Analysis." *NeuroImage* 104: 177–88. https://doi.org/10.1016/j.neuroimage.2014.10.015.

Toya, Hideki, and Mark Skidmore. 2007. "Economic Development and the Impacts of Natural Disasters." *Economics Letters* 94 (1): 20–25. https://doi.org/10.1016/j.econlet.2006.06.020.

Tubiello, Francesco N, Mirella Salvatore, Simone Rossi, Alessandro Ferrara, Nuala Fitton, and Pete Smith. 2013. "The FAOSTAT Database of Greenhouse Gas Emissions from Agriculture." *Environmental Research Letters* 8 (1): 015009. https://doi.org/10.1088/1748-9326/8/1/015009.

Wu, Jianshe, Xiaoxiao Li, Licheng Jiao, Xiaohua Wang, and Bo Sun. 2013. "Minimum Spanning Trees for Community Detection." *Physica A: Statistical Mechanics and Its Applications* 392 (9): 2265–77. https://doi.org/10.1016/j.physa.2013.01.015.

Xu, Rui, and Donald Wunsch II. 2005. "Survey of Clustering Algorithms." *IEEE Transactions on Neural Networks* 16 (3): 645–78. https://doi.org/10.1109/tnn.2005.845141.

Zhong, Caiming, Duoqian Miao, and Pasi Fränti. 2011. "Minimum Spanning Tree Based Split-and-Merge: A Hierarchical Clustering Method." *Information Sciences* 181 (16): 3397–3410. https://doi.org/10.1016/j.ins.2011.04.013.