Engineering Applications of Artificial Intelligence

Evolutionary feature selection approach for yield prediction of subtropical crops

--Manuscript Draft--

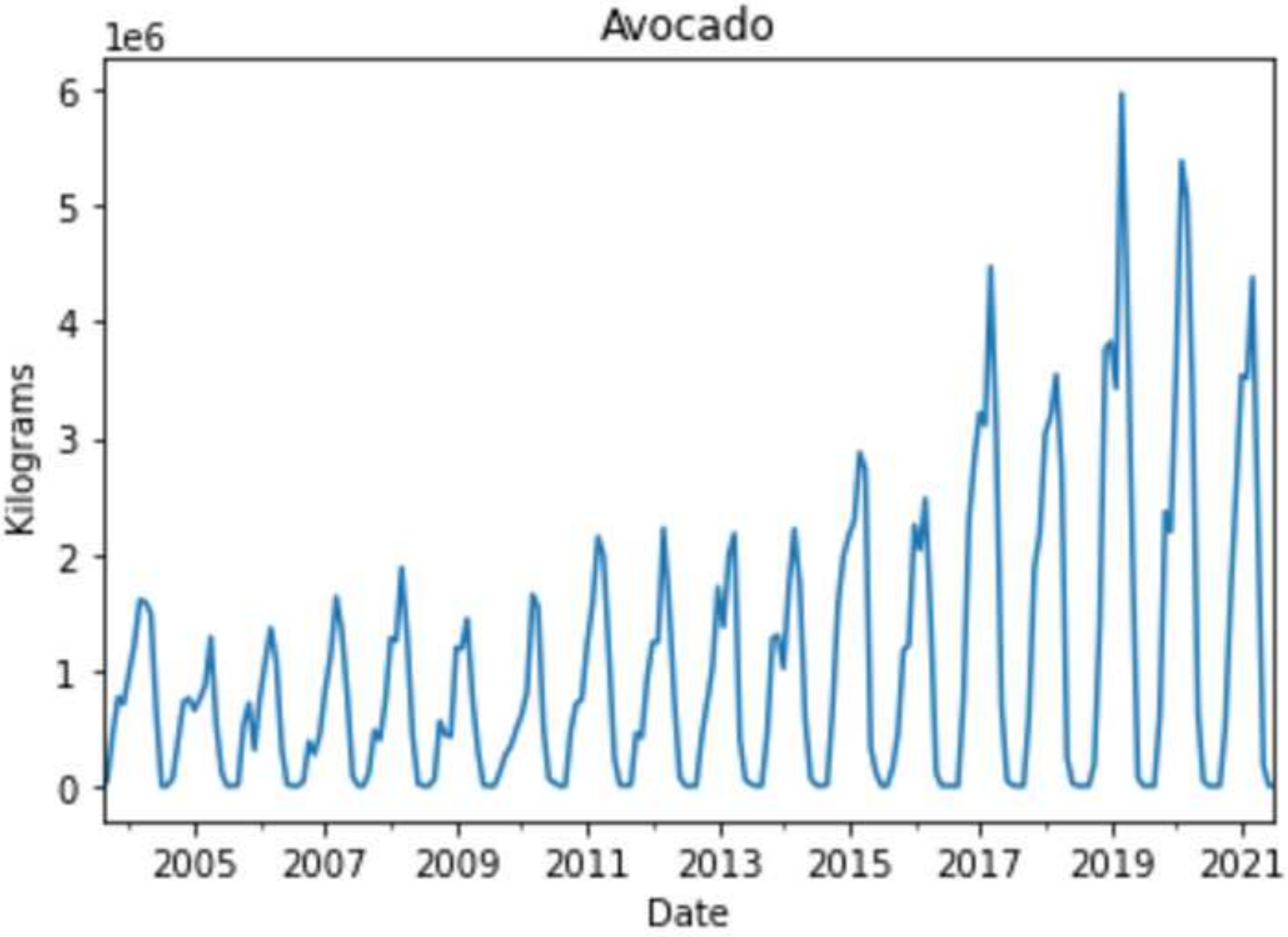
Manuscript Number:	EAAI-23-6198
Article Type:	Research paper
Keywords:	Feature selection; Data-driven; Crop yield forecasting; Subtropical crop; Correlation statistic; Evolution strategy
Abstract:	Accurate crop yield forecasting and understanding the most influential meteorological factors are essential for efficient agricultural practices and supply chain management. This study presents a data-driven methodology using a 20-year time series dataset with monthly granularity, incorporating meteorological variables and their lagged values to analyze their impact on subtropical crop yields, with avocado and mango from Axarquía (Málaga) as the case study. However, numerous features may lead to overfitting and reduced model performance. Therefore, devising techniques to manage features and their combination is pivotal. This feature selection might slightly change over time and depends entirely on the crop type and geographical location. The proposed method, Correlation Statistic with Maximum Relevance Minimum Redundancy-Discrete Weighted Evolution Strategy for Regression (CS_MRMR-DWES_R), not only enhances the model's reliability and accuracy but also offers a generic approach for identifying key drivers of crop yields. First, the correlation statistic of each feature with the target variable is calculated and the features are ranked along with Maximum Relevance Minimum Redundancy (MRMR) to select the best 10% of the best features. Then, the selected features are passed to the evolution strategy customized and optimized for regression data. This approach allows stakeholders to make informed decisions based on the most relevant meteorological factors specific to their crop type and region, leading to optimized agricultural practices and improved supply chain management. The implementation of CS-DWES is available on https://github.com/KhaosResearch/CSMRMR-DWESR.

Highlights

- To collect, preprocess, and present mango and avocado yield datasets from the Axarquía region of Málaga in Spain.
- To propose a supervised filter ranker using Correlation Statistic with Maximum Relevance Minimum Redundancy (CS_{MRMR}).
- To propose a wrapper Discrete Weighted Evolution Strategy for Regression ($DWES_R$).
- To propose a final hybrid feature selection for regression by combining CS_{MRMR} and $DWES_R$.

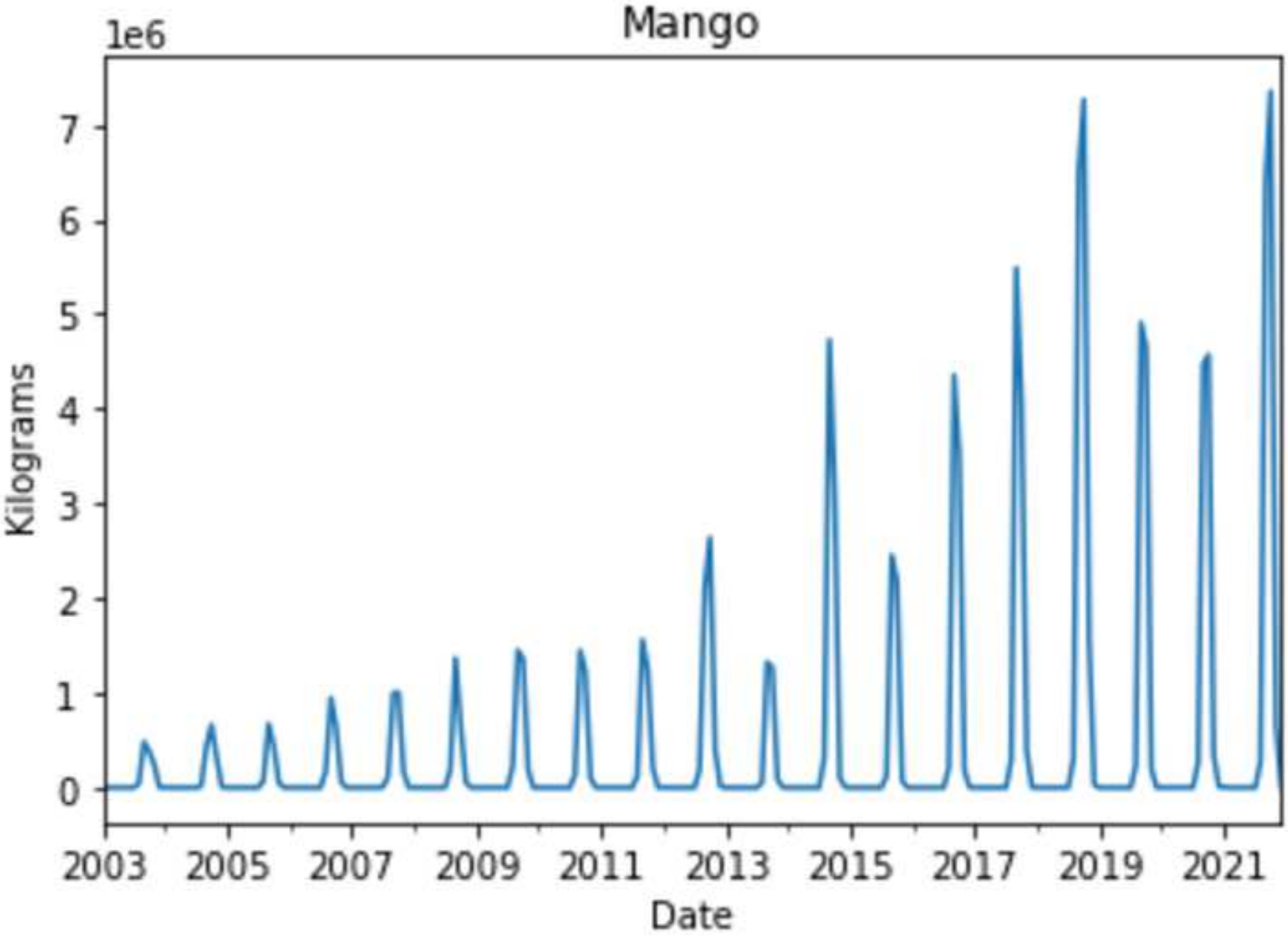
Figure

Click here to access/download;Figure(s);AvocadoTrend.png



Figure

Click here to access/download;Figure(s);MangoTrend.png



Evolutionary feature selection approach for yield prediction of subtropical crops

Hossein Nematzadeh[a,b,*], Cristian Cardas[a,b], Sandro Hurtado[a,b], Ismael Navas-Delgado[a,b], José F. Aldana-Montes[a,b], José García-Nieto[a,b]

[a]*ITIS Software, Universidad de Málaga, Arquitecto Francisco Peñalosa 18, Malaga, 29071, Spain*
[b]*Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Malaga, Spain*

Abstract

Accurate crop yield forecasting and understanding the most influential meteorological factors are essential for efficient agricultural practices and supply chain management. This study presents a data-driven methodology using a 20-year time series dataset with monthly granularity, incorporating meteorological variables and their lagged values to analyze their impact on subtropical crop yields, with avocado and mango from Axarquía (Málaga) as the case study. However, numerous features may lead to overfitting and reduced model performance. Therefore, devising techniques to manage features and their combination is pivotal. This feature selection might slightly change over time and depends entirely on the crop type and geographical location. The proposed method, Correlation Statistic with Maximum Relevance Minimum Redundancy-Discrete Weighted Evolution Strategy for Regression (CS_{MRMR}-$DWES_R$), not only enhances the model's reliability and accuracy but also offers a generic approach for identifying key drivers of crop yields. First, the correlation statistic of each feature with the target variable is calculated and the features are ranked along with Maximum Relevance Minimum Redundancy (MRMR) to select the best 10% of the best features. Then, the selected features are passed to the evolution strategy customized and optimized for regression data. This approach allows stakeholders to make informed decisions based on the most relevant meteorological factors specific to their crop type and region, leading to optimized agricultural practices and improved supply chain management. The implementation of CS_{MRMR}-$DWES_R$ is available on `https://github.com/KhaosResearch/CSMRMR-DWESR`.

Keywords: Feature selection, Data-driven, Crop yield forecasting, Subtropical crop, Correlation statistic, Evolution strategy

1. Introduction

Avocado [1] and mango [2] crops have become increasingly important in the agricultural sector in the south of Spain due to their high economic value, growing demand, and

*Corresponding author
Email addresses: `hnematzadeh@uma.es`, `hn_61@yahoo.com` (Hossein Nematzadeh), `criscardas@uma.es` (Cristian Cardas), `sandrohr@uma.es` (Sandro Hurtado), `ismael@uma.es` (Ismael Navas-Delgado), `jfaldana@uma.es` (José F. Aldana-Montes), `jnieto@uma.es` (José García-Nieto)

the special conditions for their plantation. They are typically grown in the provinces of Málaga and Granada, where the Mediterranean climate and soil conditions are suitable for their cultivation.

Their popularity and exclusivity have become one of the primary sources of income for local communities in rural areas. However, concerns have been raised about the environmental impact of intensive cultivation practices, irrigation being the most troubling factor. In that regard, sustainable agriculture practices, such as organic farming and efficient water management, are being promoted to reduce the negative impact on the environment while maintaining the economic benefits of these crops.

Efficiently managing resources and devising sustainable and profitable actions are critical to the long-term success of avocado and mango crops. In that matter, crop yield forecasting [3, 4] could provide farmers and policymakers with the information to make educated decisions about planting, harvesting logistics, and resource management. Accurate crop yield forecasts help optimize the use of seeds, fertilizers, and irrigation water. Studying the events that impact their forecasting might shed light on risks related to weather, pests, and diseases. However, the ability to estimate crop yield also prevents food shortages and price spikes by early identification of potential supply shortages and surpluses [5], allowing organisations to make strategic decisions about exporting policies.

Forecasting the yield of avocado [6] and mango [7] crops is a complex process for their known variability from year to year. There are also several external factors such as weather conditions, pollination, pests, and diseases which are not only significantly variable per annum, they also affect the quality and quantity of the yield, difficulting accurate forecastings. To overcome these challenges, crop yield forecasting models use a combination of data sources and techniques, including historical yield data, weather data, satellite imagery, and ground-based observations, among others.

To further consolidate the effects of external variables on the yield, lagged variables of past observations can be included as new variables for the data model. These variables can help to capture the dynamic relationships between different variables and improve the accuracy of predictions. For instance, a model incorporating lagged weather data can better capture the impact of temperature and rainfall on crop growth and development from past observations. Nonetheless, while including more variables can potentially provide more context and improve the accuracy of predictions, it can also improve the complexity of the model and the difficulty of the training of algorithms. Therefore, it is important to consider the relevance and correlation of each variable to the target variable being forecasted.

Feature selection and feature engineering techniques can be used to identify the most important and relevant features for forecasting crop yield [8]. By analyzing the correlation and contribution of each external variable to the model, only those with a strong relationship with the target variable and weak correlation to other factors can be

selected to build an explanatory and more straightforward dataset.

Feature selection methods are mainly divided into three basic groups: filter [9], wrapper [10], and embedded [11]. Moreover, ensemble and hybrid [12, 13, 14] feature selection methods are extensions of the basic methods. Filter methods evaluate the characteristics of individual features using statistical measures, heuristics, meta-heuristics, or the capacity of learning algorithms to rank or score features based on their relevance or importance. Filter methods can indeed use learning algorithms, but they typically do not involve splitting the data into separate train and test sets for evaluation during the feature selection process. Wrapper methods select features using a specific machine learning algorithm to train and evaluate subsets of features. These methods search for the optimal subset of features by iteratively evaluating different combinations based on the performance of the chosen algorithm. Embedded methods incorporate feature selection within the process of training a machine learning algorithm. They aim to select the most informative features by considering their relevance during the algorithm's training phase. These methods often utilize regularization techniques or specific algorithmic properties to perform feature selection. Ensemble methods combine multiple feature selection techniques to leverage their strengths and mitigate their weaknesses. They aim to improve feature selection performance by aggregating the results from different methods, often using voting or weighting schemes. Hybrid methods combine two or more feature selection techniques, such as filter and wrapper or filter and embedded methods. They seek to benefit from the advantages of each technique while compensating for their limitations, resulting in potentially improved feature selection performance.

This paper aims to provide a methodology for obtaining the most relevant features for avocado and mango crop yield forecasting per annum. This information provides farmers with an explanation of the external factors that condition and alter crop yield, allowing them to prepare against them better. To that end, a hybrid feature selection method has been developed that considers both traditional meteorological variables, terrain information, and multi-spectral vegetation indices with lagged variables to identify the most important factors that influence crop yield. By selecting the most relevant features, we aim to improve the accuracy and robustness of crop yield predictions, which can benefit farmers and other stakeholders in the agriculture industry. Additionally, most of the hybrid approaches for feature selection have been applied to classification problems, and not much attention has been devoted to agricultural time series data with continuous response variables (regression problems). This serves as one of the primary motivations in this study. As such, the contributions of this research are as follows:

1. To collect, preprocess, and present subtropical crops (mango and avocado) yield datasets with monthly granularity, incorporating meteorological variables and their lagged values.

2. To propose a supervised filter ranker using Correlation Statistic with Maximum Relevance Minimum Redundancy (CS_{MRMR}).

3. To propose a wrapper Discrete Weighted Evolution Strategy for Regression ($DWES_R$). $DWES_R$ is an enhanced version of $DWES$ [15] that offers improvements in two key aspects. Firstly, $DWES_R$ features a reduced number of hyperparameters, resulting in improved algorithmic efficiency. Secondly, $DWES_R$ has been extended to support time series regression problems, whereas the original $DWES$ was solely designed for classification problems.

4. To propose a final hybrid feature selection for regression by combining CS_{MRMR} and $DWES_R$. It is empirically evaluated on the previously elaborated datasets.

The rest of this paper is organized as follows: Section 2 introduces the related works for feature selection of regression data. Section 3 presents the continuous (1+1) Evolution Strategy. Section 4 describes how the datasets are collected and preprocessed and also explains the research methodology in detail. Section 5 shows the results with respective illustrations and tables and provides discussions on the obtained results. Finally, Section 6 presents concluding remarks and future works

2. Related works

This section highlights some widely recognized strategies (also available in Python libraries) for feature selection of datasets that involve a regression (continuous) response variable. These strategies are categorized into filter, wrapper, embedded, and hybrid methods, and are equally applicable for feature selection of time series data with a continuous response variable. We intend to implement some of these strategies on our self-collected subtropical agricultural time series datasets later to compare them with our proposed method. It should be noted that, compared to regression datasets, feature selection for classification problems has been more extensively studied, as classification tasks are prevalent in various domains, including medical diagnosis [16, 17, 15], network intrusion detection[18], and text classification[19].

2.1. Filter methods

Linear Regression Coefficients (LRC) [20] measures how much the target variable changes when each feature changes by one unit. This means the features with the strongest linear relationship with the target variable would have better rankings. LRC does not account for the interaction between features or the nonlinearity of the relationship. Thus, this may not be the best criterion for feature selection, as some features may have a nonlinear or complex relationship with the target variable that is not captured by the coefficients. The filter LRC involves fitting linear regression to the entire dataset without splitting it into training and testing sets. It then sorts the features based on the calculated coefficients, selecting the top k features. These selected features are used to construct a new dataset, which is then fed into the predictive model for performance evaluation.

Correlation Coefficients (CC) calculates the Pearson correlation coefficient [21] for each pair of features in a matrix. Alternatively, other measures such as Spearman [22] can be used instead of Pearson. The coefficients return a symmetric matrix with values ranging from -1 to 1. The diagonal elements are 1, and the off-diagonal elements represent the correlation coefficients. These coefficients can provide insights into multicollinearity. The correlation coefficient measures how closely two variables vary together, regardless of their units or slope.

Correlation Statistics (CS) is a type of feature selection method that computes the F-statistic and p-value for each feature and the target variable. The F-statistic measures how well a feature can explain the variance in the target variable, and the p-value measures how likely the feature and the target variable are independent. CS is calculated by squaring the CC and dividing it by the ratio of the unexplained variance to the degrees of freedom. The unexplained variance is the variance in the target variable that is not explained by the feature, and it is equal to 1 minus the square of the CC. The degree of freedom is the number of observations minus the number of parameters in the model, and it is equal to n - 2 for a simple linear regression with one predictor variable. The higher the CS, the more likely the feature and the target variable have a linear relationship.

Mutual Information (MI) [9, 23] is a measure of how much information one variable contains about another variable. It is equal to zero if the two variables are independent, and higher values mean higher dependency (MI ranges in [0,1]). Mutual information feature selection is a filter method that ranks features based on their mutual information with the target variable and selects the top k features with the highest scores.

Granger Causality-based Feature Selection (GCFS) [24, 25] is a method that selects features based on the Granger causality test, which determines whether one time series is useful in forecasting another. While performing the Granger causality test for all pairs of predictor variables, the p-values are also computed for each pair. The p-value indicates the probability of observing a test statistic as extreme as the one observed under the null hypothesis of no Granger causality. Finally, the top k features are selected based on the highest p-values, which means they are the least likely to be Granger-caused by any other feature.

Concerning this matter, the authors in [26] have endeavored to use the machine learning algorithm Artificial Neural Network (ANN) and statistical model Multiple Linear Regression (MLR) aimed at achieving precise yield prediction. This was accomplished by employing diverse feature selection algorithms, notably correlation-based feature selection, and mutual information, to select distinct subsets of features. These identified feature subsets were subsequently applied to the MLR algorithm to determine the most optimal feature subset.

2.2. Wrapper methods

Sequential Feature Selection (SFS) [10] is a wrapper method for feature selection since it uses the SFS estimator that wraps around a base estimator model and evaluates the

performance of different subsets of features. SFS is a greedy feature selection in which an algorithm will either select the best features one by one (forward selection) or removes the worst feature one by one (backward selection). The estimator chooses the best feature to add or remove at each stage. In [27], the authors used variable selection algorithms to find the most critical features for optimizing crop yield prediction. They found the optimum feature subset by employing the forward selection and backward elimination algorithms evaluated with the MLR algorithm. However, the obtained 85% accuracy suggests room for improvement. Exploring alternative variable selection algorithms might have resulted in better performance, as using all features yielded 84% accuracy.

Recursive Feature Elimination (RFE) [28] uses the feature weight coefficients (e.g., linear models) or feature importance (tree-based algorithms) to eliminate features recursively. RFE uses a predictive model for evaluation similar to forward/backward SFS. The model is specified by the estimator parameter, which can be any estimator that assigns importance to features. RFE ranks the features according to the model and eliminates the least important features until the desired number of features is reached. In [29], the authors employed wrapper feature selection techniques like Recursive Feature Elimination (RFE) and Modified Recursive Feature Elimination (MRFE), which involves shuffling and combining the dataset, to achieve a remarkable 97% accuracy for crop prediction using Random Forest. However, exploring other variable selection methods might have further improved the results.

Exhaustive Feature Selection (EFS) [30] evaluates all possible combinations of the input features and finds the best subset based on a scoring function. However, this can be very time-consuming and computationally expensive, especially if a large number of features or a complex model exists. Assuming the optimal set contains k features and the total number of features is N, then EFS should create $\frac{N!}{k!(N-k)!}$ predictive models.

2.3. Embedded methods

Lasso regression (Lasso) [11] is a type of linear regression that is used for feature selection and regularization. It adds a penalty term to the cost function of the linear regression model, which is proportional to the absolute value of the coefficients. This penalty term shrinks the coefficients of less important features to zero, effectively removing them from the model. *Elastic net* [12] is another feature selection method that combines L1 and L2 regularization to shrink the coefficients of less important features to zero or near zero Elastic net uses a mixing parameter to balance the L1 and L2 penalties. Elastic net performs variable selection and regularization simultaneously.

Almost all tree-based feature selection methods are embedded methods because they use the structure of the tree or the ensemble of trees to rank or select the features based on some criterion, such as impurity or importance. The most common tree-based embedded feature selection methods include *Decision Tree* (DT) [31], and its ensemble improvement algorithms *Random Forest* (RF) [32], *Bagging* [33], and *Boosting* [34]. DT learns simple decision rules from the input features. A DT performs feature selection

by measuring how much each feature contributes to reducing the error or impurity of the model. Bagging, RF, and boosting techniques try to solve the instability of regressors. Bagging fits multiple large trees to bootstrapped versions of the training data and combines their predictions through aggregation, such as averaging, to create a more robust and accurate regression model. *Bootstrap Aggregator* (Bagging) is a well-known example of this technique used in regression tasks. RF is an improvement over bagging with randomized trees. Boosting techniques fit multiple large or small trees to reweight versions of the training data and combine their predictions through a weighted majority vote. Well-known boosting techniques include *Adaptive Boosting* (AdaBoost) [34] and *eXtreme Gradient Boosting* (XGBoost) [32]. Boosting and bagging are popular ensemble learning methods (though they are categorized as embedded feature selection methods), with boosting often offering improved performance over bagging in various scenarios. On the other hand, random forests, which leverage bagging principles, tend to exhibit superior performance compared to boosting in many cases.

In this context, embedded methods have also been employed for crop yield prediction. In [35], the authors utilized various feature selection techniques, including Random Forest importance feature selection, to identify a specific subset of features. These selected features were then subjected to evaluation within multiple machine learning models, resulting in an accuracy rate of up to 90%.

2.4. Hybrid methods

Hybrid methods in feature selection are commonly proposed to leverage the benefits of multiple techniques. One approach involves reducing the feature set size through a filter method, followed by a wrapper method. By employing this approach, wrapper methods can be effectively applied to smaller feature sets. Hybrid methods are typically research-driven, as scientists endeavor to combine multiple feature selection techniques for enhanced results.

Hmamouche et al. [13] proposed the causality-Graph based Feature Selection Method (GFSM) that combines a causal filtering step with a clustering step to select the final set of predictors. The clustering step can be seen as a wrapper method as it selects features based on their effect on the predictive performance of a specific machine learning model. GFSM may exhibit limitations such as potential slowness due to its reliance on calculating the graph of causalities and a subsequent clustering step. Another challenge that can arise is selecting the appropriate lag when computing causalities. Hmamouche et al. [14] also proposed PEHAR (Predictors Extraction using Hubs and Authorities Ranking). It applied a modified version of the HITS (Hyperlink-Induced Topic Search) algorithm on a multivariate time series graph of causalities (using granger causality or transfer entropy). The lag choice is still a limitation, like GFSM. PEHAR underwent testing on both macroeconomic and sales transaction datasets, and the outcomes demonstrated promise as it enhanced prediction accuracy when compared to traditional methods such as kernel principle component analysis, factor analysis, and

the correlation-based filter method. Both GFSM and PEHAR exemplify hybrid methods that combine filter and wrapper techniques.

In contrast, Amini et al. [12] introduced a hybrid method called GA-EN, which combines a wrapper approach with an embedded method. The proposed method consisted of two layers. In the first layer, a Genetic Algorithm (GA) is utilized as a wrapper to search for the optimal subset of predictors. Since GA did not guarantee optimality, a second layer was introduced to enhance prediction accuracy by eliminating any remaining redundant or irrelevant predictors. In this layer, the Elastic Net (EN) method was employed as the embedded approach. EN is preferred for its flexibility in adjusting penalty terms during the regularization process and its time efficiency. The GA-EN hybrid method was specifically applied to agricultural datasets, particularly Maize datasets, with superior results.

Accordingly, this paper follows the principles of hybrid feature selection by employing a combination of the Correlation Statistic (CS) and Maximum Relevance Minimum Redundancy (MRMR) (referred to as CS_{MRMR}) initially. The filter CS_{MRMR} is used to discard 90% of non-informative features. Subsequently, a wrapper method called Weighted Discrete Evolution Strategy for Regression (DWES_R) is proposed for the final feature selection. While CS_{MRMR} investigates the individual merits of each feature, DWES_R effectively explores the relationships between features to achieve higher accuracy.

3. (1+1) Evolution strategy

(1+1) Evolution Strategy ((1+1) EA) is one of the variations of evolution strategy that differs from the Genetic Algorithm (GA) in two major ways [36]. First, GA utilizes both crossover and mutation, whereas (1+1) EA only employs mutation. Additionally, (1+1) EA does not require the problem to be represented in a coded form. These characteristics make (1+1) EA a highly efficient evolutionary algorithm suitable for use as a wrapper feature selection method. (1+1) EA is intrinsically a continuous evolutionary algorithm as shown in Algorithm 1. However, this paper discretizes the continuous (1+1) EA and improves Discrete Weighted Evolution Strategy (DWES) [15] by proposing Discrete Weighted Evolution Strategy for Regression DWES_R. DWES_R is specifically designed to address time series regression forecasting problems, and it has the following advantages over DWES:

1. DWES_R addresses the issue of hyperparameter selection by minimizing the number of hyperparameters and fixing the number of clusters, which is different from DWES where the user had to determine the optimal number of clusters manually. Furthermore, DWES_R replaces hierarchical clustering with KMeans, removing the need for linkage selection. These improvements lead to increased speed and reduced risk of overfitting by selecting a smaller number of features.

2. DWES$_R$ introduces an updated fitness function (R-squared (R^2)) that makes it compatible with forecasting time series regression problems initially. Secondly, DWES$_R$ substitutes Support Vector Machine and Decision Tree with SARIMAX to support residuals with seasonal trends or patterns. These modifications allow DWES$_R$ to effectively handle and forecast time series regression tasks.

Algorithm 1 Continuous (1+1) EA

Require: problem: an optimization problem
Ensure: parent population
1: Generate parent population of size n: x_1, x_2, ..., x_n
2: Calculate parent population fitness: Z_1=f(x_1, x_2, ..., x_n)
3: Generate offspring population of size n: $x'_1 = x_1 + a\ (0,\gamma)$, $x'_2 = x_2 + a\ (0,\gamma)$, . . . , $x'_n = x_n + a\ (0,\gamma)$
4: Calculate offspring population fitness: Z_2=f(x'_1, x'_2, ..., x'_n)
5: If $Z_2 > Z_1$ substitute parent population with offspring population otherwise keep parent population.
6: Repeat steps 3-5 until stopping criterion is reached.
7: **return** parent population

4. Materials and methods

The importance of this work resides in providing mechanisms and presenting a reasonable methodology to study the impact of features in subtropical crop data. This section describes the datasets under study and provides details of the work's proposal. Section 4.1 introduces the source of the data and the preprocessing (including handling missing data, generating monthly aggregated data from the original datasets, and extending univariate time series to multivariate ones). Section 4.2 explains how the variables of the preprocessed aggregated datasets are divided into different time lags. This helps to investigate if some previous years' specific variables impact production. However, it increases the dimensionality of the datasets as a side effect. Finally, Section 4.3 explains the proposed method.

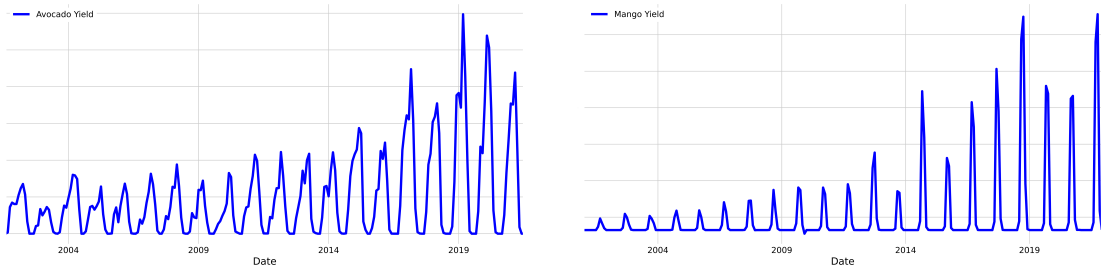
4.1. Datasets

TROPS, S.A.T. 2803 [1] is a farmers' organization specializing in producing and marketing avocado and mango crops. TROPS was founded in 1979 in Vélez-Málaga (Spain) and currently groups more than 3,000 associated growers together. Their main

[1] Available in URL `https://www.trops.es/`. Last access date 19/10/2022

production area is concentrated in the Axarquía region of Málaga and the tropical coast of Granada, following the coast of Valencia, the Portuguese Algarve, and as far as South America (Peru and Chile). This organization of producers provided this study with a historical crop yield record starting in 2001. While following specific insertion criteria, the data had missing values or different formats, so a preprocessing phase was needed. Thus, we used linear interpolation to solve these issues. The historical data initially included daily crop yield for the main mango and avocado variants cropped in the region. As the mango season starts in August and ends in November in this region, the existing records only include values for those months. Conversely, the avocado species have different growing periods, so their records vary throughout the year.

Despite processing data for distinct species and days of production, the first step consisted in grouping them in a monthly aggregated dataset due to the data's nature. Working with daily or weekly aggregated crop data with a clear yearly seasonality (as seen in Figure 1a and Figure 1b) in a long and complex series does result in poor fine-grained forecasting. Additionally, meteorological variables and terrain information were included as exogenous attributes to extend the univariate time series to a multivariate one. A sample of data corresponding to the considered variables can be observed in the example in Table 1. From left to right, the resulting time series consists of the monthly dates, the yield as the target or observation variable, and nine external variables involving terrain and meteorological data. These last values were extracted from the Spanish Agroclimatic Information System for Irrigation (SIAR[2]) and included the average (*Tave*), minimum (*Tmin*) and maximum (*Tmax*) temperatures in Celsius degrees, the average humidity in percentage (*Have*) and the average wind velocity in meters per second (*Velave*), its direction in degrees (*Dir*) and the total amount of precipitations in millimeters (*Prec*). The terrain variables are the number of trees (*NTrees*) and the planted surface in hectares (*Surface*).



(a) Plot for the complete avocado crop registry.

(b) Plot for the complete mango crop registry.

Figure 1: Plots for the complete time series of yield production (2001 - 2021).

[2]https://eportal.mapa.gob.es/websiar/Inicio.aspx

4.2. Time series data modeling

This work's experiments have been based on the dataset characterized in Table 1. While the multivariate time series includes external variables that describe the behavior of the yield data, they can be harnessed more explicitly through lags. Therefore, several datasets have been generated depending on the length of the lags, resulting in three different categories: 4, 12 and 24 lags for both crops (from this point, L4, L12 and L24). Figure 2 for L4 illustrates the lag generation procedure. One drawback of using lags is that past observations for each attribute are required at every instance, meaning that the first observations will not have any lag values. The more lags are considered, the larger the number of dates with empty values which can affect the training of forecasting models. Thus, the first two years of data have been discarded for all datasets, as the largest number of lags in this work is 24. Ultimately, the used datasets include their respective harvesting years from 2003 to 2021. Figure 1a presents a plot of the avocado yield registry. The avocado crop's harvesting year starts in October and finishes towards the end of May and the middle of June. The peak yield is usually reached in March. On the other hand, Figure 1b shows the mango yield registry. Compared to avocados, mango crop harvesting is concentrated from August to November.

Table 1: A sample year of monthly grouped mango time series.

Time	Target	Terrain		Meteorological						
Date	Yield	Surface	NTrees	Tave	Tmax	Tmin	Have	Velave	Dir	Prec
2021-01	6,982	1,012.92	1,063,753	13.3	27.4	2.7	66.0	1.78	310.98	152.2
2021-02	0	1,012.98	1,063,828	15.3	24.0	7.2	71.0	1.39	298.82	3.7
2021-03	0	1,012.98	1,063,828	16.1	25.9	5.1	64.0	1.09	275.33	34.8
2021-04	0	1,013.98	1,064,828	18.4	28.5	11.3	70.0	1.1	249.08	65.9
2021-05	0	1,015.05	1,065,981	21.1	37.0	11.7	57.0	1.28	243.08	21.5
2021-06	0	1,015.05	1,065,981	24.4	36.0	14.5	51.0	1.08	235.01	1.7
2021-07	133	1,015.05	1,065,981	28.0	42.9	19.0	47.0	1.31	227.06	0.0
2021-08	262,636	1,015.27	1,066,231	28.2	37.6	19.0	57.0	1.05	203.84	1.0
2021-09	6,409,538	1,015.27	1,066,231	25.5	36.8	16.2	61.0	1.17	261.36	11.4
2021-10	7,350,187	1,015.27	1,066,231	20.7	32.9	10.9	59.0	1.03	281.36	10.2
2021-11	657,100	1,015.27	1,066,231	15.8	29.0	5.6	57.0	1.75	305.38	9.0
2021-12	7,803	1,015.27	1,066,231	15.0	24.8	6.6	67.0	1.94	305.80	39.4

4.3. Proposed method

This section elaborates on the general phases of the proposed method, namely Correlation Statistic with Maximum Relevance Minimum Redundancy-Discrete Weighted Evolution Strategy for Regression (CS_{MRMR}-$DWES_R$) in Figure 3. Initially, the features

Date	Yield	Tmax	Tmax(t-1)	Tmax(t-2)	Tmax(t-3)	Tmax(t-4)	Tmin	...	Tmin(t-4)	...
2021-01	6,982	27.4	24.5	28.3	31.75	35.2	2.7	...	16.2	...
2021-02	0	24.0	27.4	24.5	28.3	31.75	7.2	...	12.8	...
2021-03	0	25.9	24.0	27.4	24.5	28.3	5.1	...	9.4	...
2021-04	0	28.5	25.9	24.0	27.4	24.5	11.3	...	4.7	...
2021-05	0	37.0	28.5	25.9	24.0	27.4	11.7	...	2.7	...
2021-06	0	36.0	37.0	28.5	25.9	24.0	14.5	...	7.2	...
2021-07	133	42.9	36.0	37.0	28.5	25.9	19.0	...	5.1	...
2021-08	262,636	37.6	42.9	36.0	37.0	28.5	19.0	...	11.3	...
...	...	...	...	...	...	...	...	...	...	...

Figure 2: A sample of how lagged variables are added for each feature. For each new lag, in this example up to 4, the values of the previous time step are shifted by one month, this is shown through the different colors in the figure. For instance, the value for row 2021-02 column Tmax(t-1) is the value from Tmax in 2021-01.

are sorted based on their correlation to the regression response variable in phase 1. However, only the best 10% of the features are retained and passed to phase 2. The best features in phase 1 are selected using Maximum Relevance Minimum Redundancy (MRMR) [37]. This is due to the increased diversity and quality of the selected features. Phase 1 is a supervised filter approach. It is supervised because the response variable is used in the process of sorting features. Additionally, it is a filter because it does not use any learning algorithm as a predictive model. Discrete Weighted Evolution Strategy for Regression (DWES_R) in phase 2 uses the selected features from phase 1 and automatically selects the best features that increase the accuracy of SARIMAX model. DWES_R clusters the features using the K-means clustering algorithm and intelligently assigns greater weights to the best clusters. Therefore, the clusters with more informative features always have a greater chance of being selected. DWES_R is the improved version of DWES [15] for regression datasets with fewer hyper-parameters needed to be tuned in advance by the user. Additionally, DWES_R is a wrapper approach since it directly uses the learning algorithm (SARIMAX) as a predictive model to select the final feature set. At the end of phase 2, DWES_R calculates the measurement criteria, including R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), U-Theil index (U-Theil), performance (execution time), and reports the selected features.

4.3.1. Phase 1: CS_{MRMR}

The process of supervised filter feature selection using Correlation Statistic with Maximum Relevance Minimum Redundancy (CS_{MRMR}) starts by calculating the correlation of each feature of the dataset (regressor or predictor) relative to the response

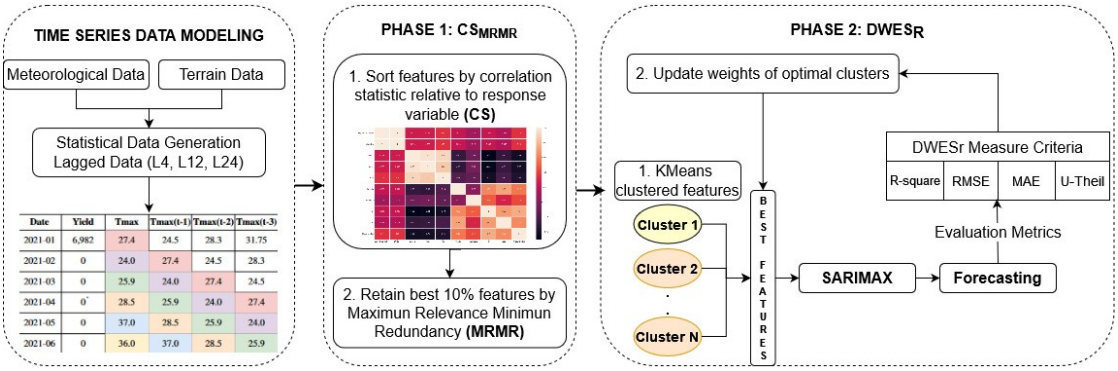


Figure 3: Data generation and phases of the proposed method.

variable (target).

The correlation is calculated using Pearson's correlation coefficient between an individual feature and the response variable. Eq.1 calculates Pearson's correlation coefficient, ρ, which can be used to calculate the F-statistic for testing the significance of a simple linear regression model with a feature x_i as the predictor variable and y as the response variable. The F-statistic follows an F-distribution with 1 and n-2 degrees of freedom. The Pearson correlation coefficient ρ measures the strength and direction of the linear relationship between two variables, x and y, and ranges from -1 to 1. A value of 1 or -1 indicates a perfect positive or negative relationship, respectively. A value of 0 indicates no linear relationship between the variables.

$$\rho = \frac{n\sum_{i=1}^{n}(x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)(n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}} \tag{1}$$

Therefore, the correlation score is calculated as cor_score in Eq. 2. It follows a t-distribution with n-2 degrees of freedom under the null hypothesis that ρ equals zero, where n is the number of samples and ρ is the Pearson correlation coefficient between x and y (calculated by Eq.1). x and y are two variables of interest, such as a feature and a target variable. cor_score can be converted to an F-statistic by squaring it and setting the numerator degrees of freedom to 1 and the denominator degrees of freedom to n-2.

$$\text{cor\_score} = \frac{\rho^2}{\frac{1-\rho^2}{n-2}} \tag{2}$$

The next step is to normalize the list of cor_score. This is important because MRMR is later calculated between correlation statistics and the Jaccard measure. Therefore, the range of cor_score should be normalized to have the same range as the Jaccard measure. Eq.3 shows how one element of cor_score is normalized where cor_score_i is the correlation score of the i^{th} feature of cor_score set, and min(cor_score) and max(cor_score)

represent the minimum and maximum correlation scores, respectively, among all features.

$$NC_i = \frac{cor\_score_i - \min(cor\_score)}{\max(cor\_score) - \min(cor\_score)} \tag{3}$$

The CS_{MRMR} in Algorithm 2 initially records the feature with the highest NC score and adds it to the selected feature set (reduced_feature_set). Then, Algorithm 2 iteratively selects features with the highest distance based on Eq. 4. The distance is a difference between the NC score of feature i (NC_i) and the Jaccard similarity of feature i (Jac_i) with respect to the mean value of reduced_feature_set. The Jaccard similarity calculates the absolute dot product of feature i (f_i) and the mean value of reduced_feature_set as shown in Eq. 6. (Jac_i) in Eq. 5 also calculates the norm of both (f_i) and the mean value of reduced_feature_set in Eqs. 7, and 8. The selected feature is added to the reduced_feature_set. This process is repeated until the desired number of features, k, is reached so that $k = 10\% \times m$ where m is the entire number of features in the input dataset.

$$dist_i = NC_i - Jac_i \tag{4}$$

$$Jac_i = \frac{C}{A + B - C} \tag{5}$$

where

$$C = |dot(f_i, mean(reduced\_feature\_set))| \tag{6}$$

$$A = (norm(f_i))^2 \tag{7}$$

$$B = (norm(mean(reduced\_feature\_set)))^2 \tag{8}$$

Algorithm 2 assures that reduced_feature_set contains features with both maximum correlation with respect to response variable as well as diversity through a supervised feature selection approach. The length of reduced_feature_set is significantly reduced in comparison with the entire set of features before supervised feature selection. However, this reduction is not enough and a wrapper feature selection in the next phase will select the best features from reduced_feature_set automatically using the SARIMAX learning algorithm.

4.3.2. Phase 2: $DWES_R$

After phase 1, the initial dataset D with size $n \times m$ will be reduced to D' with size $n \times m'$ so that $m' = 10\% \times m$. Phase 2 starts by clustering these m' features into the specified number of clusters (q). It has been experimentally shown (later in the result section) that $DWES_R$ works well when q=3 in our case of subtropical time series

Algorithm 2 CS_{MRMR} using Jaccard similarity measure

Require: Threshold for selecting features k, NC
Ensure: reduced_feature_set

1: Add the feature with the highest NC score to reduced_feature_set.
2: Calculate the mean value of the reduced_feature_set.
3: For each remaining feature not in reduced_feature_set, calculate the Jaccard index between the feature and the mean value of the reduced_feature_set.
4: Calculate the distance between each remaining feature's NC score and its Jaccard index.
5: Select the feature with the highest distance and add it to the reduced_feature_set.
6: Repeat steps 2-5 until k is reached
7: **return** reduced_feature_set

regression datasets. However, this paper proposes the formulation of $DWES_R$ generally using q. First, the m' features are clustered into q clusters using the K-means algorithm in Eq. 9. Simultaneously, a threshold is defined for each cluster in Eq. 10. A threshold specifies the selection probability of each cluster. Obviously, all clusters have an equal chance for selection at the beginning of $DWES_R$. $DWES_R$ sets this chance to 50% to implicitly mention that there is not any bias at the beginning and the $DWES_R$ algorithm is neutral to select any cluster. However, the selection probability may change during the algorithm iteration by iteration. The concept of threshold also helps automatic feature selection by intelligently paying more attention to clusters with more informative features.

$$C^q = KMeans(m', q) = c_1, c_2, ..., c_q \tag{9}$$

$$Th^q = th_1, th_2, ..., th_q, \;\; th_i = 0.5 \;\; i = 1, 2, ..., q \tag{10}$$

$DWES_R$ continues by randomly selecting one feature from each selected cluster to create the final_set in Algorithm 3 and the corresponding training and test sets accordingly. The idea is that the group of features within the same cluster have the same characteristics and there should not be a considerable difference among them (We have also experimentally tested this by assigning weights to the features of the same clusters and selecting them with respect to their weights, but we did not achieve any better results). Next, the fitness of the final_set (using the corresponding training and test sets) is calculated. The fitness is the accuracy (R^2) of the SARIMAX model for the final_set using the test set based on the fitted model by the training set as stated in Eq. 11.

$$Fitness(final\_set) = r2\_score(final\_set) \tag{11}$$

From now onwards, DWES$_R$ iteratively creates new subsets of features (temp_sets) and generates the respective training and test sets as well as the corresponding fitnesses. The temp_set would be saved as a final_set provided that it could increase the best so far fitness achieved. Likewise, the selection probability (thresholds) of those clusters that led to fitness improvement should be updated based on Eq. 12 so that $\beta = 0.1$ in this paper. Eq. 12 also guarantees that the weight corresponding to each cluster would never increase 1.

$$th_i = th_i + \beta \times (1 - th_i) \tag{12}$$

DWES$_R$ only exploits mutation by iteratively generating a new subset of features. However, this mutation is intelligently done because DWES$_R$ pays more attention and assigns more weights to more informative clusters. In fact, DWES$_R$ generates only one solution in each iteration (temp_set in Algorithm 3). This solution is directly generated through mutation (randomly selecting new features from corresponding clusters considering clusters' selection probabilities). The length of the final_set in Algorithm 3 may vary in each run of DWES$_R$. This is because clusters' selection probabilities cause automatic feature selection. Moreover, DWES$_R$ is stochastic in essence as it belongs to meta-heuristic algorithms. The stopping criteria in line 6 of Algorithm 3 is whether reaching the maximum number of iterations (50) or 5 consecutive iterations without fitness improvement. Algorithm 3 shows the general case, but in case of achieving the same fitness with fewer features, the final_set would be updated accordingly.

5. Experiments and results

This section first introduces the measurement criteria and experimental setup including specification of the SARIMAX model and the implementation platform in Sections 5.1 and 5.2, respectively. Then, the results, comparison and related discussions, and MRMR effectiveness are presented in Sections 5.3, 5.4, and 5.5.

5.1. Measurement criteria

The main measurement criteria are r2_score (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and U-Theil Index (U-Theil). In the Eqs. 13, 14, 15, and 16 y_i and $\hat{y}_i$ are actual value and forecasted value for i^{th} observation in test. $\bar{y}$ is the mean value of the response variable across all observations, and n is the number of observations in the test set. In short, R^2 generally ranges from 0 to 1 and represents the proportion of the variation in the response variable that is explained by the features in a regression model such as time series. MAE is the average of the absolute values of the errors between forecast and actual values for each observation in Eq. 14. RMSE is another common metric used to evaluate the performance of a time series model like SARIMAX as in Eq. 15 so that RMSE is similar to MAE, but it places more emphasis

Algorithm 3 DWES$_R$

Require: D', q
Ensure: final_set, best_acc
Generate C^q based on Eq. 9
Initialize Th^q based on Eq. 10
selected_clusters = Select clusters from C^q based on the corresponding selection probabilities (Th^q)
final_set = Randomly select a feature from each cluster of selected_clusters
best_acc = Run SARIMAX on the training and test sets corresponding to the final_set and record R^2
while Stopping criteria are not considered **do**
    selected_clusters = Select clusters from C^q based on the corresponding selection probabilities (Th^q)
    temp_set = Randomly select a feature from each cluster of selected_clusters
    temp_acc = Run SARIMAX on the training and test sets corresponding to the temp_set and record R^2
    if temp_acc > best_acc **then**
        best_acc = temp_acc
        final_set = temp_set
        Update selection probabilities of selected_clusters based on Eq. 12
    end if
end while
return final_set, best_acc

on large errors, as it squares the difference between the forecasted and actual values. The range of MAE and RMSE can vary depending on the scale of the data. U-Theil Index in Eq. 16 is a ratio that compares the accuracy of a forecast model with a naive forecast. The lower the ratio, the better the forecast model. Additionally, the performance of the algorithm (execution time) is also reported to give better perception of the proposed methodology.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

$$U\text{-}Theil = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n}} + \sqrt{\frac{\sum_{i=1}^{n} \hat{y}_i{}^2}{n}}} \tag{16}$$

5.2. Experimental setup

This section introduces the specification of the SARIMAX for reusability as well as the hardware/software platform of the programming environment. The execution time of SARIMAX directly depends on its parameter setting. Therefore, the best parameters for SARIMAX are specified by investigation within corresponding predefined ranges so that the (p,d,q) order are limited to p = [0, 1, 2], d = [1], and q = [0, 1, 2]. Likewise, the (P,D,Q,M) seasonal order are predefined ranges as P = [0, 1, 2], D = [1], Q = [0, 1, 2], and M = [12]. The datasets are divided into training and test sets so that the last year in the time series (the most recent year) is always a test set and the remaining years are the training set.

All the experiments have been conducted in a virtualization environment on a private, high-performance cluster computing platform. This infrastructure is located at the Ada Byron Research Center at the University of Málaga (Spain). It comprises several IBM hosting racks for storage, virtualization units, server compounds, and backup services. Our virtualization platform is hosted in this computational environment. Concretely, the virtual machine used in this work consists of 8 Intel(R) Xeon(R) Platinum 8358 CPU cores @ 2.60GHz, 2 TB of SDD, 64 GB of RAM and the OS is Ubuntu 20.04.4 LTS. All simulations of CS_{MRMR}-DWES_R are coded and executed in Python 3.9.7 environment.

5.3. Results

Table 2 shows the specification of Mango and Avocado datasets (each in three lags) including sample size, feature size, R^2, MAE, RMSE, U-Theil, and execution time before applying CS_{MRMR}-DWES_R using Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX). Additionally, the time series presented in Figure 1a and Figure 1b show a seasonality and an increasing trend for the harvested products in kilograms, which explicitly confirms using SARIMAX as a learning model. It is evident that R^2 decreases and both the errors and the execution time increase as the number of lags grows (more time lags lead to more features). The negative R^2 value in Table 2 indicates that the model is a poor fit for the data and that the features do not explain any of the variances in the response variable. This can happen when the model is overfitting the data, meaning that it is fitting the noise in the data instead of the underlying signal. In other words, the SARIMAX model completely lost its functionality when confronting the large number of features as in Mango_L24 and Avocado_L24 in Table 2 while spending considerable time of execution.

Table 3 shows the average of 50 times running CS_{MRMR}-DWES_R on the datasets. It is evident in Table 3 that R^2, MAE, RMSE, and U-Theil have been improved in comparison

with Table 2. This improvement is more impressive for datasets with a greater number of features. For example, the SARIMAX models were fitted poorly for Mango_L24 and Avocado_L24 before applying the proposed method as the R^2 became negative in Table 2 and made SARIMAX models were non-applicable at all.

Table 2: Descriptions of the datasets as well as the measurement criteria before applying CS_{MRMR}-$DWES_R$.

Dataset	Sample size	Feature size	R^2	MAE	RMSE	U-Theil	exe time (s)
Mango_L4	228	45	0.85	545712	991391	0.21	318
Mango_L12	228	117	0.81	735411	1103535	0.24	791
Mango_L24	228	225	-7.08	5291937	7227867	0.77	3248
Avocado_L4	216	45	0.89	455835	523611	0.11	659
Avocado_L12	216	117	0.13	1194724	1478665	0.29	865
Avocado_L24	216	225	-1.41	2230211	2457994	0.44	2812

Table 3: Average measurement criteria after applying CS_{MRMR}-$DWES_R$.

Dataset	R^2	MAE	RMSE	U-Theil	CS_{MRMR} exe time (s)	$DWES_R$ exe time (s)	CS_{MRMR}-$DWES_R$ exe time (s)
Mango_L4	0.99	253927	305963	0.06	1.03	403.79	404.82
Mango_L12	0.99	254851	302431	0.06	2.84	347.90	350.74
Mango_L24	0.99	249763	298152	0.05	5.90	378.86	384.76
Avocado_L4	0.96	330027	545776	0.1	2.87	390.74	393.61
Avocado_L12	0.96	218959	325313	0.07	2.88	444.66	447.535
Avocado_L24	0.96	221473	326605	0.07	5.94	424.73	430.67

However, reducing non-relevant and non-informative features from datasets improved the prediction of SARIMAX from a negative value to 0.99 and 0.96 for Mango_L24 and Avocado_L24, respectively. According to Table 2 the execution time of SARIMAX increases as the dimensionality of datasets exceeds. However, the execution time of SARIMAX after applying CS_{MRMR}-$DWES_R$ is not only more stable but also generally less in Table 3 than Table 2. Generally, the execution time of CS_{MRMR} is considerably less than $DWES_R$. This is mainly because CS_{MRMR} is a filter approach and does not consider the SARIMAX predictive model for sorting the features. In contrast, $DWES_R$ is a wrapper feature selection method and uses SARIMAX as a predictive model for feature selection. Moreover, $DWES_R$ is a meta-heuristic algorithm that optimizes through

consecutive iterations, which is also time-consuming. Table 3 also numerically shows that CS_{MRMR} spends much less execution time in comparison with $DWES_R$.

Figure 4 also shows the average length of final_set in Algorithm 3 after 50 runs on the datasets when $DWES_R$ converges and stops optimization. According to Figure 4, the average length of the selected features in final_set hardly reaches 1.5. This implicitly confirms clustering the reduced_feature_set in Algorithm 2 into $q = 3$ clusters in Eq. 9 is reasonably sufficient and the maximum length of the final_set in Algorithm 3 is limited to 3 accordingly.

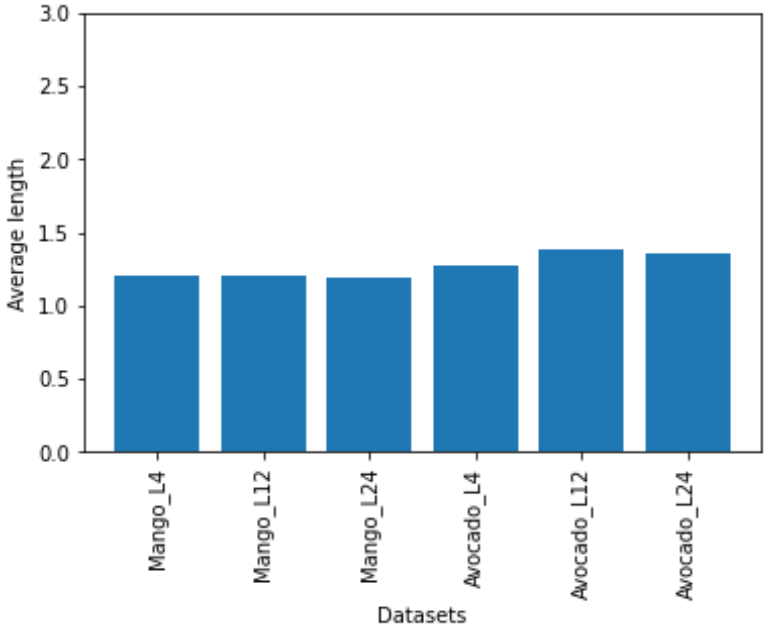


Figure 4: Average length of final_set (selected features by $DWES_R$ in Algorithm 3) after 50 runs

5.4. Comparison

Table 4 presents a comparison between CS_{MRMR}-$DWES_R$ and the state of the art. The criterion used for comparison is R^2. The results of CS_{MRMR}-$DWES_R$ represent the average of 50 runs, while the other non-stochastic techniques in Table 4 were implemented only once. To ensure a fair comparison, the average length of the selected features in CS_{MRMR}-$DWES_R$ is nearly 1, as shown in Figure 4.

In cases where random-state parameter needed to be set (e.g., DT and RF), it was set to 42. The maxlag in GCFS was set to 3. The base estimator for Bagging, SFS (Forward), SFS (Backward), and RFE is RF, while the base estimator for Adaboost is DT (although experiments indicated no significant difference between RF and DT as base estimators). The results in Table 4 clearly show that CS_{MRMR}-$DWES_R$ achieved

Table 4: Comparison of the proposed method CS_{MRMR}-$DWES_R$ (in last column of right) with state-of-the art related methods in term of R^2 achieved.

Dataset	LRC	CC	CS	MI	GCFS	Lasso	DT	RF	XGBoost	Adaboost	Bagging	SFS (Forward)	SFS (Backward)	RFE	CS_{MRMR}-$DWES_R$
Mango_L4	0.98	**0.99**	**0.99**	**0.99**	0.98	0.98	**0.99**	**0.99**	0.98	0.98	0.98	0.98	0.98	**0.99**	**0.99**
Mango_L12	0.98	**0.99**	**0.99**	**0.99**	0.98	0.90	**0.99**	**0.99**	0.98	0.98	0.98	**0.99**	**0.99**	**0.99**	**0.99**
Mango_L24	0.98	0.98	0.98	**0.99**	**0.99**	0.97	**0.99**	**0.99**	0.96	0.98	0.98	0.98	0.98	**0.99**	**0.99**
Avocado_L4	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	**0.96**
Avocado_L12	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	**0.96**	0.95	0.95	0.95	0.95	0.95	**0.96**
Avocado_L24	0.93	0.95	0.95	0.95	0.95	0.95	**0.96**	**0.96**	0.95	0.95	0.95	0.95	0.95	0.95	**0.96**

the highest R^2 in all datasets with good stability. Stability refers to the ability of the method to retain the same R^2 with the same dataset, but with different lags.

5.4.1. Reliability of the selected features by Correlation Statistic

This section investigates how the existing methods in Table 4 can effectively recognize relevant features. Domain knowledge indicates that Ntrees and Surface are individually less informative features. They represent significant changes in yield magnitude (such as massive new plantations), but they do not provide much information on a monthly basis. Figure 5 shows an experiment where we determine the percentage of the top 10% features selected by each method not comprised of NTrees and Surface. For instance, in Mango_L24 datasets, all the features selected by LRC consist of different lags of NTrees and Surface. Therefore, the corresponding blue circle indicates 0 percent. According to Figure 5, CC, CS, MI, and GCFS are the methods that successfully did not select different lags of NTrees and Surface in the top 10%. In conclusion, the following takeaways can be derived from Figure 5:

1. Figure 5 justifies the selection of CS in the first phase of CS_{MRMR}-$DWES_R$. While the percentage of selecting informative features is equally high in CC, CS, MI, and GCFS, the second criterion for selection could be the execution time. Thus, the execution times of CC and CS are insignificant and almost equal (0.02s) and quite less than MI (0.3s) and GCFS (106s) on Avocado_L24.

2. Figure 5 implicitly justifies that the high accuracy of DT and RF in Table 4 are unreliable because they may achieve this high accuracy by selecting non-informative features of NTrees and Surface.

5.4.2. Evaluation of the selected meteorological features

Given the observation that CS_{MRMR}-$DWES_R$ excludes terrain-based features (NTrees and Surface) from its selection, the subsequent experiments illustrated in Figures 6 (Mango datasets) and 7 (Avocado datasets) aim to specifically evaluate the meteorological features chosen by CS_{MRMR}-$DWES_R$. Therefore, the histograms presented in Figures 6 and 7 illustrates the frequency of selected features by $DWES_R$ after 50 runs.

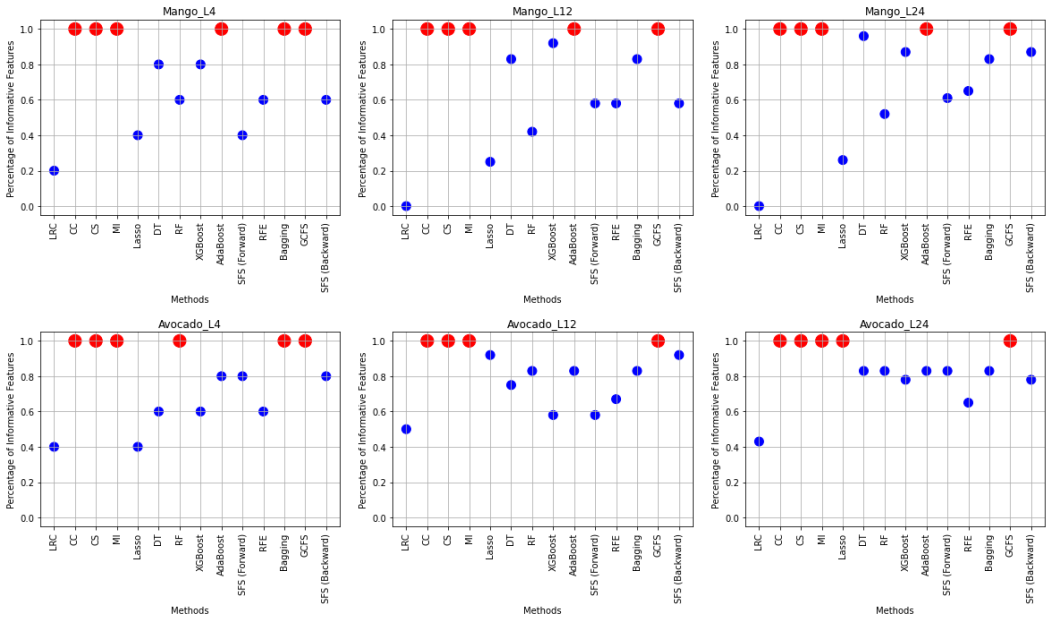


Figure 5: The percentage of selecting informative features (assesible by domain knowledge) in top 10%

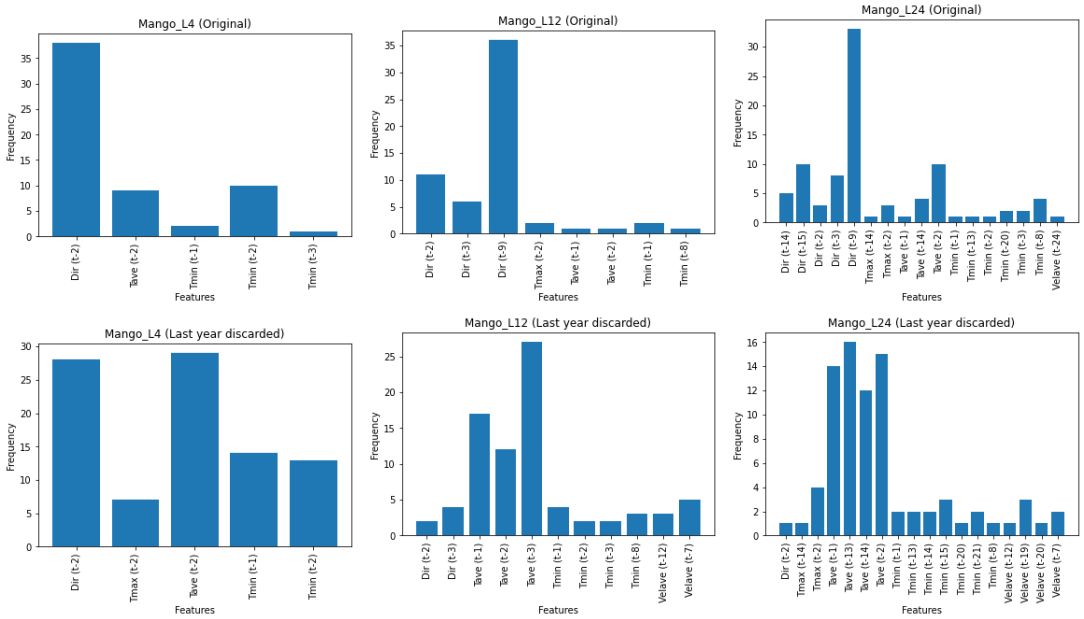


Figure 6: Frequency of selected features after 50 runs by DWES$_R$ for the original and last year discarded Mango datasets in different lags

In both Figures 6 and 7, experiments are conducted on both the original datasets and the datasets with the last year excluded. This investigation aims to determine whether there

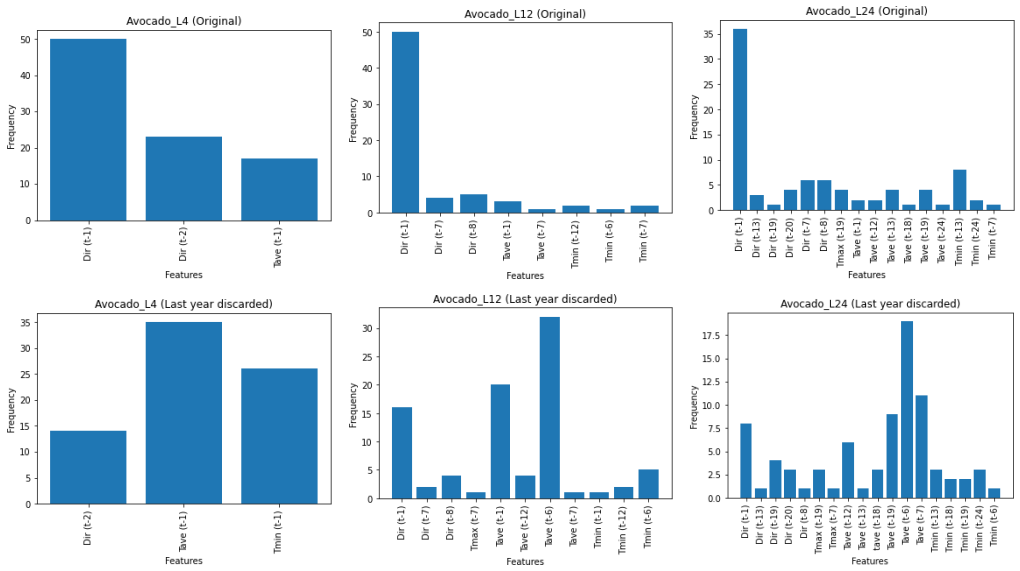


Figure 7: Frequency of selected features after 50 runs of DWES$_R$ for the original and last year discarded Avocado datasets in different lags

are any differences in the frequency of selected features between the original dataset and the dataset without the last year. Both Figures 6 and 7 clearly show that direction of the wind in degrees (*Dir*) and temperatures (*Tave, Tmin, and Tmax*) are among the most informative variables, whether in the original datasets or in the datasets without the last year. However, the concentration differs between the original dataset and the dataset without the last year, with *Dir* being the most informative variable in the original dataset and *Tave* in the dataset without the last year. Thus, it is important to note that the conclusions drawn from the analysis of the current year's data may not necessarily be applicable or valid for subsequent years. Therefore, it is recommended to repeat the experiments and analysis as each new year's data is added to the dataset. This iterative process ensures that the findings remain up-to-date and relevant over time.

Additionally, the analysis of the original Avocado datasets in Figure 7 reveals that even with the inclusion of additional lags, the selected feature (Dir (t-1)) remains consistent. This indicates the significance of the feature, as CS$_{MRMR}$-DWES$_R$ consistently chooses Dir (t-1) despite the incorporation of more features through additional time lags. But, this is not the case with the original Mango datasets in Figure 6, in which Dir (t-2) is the most informative feature in Mango_L4; however, Dir (t-9) is the selected feature in Mango_L12 and Mango_L24.

5.4.3. Interpretaion of the selected meteorological features

Besides the R^2 values reported in Table 4, it is important to assess the trustworthiness and comprehensibility of the selected features for human experts. Furthermore, it is

crucial to determine if these features can acceptably explain the results obtained from the SARIMAX model.

Investigation with agricultural domain experts revealed that both wind direction (*Dir*) and temperature (specifically, *Tave* in our case) effectively contribute to the high accuracy observed in the SARIMAX model. Wind direction (*Dir*) is known to play a critical role in pollination and pest management within agricultural systems. It influences the movement of pollen, thereby affecting the pollination process, and can also impact the spread of pests and diseases in the field. Temperature is another crucial factor that significantly influences the phenological stages of crops. Fluctuations in temperature directly impact critical growth stages, including flowering, fruit set, and ripening. Moreover, temperature profoundly influences the life cycles and activity of pests and disease vectors. The selected features can effectively explain the observed increase in the accuracy of SARIMAX.

Considering the importance of deciphering the selected features, the remaining of this section aims to analyze and interpret the selected features (along with their corresponding lags) by the CS_{MRMR}-$DWES_R$ method. Upon revisiting the original avocado datasets (including the complete time range) in Figure 7, it is evident that Dir (t -1) consistently exhibits the highest correlation with the target variable (Yield), making it the most influential feature. Consequently, one can approximately infer that the wind direction value from the preceding month significantly influences avocado production in the present month. Similarly, the same observation can be made when examining the original mango datasets in Figure 6. It becomes evident that the wind direction plays a crucial role, albeit with varying time lags, leading to distinct interpretations.

5.4.4. Execution time

Figure 8 illustrates the execution time of CS_{MRMR}-$DWES_R$ compared to existing methods, as shown in Table 4. It is generally expected that the execution time of filter and embedded methods would be lower than that of wrapper and hybrid methods (especially if the hybrid method includes a wrapper component). However, this is not always the case in every situation. For instance, the filter GCFS and the wrapper SFS (Forward) take an average of 106 and 42 seconds, respectively, for Mango-L24. In this specific case, SFS (Forward) is faster than CS_{MRMR}-$DWES_R$. This is because SFS (Forward) starts with an empty set and gradually adds the best features, allowing it to stop early once it finds the top feature (first feature). On the other hand, SFS (Backward) begins with all features and progressively eliminates the worst ones until it reaches the top feature. Consequently, in our experiment, where we only consider the best feature, SFS (backward) becomes significantly slower. Thus, the main difference in execution time lies between CS_{MRMR}-$DWES_R$ and SFS (backward). From Figure 8, the following two findings can be explicitly derived:

1. If we aim to find the optimal set of features in datasets with a high number of vari-

ables (such as high-dimensional datasets), SFS (Backward) would be significantly slower compared to CS$_{MRMR}$-DWES$_R$. This observation is supported by Figure 8. For instance, in Mango_L4 and Avocado_L4, where the number of features is not very high, SFS (Backward) exhibits faster execution time. However, in cases like Mango_L12, Mango_L24, Avocado_L12, and Avocado_L24 (where the number of features increases), CS$_{MRMR}$-DWES$_R$ outperforms SFS (Backward) in terms of execution time.

2. The execution time of CS$_{MRMR}$-DWES$_R$ does not vary significantly as the number of features increases. On the other hand, the execution time of SFS (Backward) is directly influenced by the number of features, especially when the objective is to find the optimal feature set.

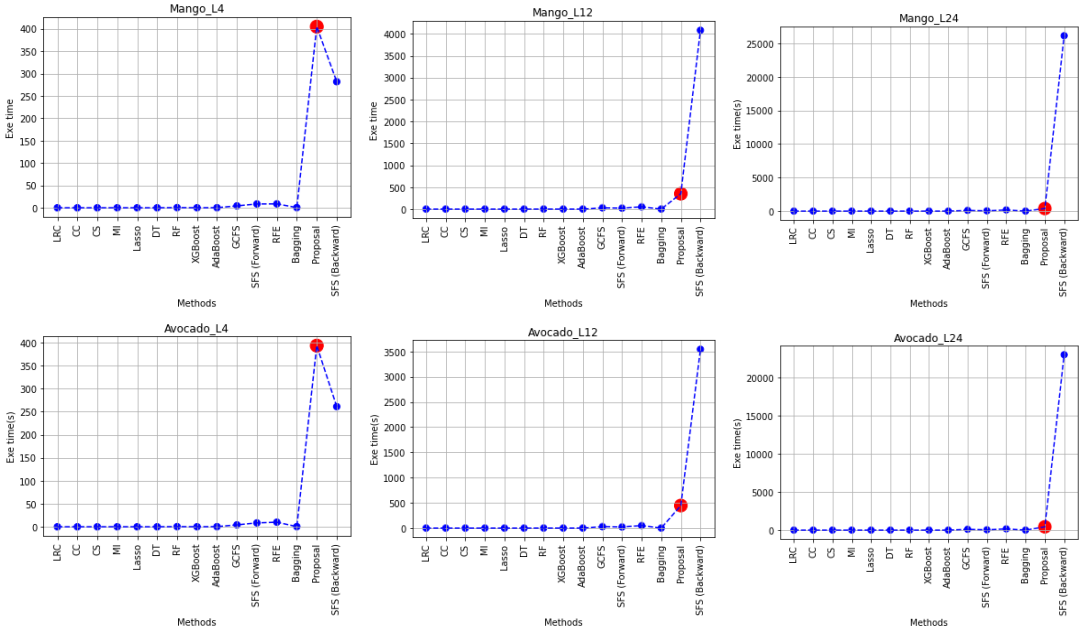


Figure 8: Execution time of automatic CS$_{MRMR}$-DWES$_R$ against existing methods

5.5. *MRMR effectiveness*

MRMR can remove redundant features with a high correlation with each other while preserving a high correlation with the target variable. This can improve the accuracy and generalization of the model while also avoiding overfitting and multicollinearity issues. We conducted and compared two experiments to demonstrate the advantages of our proposed MRMR using the Jaccard similarity measure. It can be observed that there is a significant difference in the top 10% features selected between the Correlation

Statistic (CS) and Correlation Statistic Maximum Relevance Minimum Redundancy (CS_{MRMR}) methods for Mango or Avocado datasets (with different lags). The average difference for the Mango and Avocado datasets is 0.24 and 0.40, respectively (The dotted horizontal black lines in Figure 9). Additionally, it is expected that as the number of features increases, the difference in the selected features generally becomes more prominent. For instance, this difference is particularly noticeable in the Avocado_L24 dataset, where the difference is 0.83. Moreover, considering that Dir (t-1) is identified as the most informative feature for Avocado_L24 from Figure 7, CS_{MRMR} identifies the corresponding feature earlier than CS, which is a consistent trend across all datasets. Specifically, Dir (t-1) ranks $\frac{11}{23}$ (0.48) and $\frac{8}{23}$ (0.35) in the top 10% using CS and CS_{MRMR}, respectively. This indicates that in the specific case of Avocado_L24, CS_{MRMR} efficiently discovers the best feature earlier than CS.

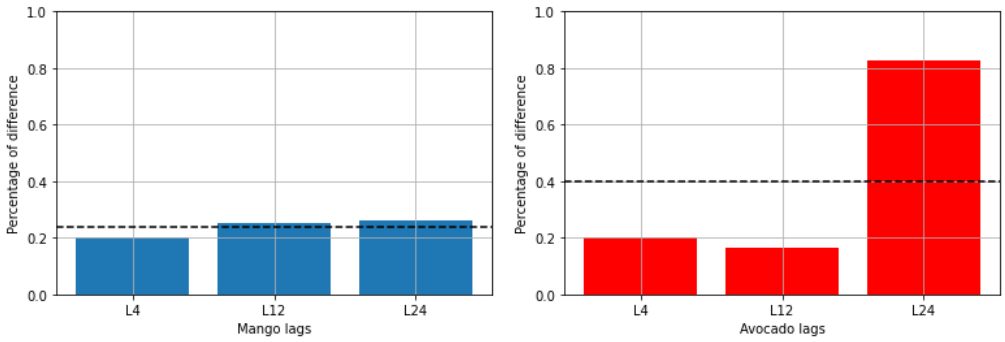


Figure 9: Percentage of difference in top 10% selected features by CS and CS_{MRMR}

6. Conclusions

This paper presents a hybrid feature selection method called Correlation Statistic with Maximum Relevance Minimum Redundancy-Discrete Weighted Evolution Strategy for Regression (CS_{MRMR}-$DWES_R$) designed for time series subtropical data with continuous response variables. The CS_{MRMR}-$DWES_R$ method consists of two phases. In the first phase, a filter supervised method is employed, utilizing the Correlation Statistic (CS) in combination with the Maximum Relevance Minimum Redundancy (MRMR) concept. After feature sorting, only the top 10% of features are retained and passed to the second phase. The second phase involves a wrapper method called Discrete Weighted Evolution Strategy for Regression ($DWES_R$), which automatically identifies the optimal features to enhance the accuracy of the SARIMAX learning model.

CS_{MRMR}-$DWES_R$ is evaluated using datasets from subtropical mango and avocado crops in Spain. The corresponding datasets have been collected and preprocessed incorporating terrain and meteorological features. The proposed method demonstrates its effectiveness in improving the R^2 value and reducing errors (MAE, RMSE, and U-Theil)

of the SARIMAX model. Unlike many existing feature selection methods, CS_{MRMR}-$DWES_R$ avoids selecting non-informative features (terrain features on a monthly basis) and instead focuses on selecting meteorological features, which have proven to explain the improved accuracy of the SARIMAX model effectively.

Future research will look into extending the applicability of the proposed CS_{MRMR}-$DWES_R$ method by exploring its effectiveness on various subtropical crops and diverse geographical regions. New datasets are also planned to be generated that will include variables regarding vegetation indexes and other images from satellital and UAVs observations. This will lead to the necessity of designing new feature selection mechanisms to deal with such heterogeneous data. Similar proposals for new data involving different crops (citrus, olive, etc.) are also considered for future work.

Acknowledgments

The authors thank SAT. 2803 TROPS for their collaboration and technical agronomic support, as well as for the provision of datasets. This work has been partially funded by grants [PYC20 RE 030 UMA] TROPICAL-PREDICT and (funded by MCIN/AEI/10.13039/501100011033/) PID2020-112540RB-C41, AETHER-UMA (A smart data holistic approach for context-aware data analytics: semantics and context exploitation) and Junta de Andalucia, Spain, under contract QUAL21 010UMA. Funding for open access charge: Universidad de Málaga / CBUA.

References

[1] V. H. D. Zuazo, L. Lipan, B. C. Rodríguez, E. Sendra, D. F. Tarifa, A. Nemś, B. G. Ruiz, Ángel Antonio Carbonell-Barrachina, I. F. García-Tejero, Impact of deficit irrigation on fruit yield and lipid profile of terraced avocado orchards, Agronomy for Sustainable Development 41 (2021) 1–16.

[2] V. Cortés, C. Ortiz, N. Aleixos, J. Blasco, S. Cubero, P. Talens, A new internal quality index for mango and its prediction by external visible and near-infrared reflection spectroscopy, Postharvest Biology and Technology 118 (2016) 148–158.

[3] D. Paudel, A. de Wit, H. Boogaard, D. Marcos, S. Osinga, I. N. Athanasiadis, Interpretability of deep learning models for crop yield forecasting, Computers and Electronics in Agriculture 206 (2023) 1–14.

[4] L. Li, B. Wang, P. Feng, H. Wang, Q. He, Y. Wang, D. L. Liu, Y. Li, J. He, H. Feng, G. Yang, Q. Yu, Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across china, Agricultural and Forest Meteorology 308-309 (2021) 1–12.

[5] R. Singh, T. Singh, U. Kaushal, Note on the crop yield forecasting methods, Asian Journal of Agricultural Research 13 (2019) 1–5.

[6] M. Mokria1, A. Gebrekirstos, H. Said, K. Hadgu1, N. Hagazi, W. Dubale, A. Bräuning, Fruit weight and yield estimation models for five avocado cultivars in ethiopia, Environmental Research Communications 4 (7) (2022) 1–16.

[7] S. Rathod, S. Vijayakumar, N. Bandumula, G. Chitikela, Prediction of mango production using machine intelligence techniques: A case study from karnataka, india, Acta Scientific Agriculture (2022) 16–22doi:10.31080/ASAG.2022.06.1174.

[8] D. C. Corrales, C. Schoving, H. Raynal, P. Debaeke, E.-P. Journet, J. Constantin, A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern france, Computers and Electronics in Agriculture 192 (2022) 1–19.

[9] J. Ircio, A. Lojo, U. Mori, J. A. Lozano, Mutual information based feature subset selection in multivariate time series classification, Pattern Recognition 108 (2020) 1–12.

[10] N. Gu, M. Fan, L. Du, D. Ren, Efficient sequential feature selection based on adaptive eigenspace model, Neurocomputing 161 (2015) 199–209.

[11] S. Shafiee, L. M. Lied, I. Burud, J. A. Dieseth, M. Alsheikh, M. Lillemo, Sequential forward selection and support vector regression in comparison to lasso regression for spring wheat yield prediction based on uav imagery, Computers and Electronics in Agriculture 183 (2021) 1–9.

[12] F. Amini, G. Hu, A two-layer feature selection method using genetic algorithm and elastic net, Expert Systems with Applications 166 (2021) 1–10.

[13] Y. Hmamouche, P. Przymus, A. Casali, L. Lakhal, GFSM: a feature selection method for improving time series forecasting, International Journal on Advances in Systems and Measurements 10 (2017) 255–264.

[14] Y. Hmamouche, L. Lakhal, A. Casali, A scalable framework for large time series prediction, Knowl Inf Syst 63 (2021) 1093–1116.

[15] H. Nematzadeh, J. García-Nieto, I. Navas-Delgado, J. F. Aldana-Montes, Automatic frequency-based feature selection using discrete weighted evolution strategy, Applied Soft Computing 130 (2022).

[16] S. Osama, H. Shaban, A. A. Ali, Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review, Expert Systems with Applications 213 (2023).

[17] S. Abasabadi, H. Nematzadeh, H. Motameni, E. Akbari, Hybrid feature selection based on sli and genetic algorithm for microarray datasets, The Journal of Supercomputing 78 (2022) 19725–19753.

[18] M. Di Mauro, G. Galatro, G. Fortino, A. Liotta, Supervised feature selection techniques in network intrusion detection: A critical review, Engineering Applications of Artificial Intelligence 101 (2021) 1–15.

[19] M. Labani, P. Moradi, F. Ahmadizar, M. Jalili, A novel multivariate filter method for feature selection in text classification problems, Engineering Applications of Artificial Intelligence 70 (2018) 25–37.

[20] M. Huang, Theory and implementation of linear regression, in: 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2020, pp. 210–217.

[21] Y. Liu, Y. Mu, K. Chen, Y. Li, J. Guo, Daily activity feature selection in smart homes based on pearson correlation coefficient, Neural Process Lett 51 (2020) 1771–1787.

[22] X. lianLi, D. Qiao, Y. Ding, Y. Shi, W. Guo, D. Wei, Design and research of statistical analysis system based on business decision field, Journal of Software 15 (6) (2020) 172–180.

[23] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.

[24] Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, R. Wang, Using causal discovery for feature selection in multivariate numerical time series, Mach Learn 101 (2015) 377–395.

[25] A. S. Chivukula, J. Li, W. Liu, Discovering granger-causal features from deep learning networks, Advances in Artificial Intelligence 11320 (2018) 692–705.

[26] P. M. Gopal, R. Bhargavi, A novel approach for efficient crop yield prediction, Computers and Electronics in Agriculture 165 (2019) 104968.

[27] M. G. PS, R. Bhargavi, Selection of important features for optimizing crop yield prediction, International Journal of Agricultural and Environmental Information Systems (IJAEIS) 10 (3) (2019) 54–71.

[28] P. Theerthagiri, Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique, Intelligent Systems with Applications 16 (2022) 200121.

[29] S. Raja, B. Sawicka, Z. Stamenkovic, G. Mariammal, Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers, IEEE Access 10 (2022) 23625–23641.

[30] F. Shahsavari, Z. Shaghaghian, Application of classification and feature selection in building energy simulations (2021). arXiv:2108.12363.

[31] I. D. Mienye, Y. Sun, Z. Wang, Prediction performance of improved decision tree-based algorithms: a review, Procedia Manufacturing 35 (2019) 698–703, the 2nd International Conference on Sustainable Materials Processing and Manufacturing, SMPM 2019, 8-10 March 2019, Sun City, South Africa.

[32] C. W. Tan, C. Bergmeir, F. Petitjean, G. I. Webb, Time series extrinsic regression, Data Min Knowl Disc 32 (2021) 1032–1060.

[33] I. K. Nti, A. F. Adekoya, B. A. Weyori, A comprehensive evaluation of ensemble learning for stock-market prediction, J Big Data 7 (20) (2020) 1–40.

[34] H. Allende, C. Valle, Ensemble Methods for Time Series Forecasting, Springer International Publishing, 2017, pp. 217–232.

[35] S. Mohapatra, N. Chaudhary, Statistical analysis and evaluation of feature selection techniques and implementing machine learning algorithms to predict the crop yield using accuracy metrics, Engineered Science 21 (2022) 787.

[36] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems, Pearson Education Limited, 2011.

[37] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238. doi:10.1109/TPAMI.2005.159.