A Feature-Based Correlation Approach for Analyzing Price Trends of Citrus in Valencia Province (Spain)

R. Arnau, J. M. Calabuig, N. Ortigosa and L. Petrosyan
Instituto Universitario de Matemática Pura y Aplicada,
Universitat Politècnica de València, Camino de Vera s/n, 46022,
Valencia, Spain
e-mail: ararnnot@posgrado.upv.es, jmcalabu@mat.upv.es, nuorar@upvnet.upv.es, luipet@inf.upv.es

***Abstract*—In this paper, we are going to present some work in progress results to study and analyze the price variation among different citric varieties in Valencia province from Comunitat Valenciana region (Spain). A data-driven approach is used to represent each citrus variety and season using 5 features for comparing its prices trends using a correlation analysis. Those findings provide the foundation for implementing clustering algorithms, such as k-Medoids, to classify citrus varieties and seasons based on profitability and market behavior.**

Keywords-Citrus; price trends; feature extraction; correlation analysis.

I. Introduction

The agriculture sector plays an important role in achieving multiple Sustainable Development Goals (SDGs). Its impact ranges from eradicating hunger (SDG2) and poverty (SDG1), to protecting the environment (SDG12) and promoting health (SDG3). Thus, the transformation towards sustainable agriculture is essential to achieve these goals [1], [2]. In this sense, the citrus fruit sector is of great importance both in the Valencia Region (Comunitat Valenciana) and in Spain. Indeed, the Comunitat Valenciana is the main citrus fruit producing region in Spain. In 2022/2023, the production was around 67% of Spain's citrus (the main varieties grown being oranges, mandarins, and lemons) with a production of around 2.8 million tonnes. In this period, 1.8 million tonnes were exported (out of 4.3 million tonnes in Spain) with an economic value of around 2 million euros (out of 3.4 million euros in Spain) [3], [4].

On the other hand, data science and Artificial Intelligence are tools that, as in other areas, are already being used in the agriculture sector [5], [6]. In this sense, it is important to highlight that the interpretability of Artificial Intelligence (AI) algorithms in general, and of machine learning algorithms in particular, is crucial when models are to be used for decision-making. Thus, interpretability not only increases confidence in the models (and their results) but also their validation. One of the most common and easy-to-interpret clustering algorithms is k-Means. In these algorithms, it is common to use the Euclidean distance because it is easy to understand the proximity of the vectors for the formation of the clusters. However, other distances (such as that associated with the correlation between vectors) and similar algorithms (such as k-Medoids) can be used.

Citrus prices can experiment changes across seasons and between seasons, depending on factors such as weather conditions, supply and demand, and production costs, among others, see for instance [7]. In this work, we study the price trends of the citrus fruits in the Valencia province over the period 2015-2022. For this purpose, we identify different varieties of these citrus fruits by means of a vector consisting of 5 features. A correlation analysis of these vectors then allows us to study the trends, which allows us to classify citrus fruits according to their profitability.

In a future work, this first trend analysis will allow us to implement clustering algorithms (such as the aforementioned k-Medoids) based on correlations. This analysis could be a valuable tool for planning, decision-making, risk mitigation and revenue optimization in the agricultural sector. It provides key information that can be used by producers, traders, policy makers, investors and other market actors to improve the efficiency and economic stability of the sector and, as mentioned above, a tool to achieve some of the SDGs.

The rest of the paper is structured as follows. Section II presents the database and the methodology used for this analysis. Obtained results are presented in Section III, while the analysis limitations, discussion and conclusions are drawn in Section IV. Finally, the stages of upcoming research are presented in Section V.

II. Materials and Methods

The starting point (and usually critical point) in all data science projects is the collection of data. In our case, we use the “Reports of the Valencian Agricultural Sector" (Informes del Sector Agrario Valenciano, in Spanish), which contain temporary information (with annual breakdown) on the agri-food situation in the Comunitat Valenciana since 2015 (actually since 1998). These reports are divided into different chapters containing both meteorological data (Chapter III) and agricultural (Chapter IV) and livestock (Chapter V) statistics, as well as agricultural prices (Chapter VI). We would like to point out here that these reports consist of 20 documents in Excel format for each of the available years, which means that they are not easy to use [8]. Fortunately, a file (in `csv` format) containing much of the necessary information can be found in [9]. This file contains weekly prices of agricultural products (not only citrus fruits) since 2017. Together with the files downloaded in [8] the data have been completed to have

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-324-8

information since 2015 (from earlier dates there was missing data). Hence, we have filtered the data (to have only citrus fruits), as well as transformed the dates. So, at the end, we have a database containing 6745 rows corresponding to the weekly price of 9 citrus products (different kinds of Oranges such as, for instance, Navel, Blancas, etc., but also Lemons and Clementines). The 9 products have 43 different varieties. Unfortunately, there were some varieties that were not priced for many weeks (even years) so we decided not to use them and only use the data of 27.

After data curation, several features were extracted in order to form a vector that will characterize each year, so that annual trends are analysed to facilitate subsequent decisions taken by producers.

In this work, we have focused on 5 data features extracted for each analysed year. Even so, we are currently analysing and working on some additional ones so that we can obtain a more custom-made characterization. Since citrus products' season in Valencia region generally starts in September and ends in July, we are not focusing on natural years. From now on, we will refer to the year of the month when the product season started as year indicator.

As each citrus product has a different season length, and we have the data information of prices sampled by weeks, the first feature to be considered is the duration d (in number of weeks) that each specific product season lasted each year. The second and the third features were the minimum (m) and the maximum (M) prices paid each season per kilo of product. Another feature to be considered was the variance, calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}, \tag{1}$$

where x_i is the price paid for the product on the i-th week of the season and N is the number of weeks that we have data information per season. Finally, we consider the coefficient

$$Q = \frac{M - m}{d}, \tag{2}$$

calculated per season, so that a measure of the distribution of the data is also considered.

With the purpose of comparing the tendencies and behaviours of the product prices during each season, we study the correlation of the vectors formed by the extracted features. That is, if $X = (x_1, x_2, \dots, x_5)$ and $Y = (y_1, y_2, \dots, y_5)$ represent a fixed variety and a season each, we compute

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{5}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{5}(x_i - \bar{x})^2 \sum_{i=1}^{5}(y_i - \bar{y})^2}}. \tag{3}$$

A correlation value close to 1 indicates a strong linear correlation between observations, while a value close to -1 signifies a linear correlation, but in opposite directions. A value near 0 suggests that the observations are not linearly correlated. This measure, related to the so called cosine similarity, allows us to identify the relation between price tendencies.

III. Results

When computing the previous features, the 1513 prices observations from [8] and [9] result in a dataframe consisting of 27 varieties of citrus, with its correspondent features for each seasons from 2015-2016 to 2020-2021, since the 2021-2022 data is uncompleted and not used. After removing the observations for which a variety in a concrete season has less than 5 prices recorded, only 19 varieties from Valencia province are considered, each one corresponding to Oranges, Mandarins, and Clementines varieties in a season, and the 5 features extracted. As an example, the first 10 rows can be seen in Table I, where the duration is computed as a portion of a year instead of number of weeks, for convenience.

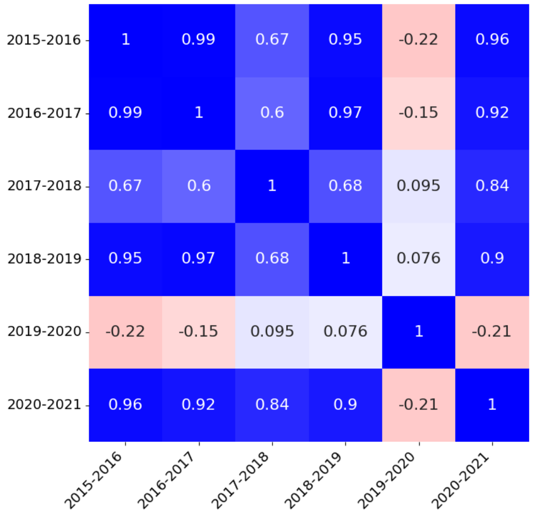


Figure 1. Autocorrelation of features corresponding to “Clausellina/Okitsu" across seasons.

Some correlation results can be seen in Figures 1 and 2. In Figure 1 we show the correlation of “Clausellina/Okitsu" variety, for which the season 2019/2020 is totally uncorrelated with the other 5 studied, probably due to the COVID-19 pandemic, which altered the prices at the end of the season. Also, 2017-2018 season has a tendency of its own, while the other four seasons are fully correlated with each other. In Figure 2, the relation between different varieties is shown. For example, it could be thought that oranges and clementines would present similar trends among them, but one can see that clementine “Clementina Clemenules" variety has a more similar price tendency to oranges than other clementines varieties, which would be the beforehand expected behaviour.

IV. Discussion and Conclusions

In this paper, we have presented some work in progress results that show how weekly prices can be analyzed in order to find significant patterns and correlations. This is done across different varieties of citrus in Valencia province in seasons from 2015 to 2021. We have found that each variety can be identified with a numeric vector of 5 features, which allows for the comparison of varieties with different numbers of recorded prices. This vector can be used as a tool to reveal the behavior of several citrus varieties along different seasons and two comparisons of its prices trends are shown.

However, it is necessary to recognize some limitations. Seasonal variations influenced by external factors, such as the COVID-19 pandemic, have affected the consistency of correlations for certain varieties, as exemplified by the 2019/2020 season of the “Clausellina/Okitsu" variety. While this effect is shown in the study, those external disruptions may make it necessary to integrate additional methods, such as time series forecasting, to adapt the study to irregular market conditions. Moreover, correlation-based similarity, while effective in capturing relative trends, may lack the intuitive clarity of Euclidean-based clustering, resulting in a less interpretable methodology.

To our knowledge, there are no similar studies in the context of agricultural price analysis, although clustering studies (comparing the aforementioned k-Medoids) do exist in the field of identification of management zones in precision agriculture [10].

V. Future Work

The vector identification found will be first step to find an AI-based model that can help to classify products in terms of profitability. More specifically, with the feature vector that identifies each product and the correlation metric (actually 1 minus the correlation to make it a dissimilarity function), we can apply clustering techniques such as k-Medoids, which is part of the on-going work. While it is true that this may produce some loss of interpretability, this technique may be useful: (1) for being less sensitive to differences in absolute magnitudes between features and, (2) to obtain a pattern of relationships between variables, rather than in absolute distances (as seems to be our case).

In this future work, other metrics added to the correlation as well as time-series analysis will be considered, including an extension of the database to include more features and products, which would be one limitation of the current study.

Acknowledgments

Authors acknowledge the support of “Ayuda a Primeros Proyectos de Investigación (PAID-06-22), Vicerrectorado de Investigación de la Universitat Politècnica de València (UPV)”. The first author was supported by “Programa de Ayudas de Investigación y Desarrollo (PAID-01-21), Universitat Politècnica de València (UPV)”.

References

[1] G. Hurduzeu, R. L. Pânzaru, D. M. Medelete, A. Ciobanu, and C. Enea, “The development of sustainable agriculture in eu countries and the potential achievement of sustainable development goals specific targets (SDG 2),” *Sustainability*, vol. 14, no. 23, p. 15 798, 2022, ISSN: 2071-1050. DOI: 10.3390/su142315798.

[2] C. M. Viana, D. Freire, P. Abrantes, J. Rocha, and P. Pereira, “Agricultural land systems importance for supporting food security and sustainable development goals: A systematic review,” *Science of The Total Environment*, vol. 806, p. 150 718, 2022, ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2021.150718.

[3] The Valencian Institute for Business Competitiveness - Institut Valencià de Competitivitat Empresarial (IVACE), *Cítricos de la Comunitat Valenciana*, https://www.ivace.es/index.php/es/component/weblinks/weblink/545-internacional-documentos/548-sectores/357-citricos?Itemid=100096&task=weblink.go, 2023. Retrieved: October, 2024.

[4] Ministry of Agriculture, Fisheries and food - Ministerio de Agricultura, Pesca y Alimentación (MAPAMA), *Análisis de la Campaña de Cítricos 2023/2024*, https://precioscitricos.com/wp-content/uploads/2024/06/informecampana2023-24citricossept-feb_tcm30-684049.pdf, 2024. Retrieved: October, 2024.

[5] R. Khan, N. Dhingra, and N. Bhati, “Role of artificial intelligence in agriculture: A comparative study,” in *Transforming Management with AI, Big-Data, and IoT*, F. Al-Turjman, S. P. Yadav, M. Kumar, V. Yadav, and T. Stephan, Eds. Cham: Springer International Publishing, 2022, pp. 73–83, ISBN: 978-3-030-86749-2. DOI: 10.1007/978-3-030-86749-2_4.

[6] D. Mhlanga, “The Role of FinTech and AI in Agriculture, Towards Eradicating Hunger and Ensuring Food Security,” in *FinTech and Artificial Intelligence for Sustainable Development: The Role of Smart Technologies in Achieving Development Goals*. Cham: Springer Nature Switzerland, 2023, pp. 119–143, ISBN: 978-3-031-37776-1. DOI: 10.1007/978-3-031-37776-1_6.

[7] R. J. S. Izquierdo, G. García-Martínez, N. Lajara-Camilleri, and G. Orea-Vega, “Price evolution of ’clemenules’ and ’navelina’ in spain during the period 2007-2012,” 2015.

[8] Valencian Department of Agriculture, Water, Stock and Fisheries - Conselleria de Agricultura, Agua, Ganadería y Pesca, *Informes del Sector Agrario Valenciano*, https://portalagrari.gva.es/es/pye/informes-del-sector-agrario-valenciano, 2024. Retrieved: October, 2024.

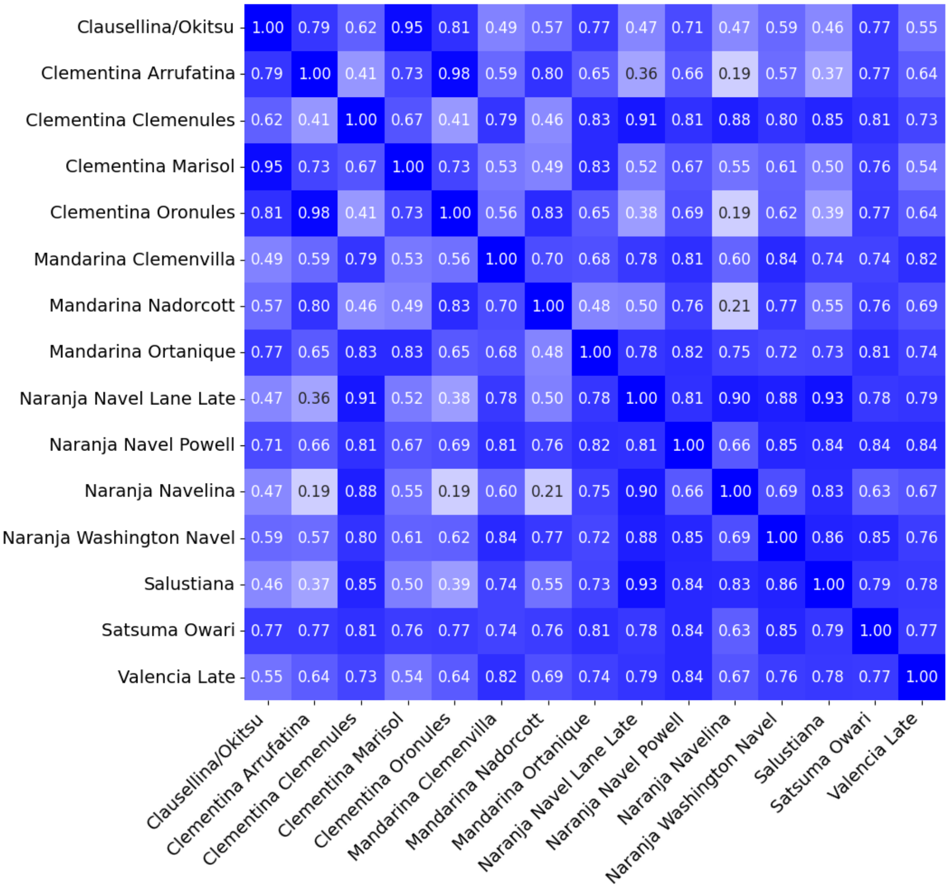
[9] Ministry for the Digital Processing and civil service - Ministerio para la transformación digital y de la función pública, *Precios agrarios de origen de la Comunitat Valenciana*, https://datos.gob.es/es/catalogo/a10002983-precios-agrarios-de-origen-de-la-comunitat-valenciana, 2024. Retrieved: October, 2024.

[10] A. Gavioli, E. G. de Souza, C. L. Bazzi, K. Schenatto, and N. M. Betzek, “Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods,” *Biosystems Engineering*, vol. 181, pp. 86–102, 2019, ISSN: 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2019.02.019.

ISBN: 978-1-68558-324-8

TABLE I. FEATURES FROM THE 10 FIRST OBSERVATIONS OF THE DATAFRAME (DETAILS IN SECTION II).

Variety	Season	Duration	m	M	Q	σ^2
Clausellina/Okitsu	2015-2016	0.12	0.20	0.27	0.61	0.00089
Clausellina/Okitsu	2016-2017	0.15	0.16	0.26	0.65	0.00123
Clausellina/Okitsu	2017-2018	0.13	0.23	0.26	0.22	0.00019
Clausellina/Okitsu	2018-2019	0.21	0.15	0.25	0.47	0.00138
Clausellina/Okitsu	2019-2020	0.87	0.20	0.29	0.10	0.00093
Clausellina/Okitsu	2020-2021	0.10	0.23	0.27	0.42	0.00023
Clementina Arrufatina	2015-2016	0.12	0.27	0.34	0.61	0.00079
Clementina Arrufatina	2016-2017	0.15	0.22	0.38	1.04	0.00189
Clementina Arrufatina	2017-2018	0.21	0.29	0.36	0.33	0.00045
Clementina Arrufatina	2018-2019	0.12	0.23	0.30	0.61	0.00054



	Clausellina/Okitsu	Clementina Arrufatina	Clementina Clemenules	Clementina Marisol	Clementina Oronules	Mandarina Clemenvilla	Mandarina Nadorcott	Mandarina Ortanique	Naranja Navel Lane Late	Naranja Navel Powell	Naranja Navelina	Naranja Washington Navel	Salustiana	Satsuma Owari	Valencia Late
Clausellina/Okitsu	1.00	0.79	0.62	0.95	0.81	0.49	0.57	0.77	0.47	0.71	0.47	0.59	0.46	0.77	0.55
Clementina Arrufatina	0.79	1.00	0.41	0.73	0.98	0.59	0.80	0.65	0.36	0.66	0.19	0.57	0.37	0.77	0.64
Clementina Clemenules	0.62	0.41	1.00	0.67	0.41	0.79	0.46	0.83	0.91	0.81	0.88	0.80	0.85	0.81	0.73
Clementina Marisol	0.95	0.73	0.67	1.00	0.73	0.53	0.49	0.83	0.52	0.67	0.55	0.61	0.50	0.76	0.54
Clementina Oronules	0.81	0.98	0.41	0.73	1.00	0.56	0.83	0.65	0.38	0.69	0.19	0.62	0.39	0.77	0.64
Mandarina Clemenvilla	0.49	0.59	0.79	0.53	0.56	1.00	0.70	0.68	0.78	0.81	0.60	0.84	0.74	0.74	0.82
Mandarina Nadorcott	0.57	0.80	0.46	0.49	0.83	0.70	1.00	0.48	0.50	0.76	0.21	0.77	0.55	0.76	0.69
Mandarina Ortanique	0.77	0.65	0.83	0.83	0.65	0.68	0.48	1.00	0.78	0.82	0.75	0.72	0.73	0.81	0.74
Naranja Navel Lane Late	0.47	0.36	0.91	0.52	0.38	0.78	0.50	0.78	1.00	0.81	0.90	0.88	0.93	0.78	0.79
Naranja Navel Powell	0.71	0.66	0.81	0.67	0.69	0.81	0.76	0.82	0.81	1.00	0.66	0.85	0.84	0.84	0.84
Naranja Navelina	0.47	0.19	0.88	0.55	0.19	0.60	0.21	0.75	0.90	0.66	1.00	0.69	0.83	0.63	0.67
Naranja Washington Navel	0.59	0.57	0.80	0.61	0.62	0.84	0.77	0.72	0.88	0.85	0.69	1.00	0.86	0.85	0.76
Salustiana	0.46	0.37	0.85	0.50	0.39	0.74	0.55	0.73	0.93	0.84	0.83	0.86	1.00	0.79	0.78
Satsuma Owari	0.77	0.77	0.81	0.76	0.77	0.74	0.76	0.81	0.78	0.84	0.63	0.85	0.79	1.00	0.77
Valencia Late	0.55	0.64	0.73	0.54	0.64	0.82	0.69	0.74	0.79	0.84	0.67	0.76	0.78	0.77	1.00

Figure 2. Mean correlation of Orange (“Naranja"), Clementines (“Clementina") and Mandarins (“Mandarina") varieties on the different seasons.