INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Determination of crop coefficient (kc) based on machine learning NDVI Time Series models

Guilherme Filipe Tomé Duarte

Master Degree in Data Science

Orientador(a):
PhD, Diana Elisabeta Aldea Mendes, Associate Professor,
Iscte-IUL

Co-Orientador(a):
PhD, Pedro Chambel Filipe Lopes Leitão

October, 2022

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Determination of crop coefficient (kc) based on machine learning NDVI Time Series models

Guilherme Filipe Tomé Duarte

Master Degree in Data Science

Orientador(a):
PhD, Diana Elisabeta Aldea Mendes, Associate Professor, Iscte-IUL

Co-Orientador(a):
PhD, Pedro Chambel Filipe Lopes Leitão

October, 2022

Acknowledgments

I would like to thank my supervisors Professor Diana Mendes and Doctor Pedro Chambel for their guidance and patience during the development of this dissertation. I also want to thank Iscte-Iul, for providing this opportunity. To my friends, thank you for the conscious decision of sharing your time with me. To my family, words cannot express how deeply and truly grateful I am for all your love, guidance, and support.

Resumo

O objetivo da dissertação é enfrentar o desafio necessário de fornecer modelos de coeficiente de cultura (K_c) baseados em refletância para reduzir o consumo de água na irrigação agrícola. Neste trabalho, foram criados 6 modelos diferentes para cada uma das culturas usando o índice de vegetação por diferença normalizada (NDVI) para estimar os coeficientes de cultura para milho, tomate, batata e girassol na região da Lezíria do Tejo combinando diferentes métodos de pré-seleção de séries temporais e usando a média e k-*means* para criar novas séries temporais, bem como usar regressão linear e polinomial para ajustar as novas séries temporais geradas com curvas K_c teóricas com o objetivo de usar esses modelos para determinar K_c nesta região. O desempenho desses modelos foi avaliado usando o coeficiente de determinação (R^2), a raiz quadrada do erro quadrático médio (RMSE) e uma inspeção visual das previsões no conjunto de teste.

Os resultados mostram que os modelos K_c-NDVI criados conseguiram capturar bem as curvas teóricas de K_c, bem como o uso de uma pré-seleção das séries temporais, média e k-means para estas culturas são úteis para capturar as curvas dos coeficientes de cultura, uma vez que alguns dos melhores resultados obtidos foram quando estas foram utilizadas. As melhores metodologias dependem de cada cultura e não existe uma que seja globalmente melhor. Estes resultados obtidos são promissores e podem ser vistos como métodos potenciais para melhor determinar os coeficientes de cultura e os modelos são adequados para seu uso pelo menos na região estudada.

Palavras-chave: NDVI, K_c, machine learning, séries temporais, sensoriamento remoto.

Abstract

This dissertation aims to meet the required challenge of providing reflectance-based crop coefficient models to reduce water consumption in agriculture irrigation. In this work, 6 different models were created for each crop by using normalized difference vegetation index (NDVI) to estimate crop coefficients (K_c) for maize, tomato, potato and sunflower for Lezíria do Tejo region combining different pre-selection methods of time series and mean and k-means to create new time series and use linear and polynomial regression to fit the new generate time series with theoretical K_c curves to use these models to determine K_c in this region. These models' performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE) and a visual inspection of test set predictions.

The results show that the K_c-NDVI models created were able to capture the theoretical curves of K_c well, and the use of a pre-selection of time series, mean and k-means for these crops is useful to capture the curves of the crop coefficients since some of the best results obtained were when they were used. The best methodologies depend on each crop; no one is globally better than the others. The results shown are promising and can be seen as potential methods to better determine crop coefficients and the models are suitable for their use at least in the region of this study.

Keywords: NDVI, K_c, machine learning, time series, remote sensing.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Motivation

The development of remote sensing in recent years has created the possibility of acquiring and storing large amounts of data, obtaining satellite images with a higher spatial, spectral, and temporal resolution, and enabling satellite data access to everyone in society (Cracknell, 2018). Therefore, the increased research in remote sensing is not surprising, evapotranspiration included (Z. Zhu et al., 2019).

The measurement of evapotranspiration (ET) is extremely important since it is a significant component for estimating crop water requirement. Crop water requirement can be defined as the amount of water necessary to compensate for evapotranspiration loss. If we subtract effective precipitation from it, it fundamentally represents the irrigation water requirement, that is, the amount of water necessary from irrigation to compensate for crop evapotranspiration and supplementary water needs (R. G. Allen et al., 1998).

Its importance has grown in the last years due to climate change causing fluctuation in its measurement and the necessity for water resource management (Piticar et al., 2016). An essential factor to consider is the complexities in the land-plant-atmosphere system, which makes the measurement of actual evapotranspiration rather difficult (Ahmed et al., 2021). On the other hand, forecasting interpretation and accuracy depend heavily on the appropriate selection of predictor variables (Prasad et al., 2018).

Irrigated agriculture accounts for around 70% of all available freshwater, making it the principal water consumer worldwide (AQUASTAT - FAO's Global Information System on Water and Agriculture, n.d.). Due to climate change, population growth, and increasing pressure on available water resources, the importance of irrigation water management in agricultural fields and sustainable agricultural development has become apparent (Kharrou et al., 2021). Despite its importance, a lack of spatiotemporal patterns of evapotranspiration, mainly in large irrigated areas, as well as the factors which impact its measurement, has been frequently noted (Cheng et al., 2021; Saboori et al., 2021). For that reason, we suggest that more studies are needed to explore evapotranspiration.

One way of determining ET is by using reflectance-based crop coefficient methods. One of the methods is called single crop coefficient where the ET can be obtain through the product of a crop coefficient (K_c) with a reference evapotranspiration (ET_0), which is measured using the FAO-Penman-Monteith equation (R. G. Allen et al., 1998). K_c can be estimated using vegetation indexes (VI) which are mathematical representations of multiple spectral bands.

This dissertation focuses on the study of creating models by fitting normalized difference vegetation index (NDVI) to K_c called NDVI-K_c models, which are sub-group of VI-K_c models. That means that we aim to be able to estimate K_c, as well as evaluate different approaches for time series pre-selection. VI-K_c models are known for the easily of obtaining VI values through satellites like Sentinel-2, where the data obtained from these satellites are open to the public (Pôças et al., 2020). However, these models are generally region-specific and affected by climate conditions (Richard G. Allen et al., 2011a) and due to the effect of climate change in the last years (Piticar et al., 2016), these models need to be constantly updated. To the best of our knowledge, at least in the last two years, no NDVI-K_c models for maize, tomato, potato, and sunflower for Lezíria do Tejo region were created. To address this issue, the dissertation focuses on building NDVI-K_c models for these crops and this region from data obtained from Sentinel-2 using linear and polynomial regression, as other studies have done with success (Pôças et al., 2020). We also aim to evaluate the use of different pre-selection techniques and k-means, which have not been addressed in the literature, at least for these crops in this region.

1.2. Goals

The main focuses of this dissertation are the followings:

- Development of NDVI-K_c models for maize, tomato, potato, and sunflower for the Lezíria do Tejo region
- Study the use of a pre-selection of time series and k-means to discover new temporal patterns in the data and the overall impact on the models.

1.3. Dissertation organization

This dissertation is organized into the following chapters. A literature review is presented in Chapter 2, which addresses the fundamental concepts related to this work, namely: remote sensing, evapotranspiration, crop coefficient, vegetation indexes and different methodologies do determine evapotranspiration. Chapter 3 explains the methodology from how the data was acquired, then pre-processed, and right after how the NDVI-K_c models were created and evaluated. In Chapter 4, the performance of each model is evaluated and compared using the coefficient of determination (R^2) and root mean square error (RMSE) and a visual inspection of test set predictions. The conclusions and future work are presented in Chapter 5.

1.4. Main contributions

We consider that this work has the following main contributions:

- Implementation and validation of the effectiveness of NDVI-K_c models for maize, tomato, potato, and sunflower for the Lezíria do Tejo region that may translate well to other regions with similar climate conditions.
- Demonstrating that using a pre-selection of time series and using k-means provide better results in some cases.

CHAPTER 2

Literature Review

2.1 Remote Sensing

Since 1970, satellites with medium-high spatial resolution and with multispectral observations in the near-infrared (NIR) and visible (VIS) spectra started to be available with satellites like the Landsat series and later SPOT satellites being able to capture data with a spatial resolution of 30 m (with 16-day revisit time) and 20 m, respectively. Meaning that, in the case of the Landsat, each pixel in the image corresponds to a 30x30 meter square on the ground. Unfortunately, this satellite imagery was costly, with low availability and long revisiting times. Nowadays, satellites like ESA Sentinel-2 have free and open access policy and with better spatial, temporal, and spectral resolutions. Sentinel-2 is composed of two satellites (Sentinel-2A and Sentinel-2B) with a multispectral sensor being able to capture data at 10 m (VIS and Broad NIR), 20 m (red edge, narrow NIR, and Short-Wave InfraRed (SWIR) bands), and 60 m (atmospheric bands) with 13 spectral bands in total and with a temporal resolution of 5 days (Transon et al., 2018). These new satellites open doors for society and promotes increased research in multiple fields related to the study of the earth.

2.2 Evapotranspiration

The Evapotranspiration concept consists of two parts that occur simultaneously and are hard to distinguish: evaporation and transpiration. The first one is the conversion of liquid water to water in a gaseous state (vapor) called vaporization and the consequent removal from the surface where it was. The second one consists of the release of water in a liquid state from the plant tissues through vaporization into the atmosphere (R. G. Allen et al., 1998) and plays a key role in water management in agriculture (Zhao et al., 2019).

Initial studies mentioning evapotranspiration can be traced back to 1937, however, without a proper definition and explanation. It was then first defined in 1944 by Thornthwaite, which quickly was adopted, primarily thanks to the efforts of Penman and Monteith at the time (Stanhill, 2005).

There are two main groups in terms of methodology that are used to measure ET. First, we have *in-situ* techniques which measure ET at the local scale, including micrometeorological methods (e.g., eddy covariance), hydrological methods (e.g., lysimeters), and physiological methods (e.g., sap flow) (Rana & Katerji, 2000). Due to the high accuracy of these methods, they have become the standards of ET measurement and are used to provide baseline information on how accurate the remote sensing approaches are to the measurement of ET. However, there are some limitations: first, they

have a hard time simulating mixed terrain, soil features, water content, and meteorological conditions, there fourth not suited for larger fields, and for last the instrumentation is expensive (Bhattarai et al., 2016; Drexler et al., 2004; Wagle et al., 2017). The other group is the remote sensing one which can be divided into three different categories according to the variables used:

- First, we have the remote sensing-based Penman-Monteith direct methods, which are used in seasonally varied vegetation (R. Zhu et al., 2013) and are known for being physically rigorous models that take into account the potential ET's relationships with the soil-surface temperature and net radiation heat flux with the counterpart of requiring an enormous number of climatic variables (Kazemi et al., 2021; Tikhamarine et al., 2020).
- Second, we have the reflectance-based crop coefficient method, the one followed in this work. This method use vegetation indexes, which is possible due to the high correlation between the spectral response of the vegetation and vegetation transpiration, obtained through the fraction of active photosynthetic radiation absorbed by the canopy (Glenn et al., 2011). They have the advantage of providing continuous and robust estimates of ET since crop transpiration can firmly be extrapolated between satellite overpass times and with a smoother curve over time (Glenn et al., 2011; Hunsaker et al., 2003; Nagler et al., 2005). Unfortunately, this method needs corrections, depends on crop-specific relationships, and the accuracy is reduced after wet events like rain or differences in soil moisture (Richard G. Allen et al., 2011a).
- Finally, the surface energy balance (SEB) method, where the latent flux, calculated as the residual term of the surface energy balance equation, is used to estimate ET (Kharrou et al., 2021). These models use remote sensing to obtain albedo, land surface temperature, leaf area index, and vegetation cover fraction and have the advantage of most of them being independent of ground measurements. Consequently, indicators related to the measurement of ET can be calculated in extensive areas. The downside is that the ones that use thermal and visible/near-infrared data can only give precise measurements during clear-sky conditions (Jurečka et al., 2021). The main advantage of all remote sensing methods is that they are generally cost-effective and enable a good estimative of ET both in spatial and temporal aspects (Kharrou et al., 2021).

2.3 Crop coefficient

One the most used reflectance-based crop coefficient methods is the FAO-56 method (R. G. Allen et al., 1998), which says that crop evapotranspiration (evapotranspiration of a specific crop), denoted as ET_c (mm day^{-1}), is determined by the product of a crop coefficient K_c by the reference evapotranspiration ET_0 (mm day^{-1}) as shown in following formula:

$$\boldsymbol{ET_c = K_c ET_0} \quad (2.1)$$

Where ET_c is the evapotranspiration of a crop under standard conditions (crops grown in large fields in optimal soil water and agronomic conditions). K_c is a coefficient that represents the physical and physiological differences such as stomatal characteristics, aerodynamic properties, leaf anatomy and albedo between a specific crop and the reference crop. On the other hand ET_0 is represented as the hypothetical grass reference crop evapotranspiration with some underlying assumptions such as an albedo of 0.23, a crop height of 0,12 m and a fixed surface resistance of 70 s m^{-1} representative of a moderately dry soil surface as a consequence of a weekly irrigation frequency and is determined using the FAO Penman-Monteith equation through meteorological data (R. G. Allen et al., 1998) and calculated as:

$$\boldsymbol{ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34 u_2)}} \quad (2.2)$$

where R_n is the net radiation at the crop surface (MJ m^{-2} day^{-1}), G is the soil heat flux density (MJ m^{-2} day^{-1}), T the mean daily air temperature at 2 m height (°C), u_2 is the wind speed at 2 m height (m s^{-1}), e_s the saturation vapor pressure (kPa), e_a the actual vapor pressure (kPa), e_s - e_a the saturation vapor pressure deficit (kPa), Δ the slope vapor pressure curve (kPa $°C^{-1}$) and γ the psychrometric constant (kPa $°C^{-1}$). The advantage of estimating ET_0 using FAO Penman-Monteith equation is to provide a standard so that evaporation in a different spatial and temporal setting and between crops can be compared (R. G. Allen et al., 1998).

The K_c can be used in two different ways. A single approach (single crop coefficient), as shown before, where a single K_c integrates the relationship between ET_c and ET_0, and a dual approach (dual crop coefficient) where K_c is constituted of the sum of two individual components: one that represents the evapotranspiration from a well-irrigated crop in a dry soil named basal crop coefficient (K_{cb}) and one that represents the evaporation from the soil that is exposed named soil evaporation coefficient (K_e) (Wang et al., 2021) such as:

$$K_c = K_{cb} + K_e \quad (2.3)$$

A way to determine K_c is by using vegetation indexes obtained through remote sensing data that are mathematical transformations of multiple spectral bands. Studies have shown that these are good indicators of relative abundance and growth stage of green vegetation and, consequently, radiation absorption. This data can be obtained from multiple sources such as satellites and

unnamed aerial vehicles (UAV) and due to scientific development, its use has continued to increase (Pôças et al., 2020).

The first studies, in this specific area, can be found in the works done by Bausch & Neale, (1987). Since then, multiple ways of using VI such as the NDVI, Soil Adjusted Vegetation Index (SAVI), and Enhanced Vegetation Index (EVI) to estimate K_c has been used. From those, NDVI is the most used one (Goward et al., 1991) as well as provides a better basis for the K_c -NDVI relationship than other VI (e.g SAVI) (Richard G. Allen et al., 2011b) and for that reason is the one used on this study. The authors Bausch & Neale (1987) found that NDVI is highly correlated to leaf area index (corresponding to the density of the vegetation cover) and fractional cover (corresponding to the fraction of ground covered by green vegetation), making it a good indicator of vegetation growth. NDVI is determined as:

$$NDVI=\frac{(\rho_{NIR}-\rho_{red})}{(\rho_{NIR}+\rho_{red})} \qquad (2.4)$$

where ρ_{NIR} is the reflectance at Near-infrared (*NIR*) spectral domain and ρ_{red} the reflectance at red spectral domain. Although VI has been shown to provide better relationships with K_{cb} than K_c (Choudhury et al., 1994), due to the simplify of the eventual intention of the determination of ET_c, the K_c values of R. G. Allen et al. (1998) study were used.

The primary reason for their use in estimating K_c (K_c-VI models) is their interpretability, computational simplicity and availability in most constellations of satellite missions with moderate resolution (10-300 m) measurements in short periods. It means that they can capture the time and space variation in the development of crops with more straightforward automation of data pre-processing with low latency and consequently suited for irrigation management in an agricultural setting (Calera et al., 2017; Johnson & Trout, 2012; Murray et al., 2009; Rafn et al., 2008). However, these models come with some difficulties since some of them are local and crop-dependent, with the necessity of complementary methods or failing due to the presence of clouds (Pôças et al., 2020).

There are many K_c -VI models. Some of them can be created from relationships based on canopy development information created from vegetation indexes and empirical (linear and no-linear) relationships between crop coefficient and vegetation indexes and others that combine with soil water balance models (SWB) or thermal-based approaches which can be found in Pôças et al. (2020).

Since previous studies using linear and polynomial regressions presented good results, in this work, we focus on bringing some contribution to the establishment of K_c -NDVI models for tomato, potato, maize, and sunflower for the considered region (Lezíria do Tejo), since at least in last two years, no research on this topic were found. We also evaluate the use of different pre-selection techniques and k-means, which to the best of our knowledge, has not been addressed in the literature.

CHAPTER 3

Methodology

In this chapter, a description of the methodology taken will be presented. It is subdivided into three sections. In the first section (3.1) an explanation of the data and how was acquired can be found. In the second section (3.2) a description of the pre-processing of the data process to increase the quality of the previously obtained data can be found. Finally, a specification of which models were used and how they were evaluated is in section 3.3.

3.1 Acquisition and Selection of the Data

The data was obtained from the AQUAFARM database (Aquafarm Database, n.d.) accessed on 24/01/2022. The data consisted of the identifier of each polygon/crop field (id), type of culture, latitude (lat) e longitude (long), and a mean normalized difference vegetation index (NDVI) for the polygon registered with an interval of 5 days between each record (date) from 01/09/2020 to 24/01/2022. The NDVI values were obtained from the Sentinel-2 mission from the Copernicus Programme (Homepage | Copernicus, n.d.). All polygons are in the Lezíria do Tejo in Portugal, according to the Nomenclature of Territorial Units for Statistics (NUTS III) (Figure 3.1). Although many crops were available only 4 crops (potato, maize, tomato, sunflower) were studied. These were chosen because they are some of Portugal's most important and cultivated crops in the country. Thus, mitigating the waste of water on them would be beneficial to fighting the soil drought since the impact of climate change in the country has been increasing in recent years. In total, 7771 polygons were used (4779 for maize, 2380 for tomato, 445 for potato, and 168 for sunflower), which in turn created 7771 NDVI time series as illustrated in Figures 3.2-3.5. It´s possible to see in Figures 3.2-3.5 that there isn't a common temporal pattern for the time series and therefore defining precisely in time the development period of a crop it´s hard.

The theoretical K_c values in each stage used for each crop and the length in each development stage were extracted from the FAO-56 paper guidelines (R. G. Allen et al., 1998) and are shown in Table 3.1 and Table 3.2, respectably.

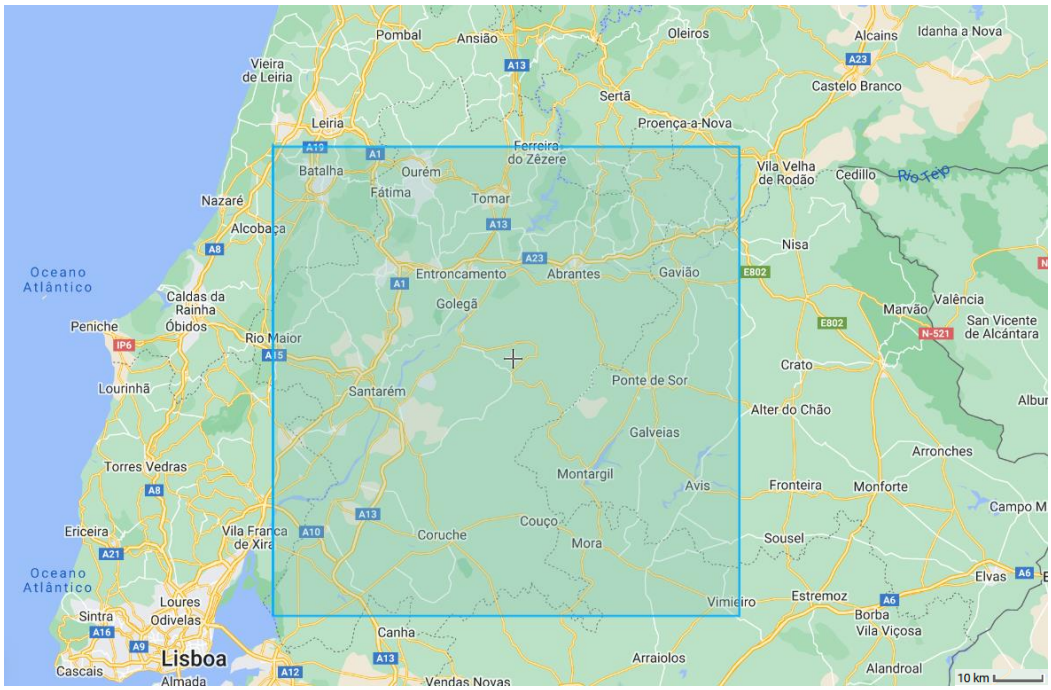

Figure 3.1. Lezíria do Tejo region. The marked zone represents the studied area.

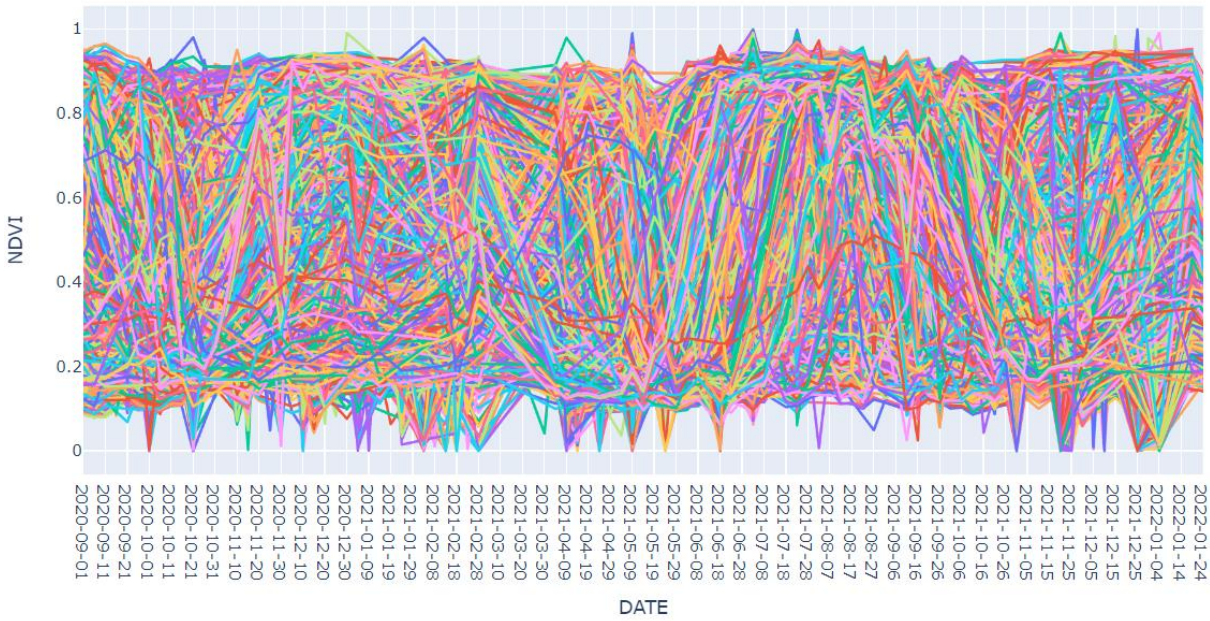

Figure 3.2. NDVI time series for the maize crop.

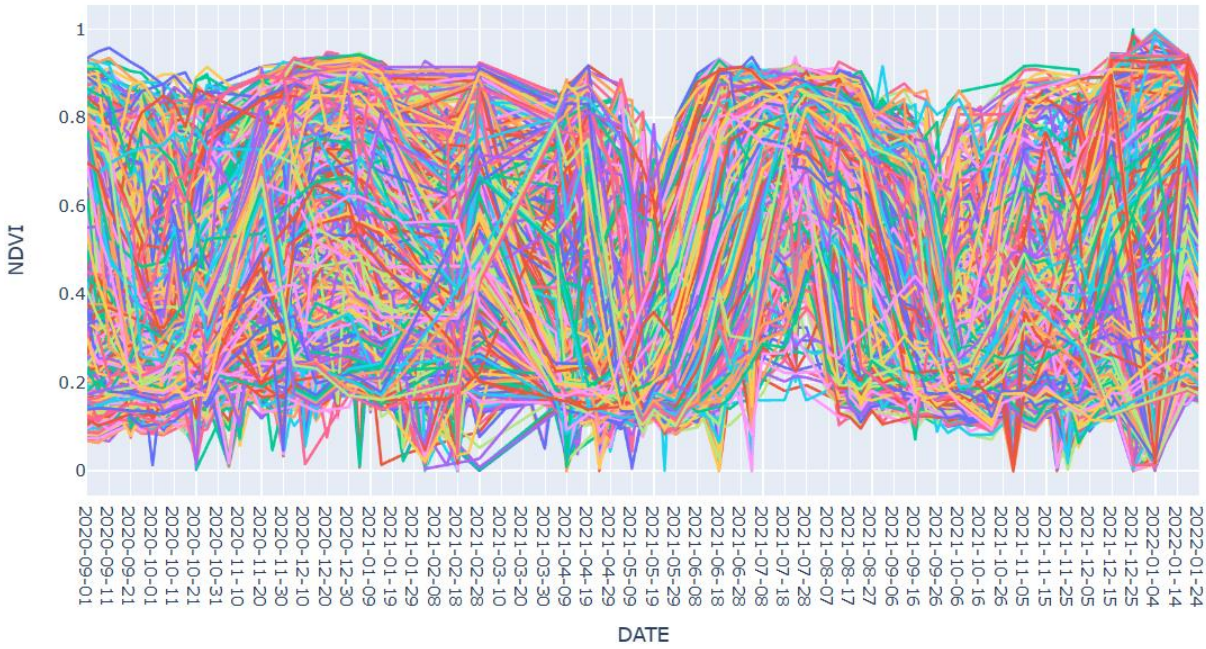

Figure 3.3. NDVI time series for the tomato crop.

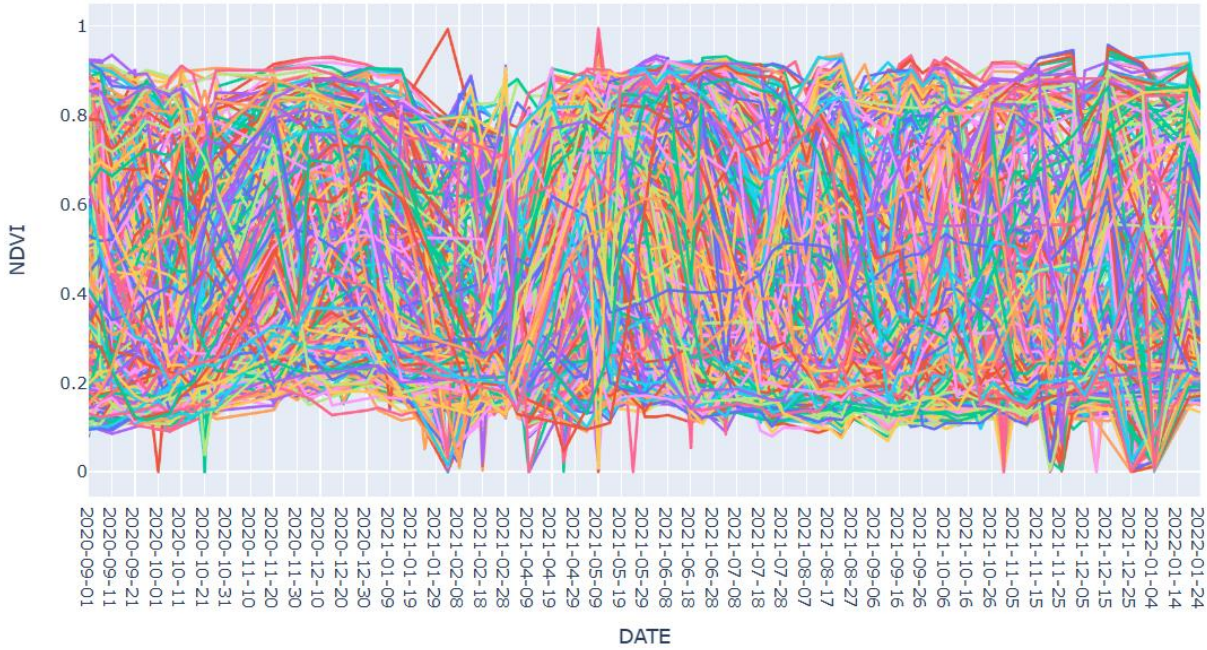

Figure 3.4. NDVI time series for the potato crop.

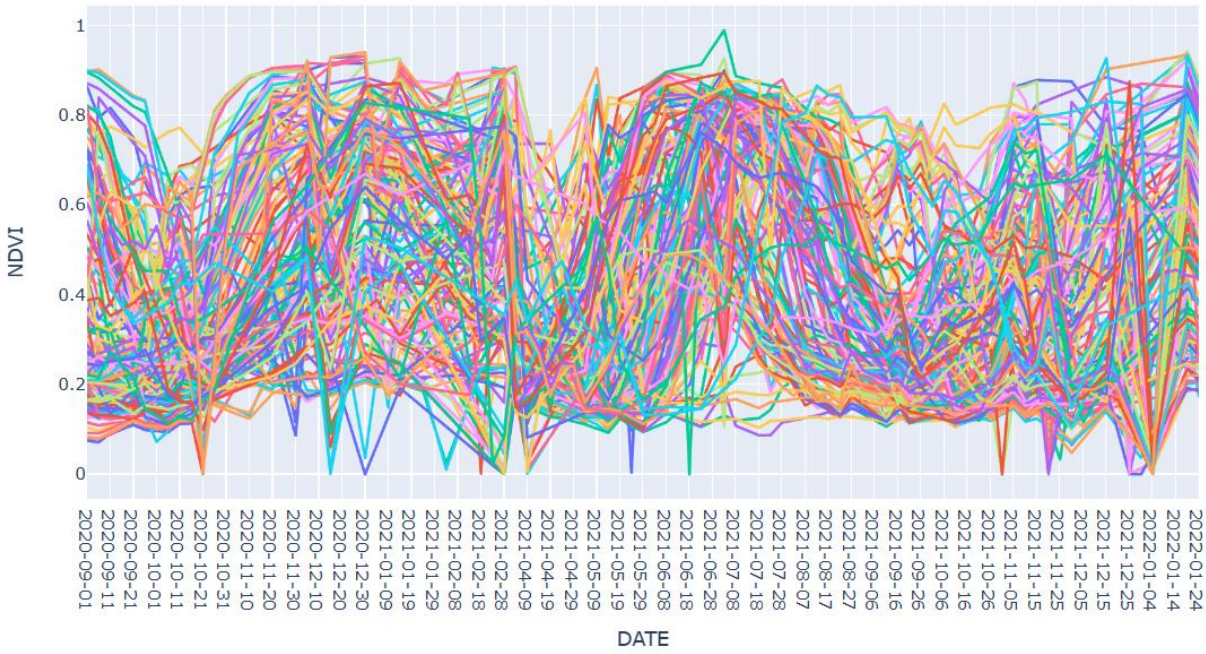

Figure 3.5. NDVI time series for the sunflower crop.

Table 3.1. The theoretical K_c values in each stage used for each crop ($K_{c\ ini}$ – K_c values in the initial stage, $K_{c\ mid}$ – K_c values in the mid-season stage, $K_{c\ end}$ – K_c values in the end of the late season stage)

Crop	$K_{c\ ini}$	$K_{c\ mid}$	$K_{c\ end}$
Maize	0.3	1.2	0.48
Sunflower	0.35	1.08	0.35
Potato	0.5	1.15	0.75
Tomato	0.6	1.15	0.8

Table 3.2. The length of each stage used for each crop (L_{ini} – number of days in the initial stage, L_{dev} – number of days in the development stage, L_{mid} – number of days in the mid-season stage, L_{late} – number of days in the late season stage)

Crop	L_{ini} (days)	L_{dev} (days)	L_{mid} (days)	L_{late} (days)	Total (days)
Maize	30	40	50	30	150
Sunflower	25	35	45	25	130
Potato	30	35	50	30	145
Tomato	30	40	45	30	145

3.2. Data Preparation

First, the data was analyzed to verify if no incorrect values were present (NDVI values inferior to -1 and superior to 1) as well as the presence of Not a Number (NaN) values. Between these two, only the last one was detected. The replacement of these values was concluded using linear interpolation since, in the few series where they were present, this type of replacement would enable the conservation of the tendency of the time series in that period. Although the NDVI values were within the correct range, the temporal patterns were very different from each other's, meaning that existing crops were registered incorrectly in the database. Unfortunately, there isn't an easy solution to solve this problem and for that reason, all time series were taken into consideration.

Afterward, a smoothing was applied to the newly generated time series to filter the noise they might contain due to, for example, the presence of clouds. The Savitzky-Golay (S-G) filter (Savitzky & Golay, 1964) was applied since other studies have used it successfully (Han et al., 2020). The S-G filter is a weighted moving average filter algorithm, where the weights coefficients are obtained within the moving filter window by the least-squares fitting of a polynomial (Chen et al., 2004) and it was implemented using the scipy python library. The general equation can be represented as follows:

$$Y_j^* = \frac{\sum_{i=-m}^{i=m} C_i Y_{j+i}}{N} \qquad (3.1)$$

where Y is the original NDVI value, Y* is the resultant NDVI value, C_i is the coefficient for the ith NDVI value of the filter (smoothing window), and N is the number of convoluting integers and is equal to the smoothing window size (2m+ 1). The index j is the running index of the original ordinate data table. The smoothing array (filter size) consists of 2m+ 1 points, where m is the half-width of the smoothing window (Chen et al., 2004).

The filter has two essential parameters: the window size (w) of the moving window and the degree of the smoothing polynomial (p). The filter was tested with different ratios of w and p, starting with a ratio w/p of 2.5 and incrementally raising by 1 (3.5; 4.5; …) until, through visual inspection, the filters seemed to be able to remove most of the noise. An example of a time series for the maize crop is illustrated in Figure 3.6.

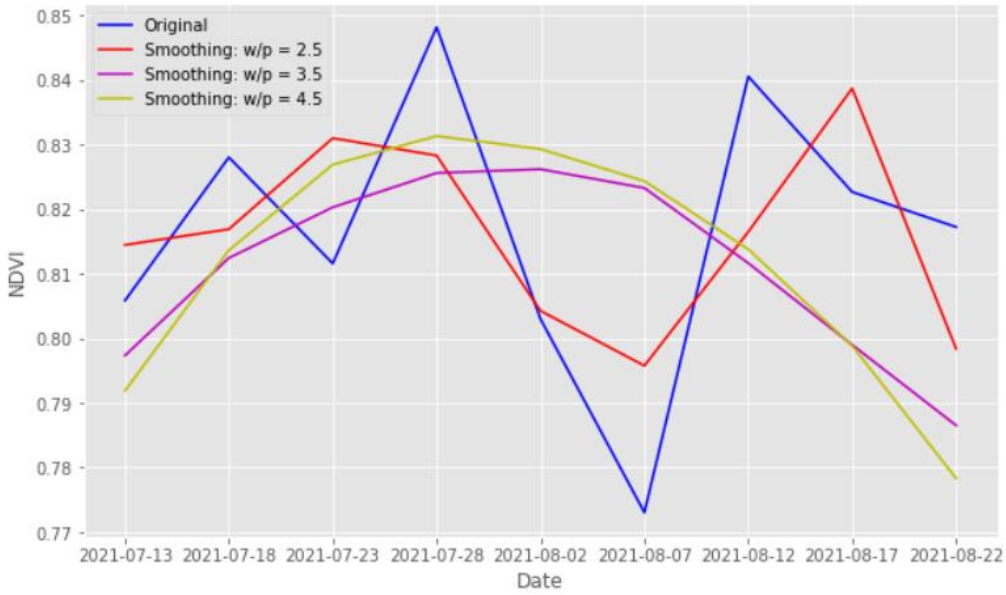

Figure 3.6. Original time series (blue) and the same time series after applying different Savitzky-Golay filters (red, purple, yellow)

To determine the mathematical relationship between K_c-NDVI a few steps were necessary. First, the K_c values assumed that the crops were well irrigated. For that reason, for each time series, only the section where the pattern was similar to or the closest to a proper irrigated culture was used by ensuring the same number of days of the theoretical K_c curve for that crop. The period was selected by comparing each consecutive interval of days (with the same number of days of the theoretical K_c curve) with the theoretical K_c curve and the period where the Pearson correlation coefficient (*r*) was the highest since some studies found a high level of correlation between the two (Alam et al., 2018; Kukal et al., 2017). The Person correlation coefficient measures the linear correlation between two variables, where the values are always between -1 and +1, corresponding to a perfect negative and positive correlation, respectively.

The overall look of the time series after all pre-processing (use of linear interpolation to remove NaN values, using S-G filter to filter the noise on the time series and selection of the most representative period for each time series using Pearson correlation) can be seen in Figure 3.7.

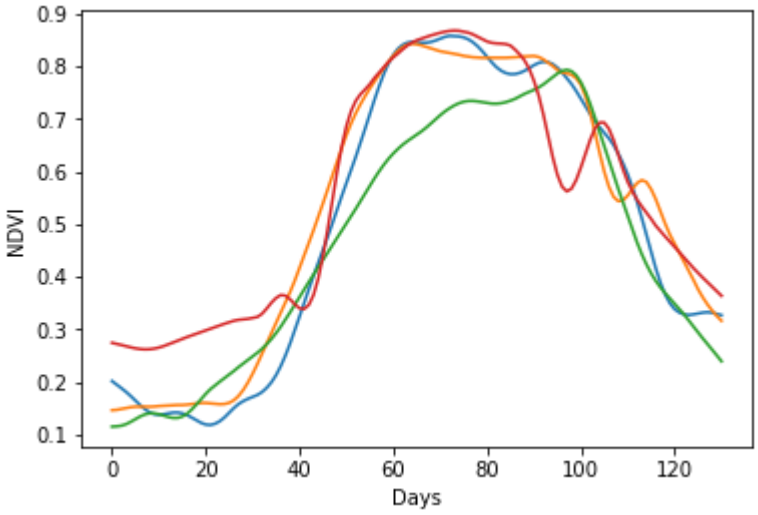

Figure 3.7. Some of the time series after the pre-processing

Afterward, the data was divided into two subsets: train (70%) and test (30%) for maize and tomato and train (60%) and test (40%) for potato and sunflower since there are fewer time series for these two crops compared with maize and tomato. The train set was used to fit the data and the test set to evaluate the model. The following approaches for the analysis of the time series for each culture are mentioned below:

-Use the mean of the time series values given at each point to create a single time series.

-Use the k-means clustering algorithm to create clusters where the values of the centroids of each cluster would be used as time series. The purpose of using k-means was to see if new temporal patterns in the data were found.

-Two other procedures used the same two previously mentioned, but instead of using all the time series, only 10% were selected (before the division in train and test set). The time series were chosen by having the highest values of NDVI at the global maximum of all the time series for that crop. In Figure 3.8 can be seen the global maximum of a time series. The intention behind these approaches is that the theoretical K_c curve was established for cultures that had achieved their full development. However, it´s difficult for the cultures in practice to achieve this stage of growth due to diverse reasons (e.g. climate change), and for that reason using only the crops that are the closest to this stage for modeling should give, in theory, better results. The drawback is that since we are working with fewer series there is a higher chance of having a lower proportion in terms of real signal/noise, and so, the new series created can contain more noise comparably with the other methods.

-The last two approaches followed the same methodology as the two first initially mentioned, but beforehand (before the division in train and test set) a pre-selection of the time series was done where only the ones with a Pearson correlation higher or equal to 0,9 with the K_c time series were used. The motivation behind this was that some time series appeared to have a different pattern from the majority, meaning that the presence of different crops was due to some time series being misattributed. For that reason, limiting the amount of series to the ones where only a strong linear relationship was present should allow the removal of some noise in the data. The drawbacks are that there is a chance of removing time series from the same culture and preserving time series that coincidentally have a high correlation with the K_c curve.

These six approaches were applied and compared since a massive number of series makes the one-by-one analysis difficult and time-consuming. So, by using only the centroids values of each cluster or the mean of the time series, these difficulties would be solved and enable us to understand the main patterns that these time series have and see the overall quality of the data.

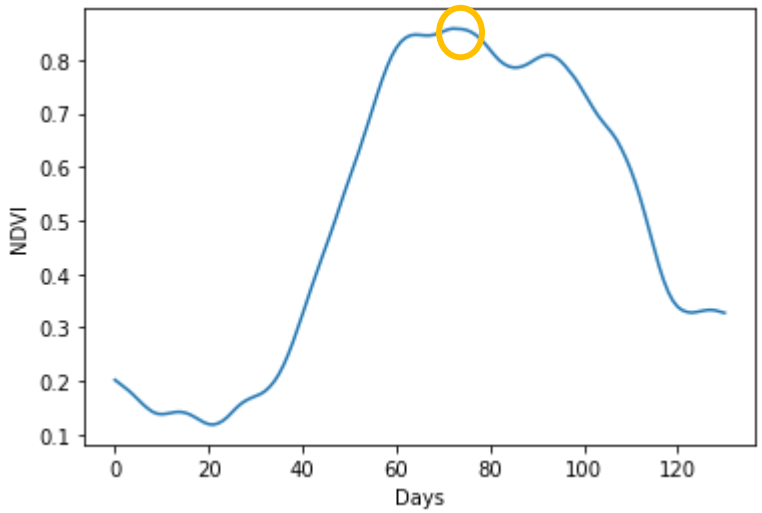

Figure 3.8. NDVI time series of a crop. In orange the global maximum of this series is shown.

The k-means algorithm was applied using the tslearn python library, which is based on the scikit-learn python library implementation of the algorithm. According to the documentation of scikit-learn:

"the k-means algorithm clusters data by trying to separate samples in n groups of equal variance: minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. The k-means algorithm divides a set of N samples X into K disjoint clusters C , each described by the mean μ_j of the samples in the cluster. The k-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left| \left| x_i - \mu_j \right| \right|^2 \right) \qquad (3.2)$$

Inertia can be recognized as a measure of how internally coherent clusters are." (2.3. Clustering, n.d.). The algorithm was implemented by considering the following steps:

1. Choose a value for k (number of clusters)
2. Initialize the k clusters centers (centroids) using k-means++ (Arthur & Vassilvitskii, 2007)
3. Use Euclidian distance to assign the remaining instances to the nearest cluster center
4. Re-estimate the k cluster centers
5. If convergence it´s reached by the relative tolerance with regards to Frobenius norm of the difference in the cluster centers of two consecutive iterations or the maximum number of iterations of the k-means algorithm for a single run it´s reached, then the algorithm stops. Otherwise, use the new k centroids and repeat steps 3-5.

The Euclidian distance was used to reduce the computation time since a large set of time series was considered, and the Euclidian distance is fast to compute (Ratanamahatana & Keogh, 2004). Every culture was clustered with a maximum of 1000 iterations of the algorithm occurring on each run to increase the probability of convergence being reached and k-means++ was used to speed up the convergence (Arthur & Vassilvitskii, 2007). The clustering process was the following for each culture:

1. Create 2 clusters from the time series
2. Calculate the mean silhouette score of all samples
3. Repeat the step (1) and (2) for 3 clusters
4. The n clusters that would give a higher silhouette score would be chosen for the next step of the analysis

The silhouette score is calculated for each sample as (b-a)/max(a,b), where a is the mean intra-cluster distance and b is the mean nearest cluster distance for each data point. The score can range between -1 and 1, where a data point with a score close to -1 has a high probability of being in the wrong cluster and a score close to 1 have a high chance of being in the right cluster (Shahapure & Nicholas, 2020). A maximum of 3 clusters was established since when more than 3 clusters were tested, new patterns in the time series weren´t visually detected.

An example of the maize crop patterns for the k-means algorithm and mean is illustrated in Figure 3.9. In the case of the time series created from the k-means algorithm, for model evaluation only the time series with centroids with the highest values of NDVI at the global maximum with a difference not higher than 0,1 when compared to the global maximum of the mean time series were used. At the same time, these time series need to preserve the overall shape when compared with the mean time series, since it is likely to be the best representative of a well-developed crop while reducing the possibility of the time series, due to noise, being the representative of another crop and so for the example illustrated in Figure 3.4 only the second cluster (cluster_2) was used.

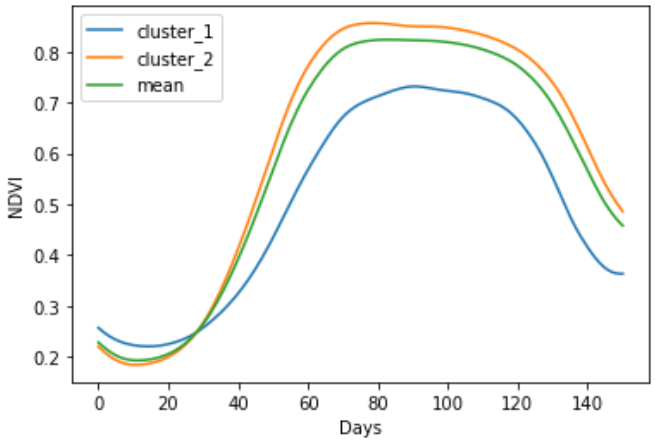

Figure 3.9. Mean (green) and clusters (orange and blue) created by the k-means algorithm for the maize culture after the smoothing was applied.

3.3. Modeling and Evaluation

The determination of the relationship between the two (K_c and NDVI) was done using simple linear and polynomial regression.

The following expression defines the simple regression model:

$$Y = \beta_0 + \beta_1 X, \qquad (3.3)$$

where Y is the dependent variable that we want to predict (target), X is the independent variable (predictor), and β_0 and β_1 are the intercept and slope, respectively. They are unknown parameters or coefficients and can be estimated. The training data can be used to generate estimates β_0 and β_1 for the model coefficients by the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \qquad (3.4)$$

where $\hat{y}$ is a prediction of Y based on $X = x$, and $\hat{\beta}_0$, $\hat{\beta}_1$ are the estimated parameters. The employed algorithm will determine the line that minimizes the distance to each data pair according to a criterion. The method used in this work was the ordinary least squares estimation (OLS). The algorithm consists of calculating the residual sum of squares (RSS) as:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad (3.5)$$

where e_i is the *i*th residuals that is, the difference between the *i*th observed value ($y_i = \beta_0 + \beta_1 x_i + \epsilon_i$) and the *i*th predicted value by the linear model ($\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) (Hastie et al., 2021).

The polynomial regression is an extension of the linear regression and can be represented as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i, \qquad (3.6)$$

where the coefficients can be estimated as well by the OLS method. The polynomial regression has the advantage of being able to capture non-linear relationships of the data (Hastie et al., 2021).

For the polynomial regression, we tested polynomial degrees up to 5 since no significant improvements were observed when further increasing the higher polynomial degree; and solely cases achieving the best results are shown. The models were evaluated using the coefficient of determination (R^2), and root mean square error (RMSE) and a visual inspection of test set predictions.

R^2 is a statistic used to assess the goodness-of-fit (degree of fit) ranging from 0 (no fit) to 1 (perfect fit). In other words, R^2 explains the proportion of variability in K_c that may be explained by the independent variable NDVI and is given by the following equation.

$$R^2 = \left[\frac{\sum_{i=1}^{n} (K_c - \overline{K_c})(NDVI - \overline{NDVI})}{\sum_{i=1}^{n} \sqrt{(K_c - \overline{K_c})^2 (NDVI - \overline{NDVI})^2}} \right]^2 \quad (3.7)$$

On the other hand, RMSE measures the variation of the predicted values around observed values where the lower its value, the better the fit. As opposed to R^2, RMSE is an absolute measure of fit. RMSE is the square root of the average of all the squared residuals (the difference between the observed and predicted values) and can be measured by the following expression:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}} \qquad (3.8)$$

where yi is the ith observation of y and ŷ is the predicted y value given the model.

3.4. Summary

In summary, the methodology starts by acquiring the data from the Aquafarm database, pre-processing it to become useful in the following steps, using or not a pre-selection of the time seriesto compare in the evaluation step, dividing the data in train and test set, using mean and k-means to create a single time series in each case, fitting the time series to linear and polynomial regression and finally evaluate these models. In the Figure 3.10 a representation of every step of the methodology is illustrated.

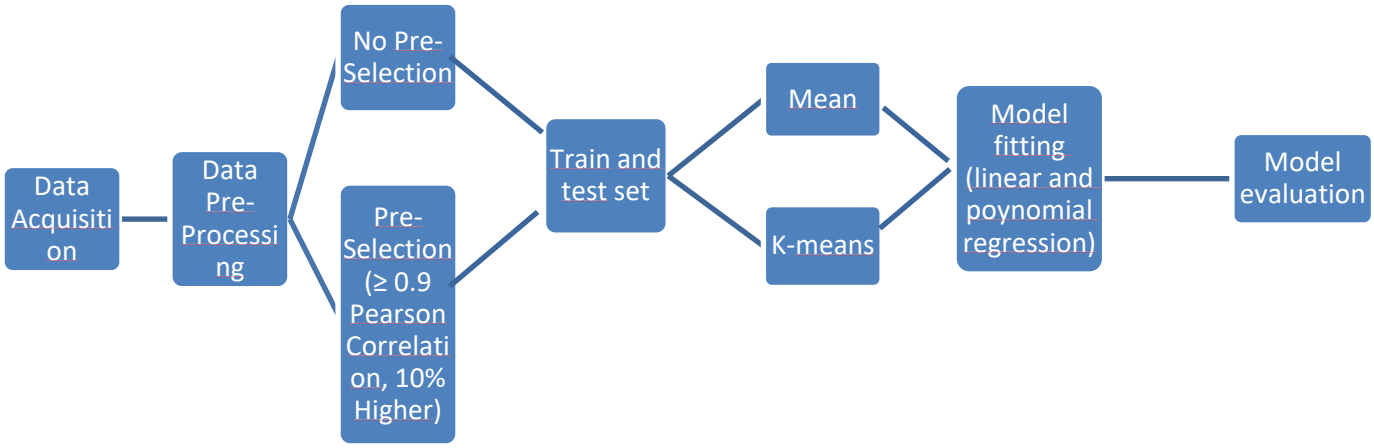

Figure 3.10. Representation of every step of the methodology from data acquisition to model evaluation.

CHAPTER 4

Results and Discussion

This chapter shows the models for each crop in each section (4.1 to 4.4). A comparison is made between linear and polynomial regression; and between using the mean and k-means, followed by a comparison with previous studies that used linear and non-linear relationships to estimate K_c. In section 4.5 a summary of the results is presented.

4.1. Maize

Table 4.1 and Table 4.2 present the results obtained for using mean and k-means, respectively, without pre-selection, only selecting the ones with at least 0,9 Pearson correlation and finally, the time series with the 10% highest values at the global maximum of all the time series for that crop are presented.

Table 4.1. Results obtained for Maize using mean

MAIZE						
mean						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.981	0.98	0.978	0.995	0.995	0.995
R squared (test)	0.979	0.979	0.976	0.993	0.994	0.991
RMSE (train)	0.049	0.050	0.049	0.026	0.026	0.026
RMSE (test)	0.051	0.052	0.050	0.028	0.028	0.035

Table 4.2. Results obtained for Maize using k-means

MAIZE						
k-means						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.971	0.965	0.955	0.993	0.993	0.991
R squared (test)	0.970	0.964	0.954	0.991	0.991	0.985
RMSE (train)	0.060	0.067	0.076	0.030	0.031	0.033
RMSE (test)	0.062	0.068	0.078	0.033	0.034	0.044

From these results, we can observe that the difference between the train and test set metrics values is minimal in all cases in both tables (indicating no overfitting). Another aspect in common in both tables is the fact that the polynomial regression achieves better results than linear regression.

From Table 4.1, we can see that using pre-selection or not provides pretty much the same results, both for linear and polynomial regression. However, in Table 4.2, a higher discrepancy is detected for the linear regression, compared with Table 4.1.

Since polynomial regression provided better results a close inspection was done in the graphical representation of the predictions done in the test set illustrated in Figure 4.1.

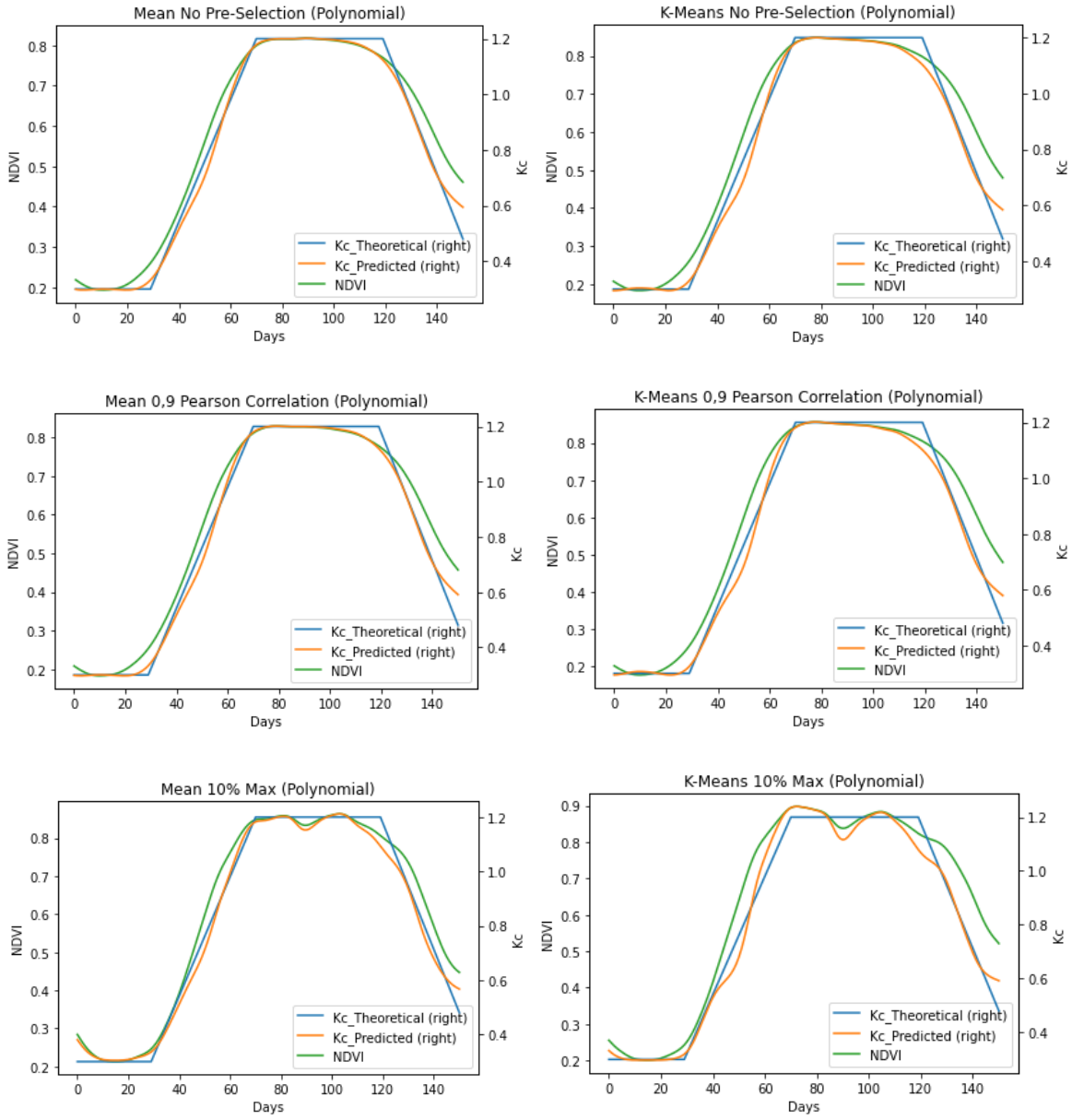

Figure 4.1. Test set predictions for different models for Maize

We can see from Figure 4.1 that the mean provides, in general, a better fit compared to k-means. Another pattern that can be seen is that k-means NDVI time series reach higher maximum values than the mean NDVI time series which can, in theory, be a better representation of a well-developed crop. Finally, it is possible to observe that using mean without pre-selection and mean with a pre-selection of the cases with a Pearson correlation higher or equal to 0.9 with the K_c crop coefficient - provides the best fit. In contrast, mean and k-means with a selection of the time series with the 10% highest values at the global maximum of all the time series for that crop provided the worst fit.

4.1.1. Other Studies

A study made in Évora (Portugal) achieved an R^2 of 0,82 through NDVI obtained by the Landsat 5 images and K_c coefficients measured using water balance and in situ observed data (Toureiro et al., 2017). In another study made by Beeri et al. (2019), the best model achieved an R^2 of 0,95 and an RMSE of 0,057. Showing that the models created in this work show better results, it is essential to know that the conditions in their studies are different from those in this work.

4.2. Tomato

In Table 4.3 and Table 4.4 the results obtained for using mean and k-means, respectively, without pre-selection, only selecting the ones with at least a 0,9 Pearson correlation and finally, the time series with the 10% higher values at the global maximum of all the time series for that crop are presented.

Table 4.3. Results obtained for Tomato using mean

TOMATO						
mean						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.984	0.985	0.989	0.998	0.998	0.998
R squared (test)	0.984	0.984	0.982	0.998	0.998	0.996
RMSE (train)	0.027	0.027	0.023	0.010	0.010	0.009
RMSE (test)	0.028	0.027	0.029	0.010	0.010	0.014

Table 4.4. Results obtained for Tomato using k-means

TOMATO						
k-means						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.965	0.991	0.971	0.994	0.998	0.994
R squared (test)	0.960	0.990	0.962	0.990	0.998	0.989
RMSE (train)	0.041	0.022	0.037	0.017	0.009	0.017
RMSE (test)	0.044	0.022	0.042	0.021	0.009	0.023

From the results, it's possible to observe that in all cases in both tables, the differences between the train and test set metrics values are minimal, although the cases where a pre-selection of the time series with the 10% highest values at the global maximum of all the time series for that crop show in general a bigger discrepancy. Another aspect in common in both tables is where polynomial regression is used, better results are detected when compared with the same methodology when using linear regression. In Table 4.3 it's possible to see either using or not a pre-selection of the time series provided similar results for linear and polynomial regression. However, in Table 4.4 it's possible to see that applying a pre-selection by selecting the time series with at least a 0,9 Pearson correlation with K_c provides better results.

Finally, it's possible to see that the methodologies that provided better results were using the mean with polynomial regression, with no pre-selection and with a pre-selection of the time series that have at least a 0,9 Pearson correlation with K_c and a pre-selection of the time series with the 10% highest values at the global maximum of all the time series for that crop as well as k-means with polynomial regression with a pre-selection of the time series who have at least a 0,9 Pearson correlation with K_c. For that reason, a closer inspection was done of the graphical representation of the predictions in the test set illustrated in Figure 4.2.

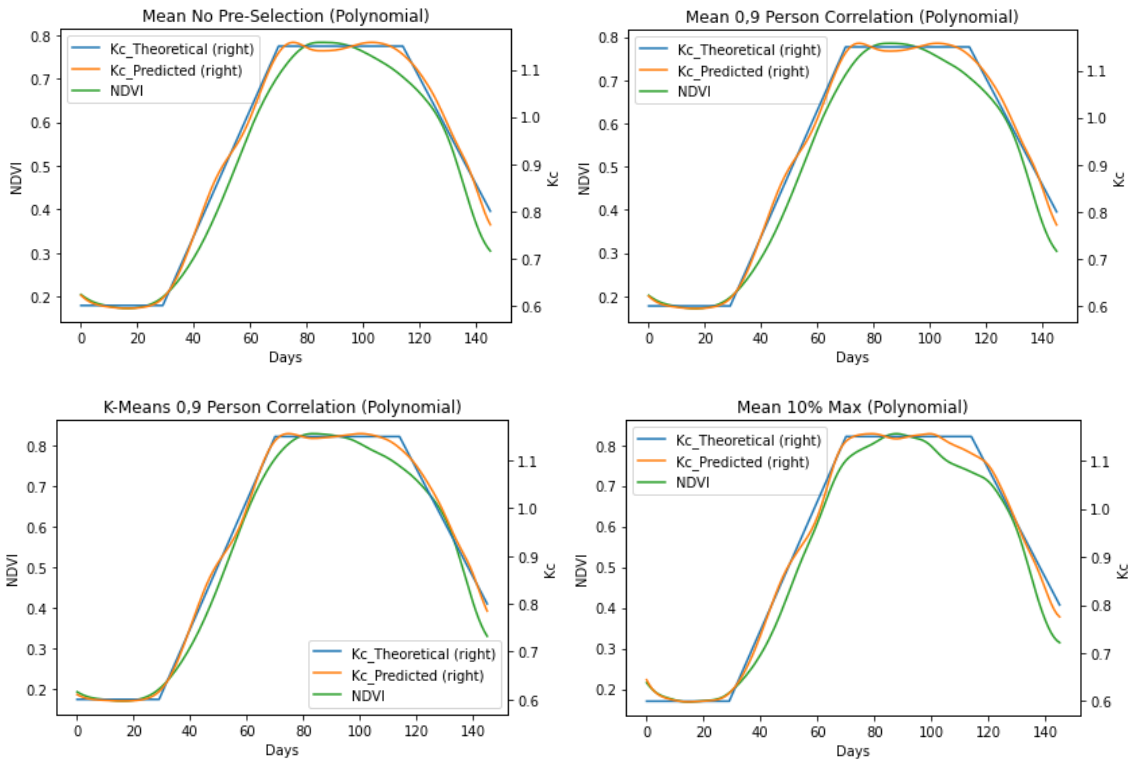

Figure 4.2. Test set predictions for different models for Tomato

From Figure 4.2 it's possible to see that all four provided a great fit to the K_c curve, although using mean with a selection of the time series with the 10% highest values at the global maximum of all the time series for that crop provided a slightly worse fit but almost insignificant.

4.2.1. Other Studies

A study made by Ihuoma et al. (2021) in Canada established two equations using NDVI using Sentinel-2 and PlanetScope, obtaining an R^2 of 0,98 and 0,78, respectively. Although the study was done in a different location the result using Sentinel-2 were similar to those obtained in this work.

4.3. Potato

In Table 4.5 and Table 4.6, the results obtained from using mean and k-means, respectively, without pre-selection, only selecting the ones with at least a 0,9 Pearson correlation and the time series with the 10% higher values at their maximum, are presented.

Table 4.5. Results obtained for Potato using mean

POTATO						
mean						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.984	0.992	0.993	0.998	0.999	0.998
R squared (test)	0.983	0.985	0.986	0.993	0.989	0.995
RMSE (train)	0.033	0.023	0.021	0.012	0.010	0.011
RMSE (test)	0.034	0.030	0.030	0.022	0.028	0.018

Table 4.6. Results obtained for Potato using k-means

POTATO						
k-means						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.995	0.995	0.989	0.998	0.998	0.997
R squared (test)	0.992	0.993	0.971	0.996	0.997	0.985
RMSE (train)	0.019	0.019	0.027	0.011	0.011	0.013
RMSE (test)	0.023	0.022	0.044	0.017	0.013	0.032

Compared to Maize and Tomato, the results show, overall, a bigger discrepancy between train and test set metrics in both tables. Another aspect in common in both tables is where polynomial regression is used, better results are detected when compared with the same methodology when using linear regression. On other hand, in contrast to Maize and Tomato, using k-means generally provided better test results than the mean.

Finally, the methods that offered better results were: mean with polynomial regression without pre-selection and with a selection of the time series with the 10% highest values at the global maximum of all the time series for that crop and using k-means with linear and polynomial regression without pre-selection and with a selection of the time series who have at least a 0,9 Pearson correlation with K_c and for that reason a closer inspection was done of the graphical representation of the predictions in test set illustrated in Figure 4.3.

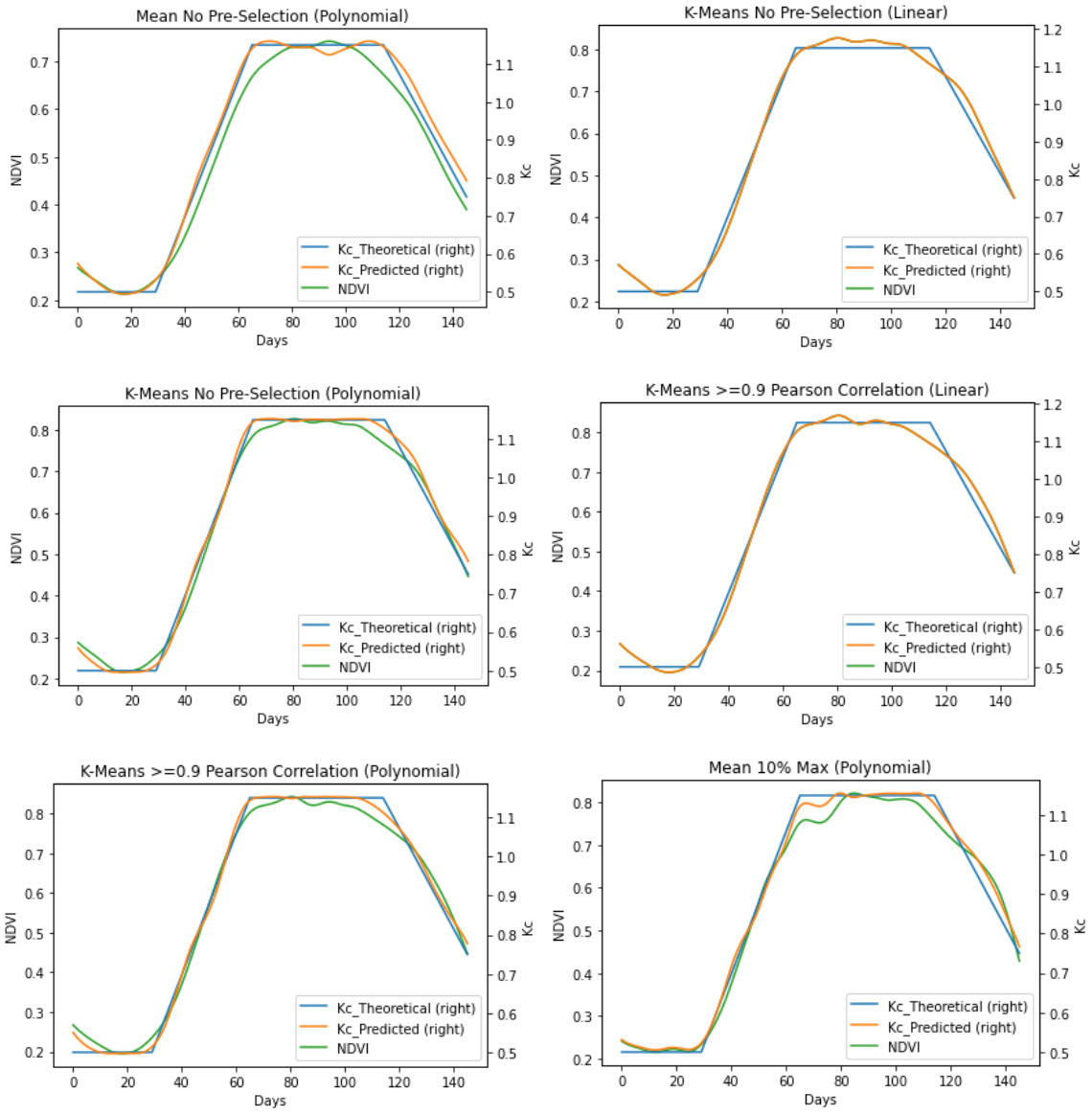

Figure 4.3. Test set predictions for different models for Potato

From Figure 4.3 it's possible to see that using mean without a pre-selection had lower values of NDVI at his global maximum compared to the other cases, which may be less representative of a fully developed crop than others.

Figure 4.3 suggests that using k-means with a selection of the time series that have at least a 0,9 Pearson correlation with K_c and fitted with a polynomial regression, a better fit is obtained, although all cases seemed to have achieved a good fit.

4.3.1. Other Studies

A study by Alataway et al. (2019) established an equation that measures K_c according to the nº of days after planting, obtaining an R^2 of 0,9629. The study by Malachy et al. (2022) used crop height methods using four different methods where the best model obtained an R^2 of 0,84 with an RMSE of 0,049. Although the models created in this work show better results, it is important to know that the conditions in their studies differ from those in this work.

4.4. Sunflower

Table 4.7 and Table 4.8 present the results obtained for using mean and k-means, respectively, without pre-selection, only selecting the ones with at least a 0,9 Pearson correlation, and finally, the time series with the 10% higher values at their global maximum.

Table 4.7. Results obtained for Sunflower using mean

SUNFLOWER						
mean						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.973	0.980	0.969	0.983	0.986	0.982
R squared (test)	0.968	0.978	0.934	0.972	0.982	0.954
RMSE (train)	0.049	0.042	0.053	0.039	0.035	0.039
RMSE (test)	0.056	0.045	0.076	0.050	0.040	0.064

Table 4.8. Results obtained for Sunflower using k-means

SUNFLOWER						
k-means						
	Linear			Polynomial		
	No Pre-selection	≥0,9 P. Corr.	10% Higher	No Pre-selection	≥0,9 P. Corr.	10% Higher
R squared (train)	0.959	0.969	0.978	0.980	0.984	0.989
R squared (test)	0.948	0.968	0.938	0.970	0.981	0.960
RMSE (train)	0.060	0.053	0.044	0.042	0.038	0.032
RMSE (test)	0.068	0.053	0.074	0.051	0.041	0.059

Comparing these results with Maize and Tomato shows a higher discrepancy between train and test set metrics in both tables. Another aspect in common in both tables is where polynomial regression is used, better results are detected when compared with the same methodology when

using linear regression. On the other hand, in contrast to Potato, in general, using mean provided better test results compared to k-means.

Finally, the methods that offered better results were using mean and k-means with no pre-selection using polynomial regression and using mean and k-means with a pre-selection of the time series who have at least a 0,9 Pearson correlation with K_c with linear and polynomial regression and for that reason a closer inspection was done in the graphical representation of the predictions done in test set illustrated in Figure 4.4.

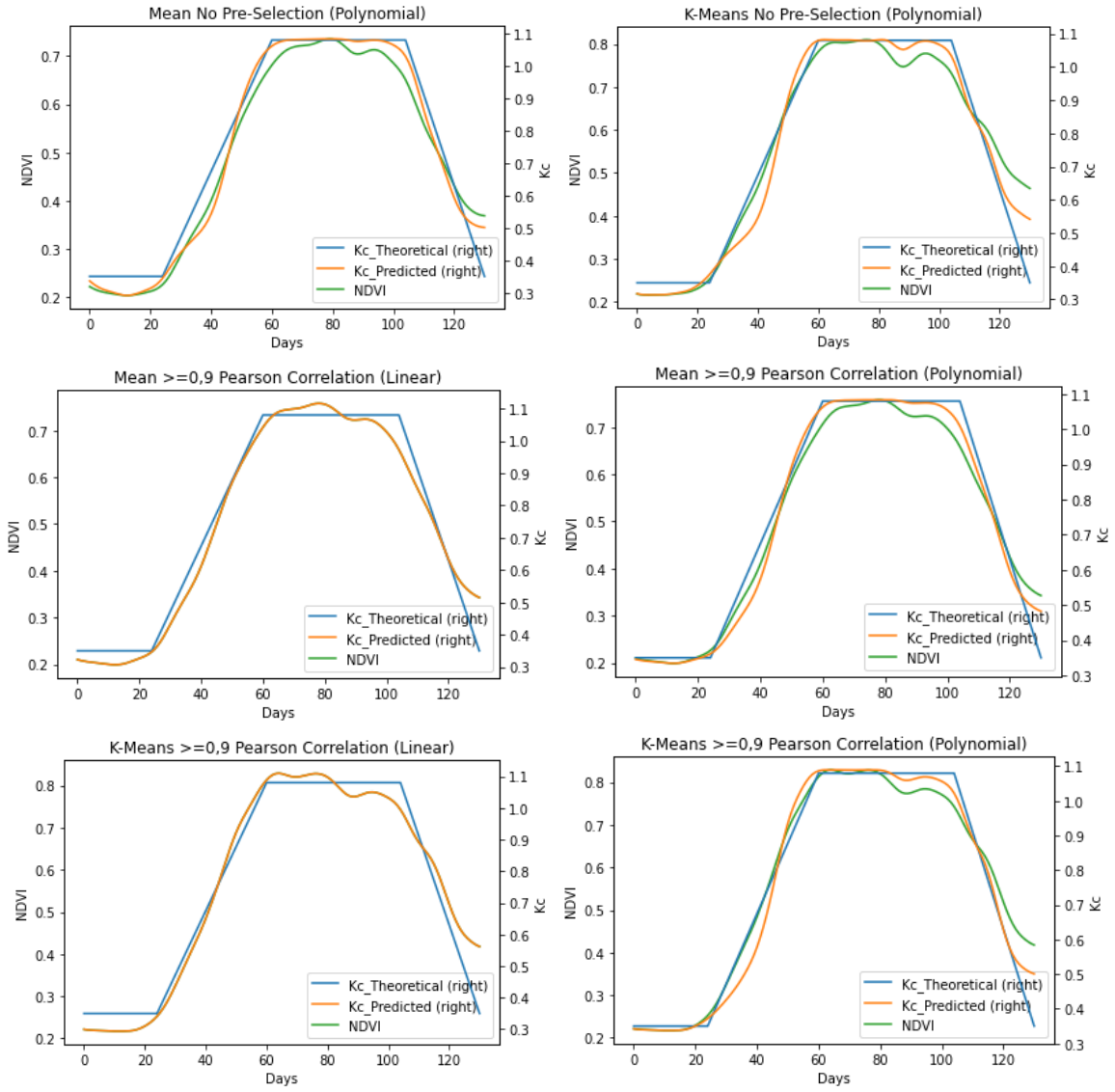

Figure 4.4. Test set predictions for different models for Sunflower

From Figure 4.4 it's possible to see that in most cases, the models are a good fit. In all models, we can see that there is a struggle to capture the end of the K_c curve (approx. after 120 days). It may be explained by the fact that the theoretical K_c curve is not the best representative for the sunflower development in Portugal or due to climate conditions in the country during this study. Another

interesting pattern is that in all models that used k-means, the NDVI time series have a higher value at their global maximum which may mean they are a better representation of a well-developed crop.

Finally, from Figure 4.4, it seems that the model that fits better overall to the K_c curve uses mean with a selection of the time series that have at least a 0,9 Pearson correlation with the K_c curve and fitted with a polynomial regression.

Just like in the potato case, in the cases where linear regression was used the NDVI curve is not possible to see since the predicted K_c curve it´s overlapping it.

4.4.1. Other Studies

A study in China made by Hong et al. (2017) established relationships between K_c in different four different stages of growth (initial, rapid growth, middle and mature stage) with salinity levels for sunflowers under salt stress obtaining an R^2 of 0,860, 0,003, 0,225 and 0,312, respectably. Since the conditions are very different from this study and an adequate study was not found, a comparison will not be made.

4.5. Summary

Taking the results obtained into account we conclude that:

- For maize the use or not of a pre-selection of time series provides no major differences with the exception of using k-means with linear regression where using no pre-selection provides better results. Using mean instead of k-means provides better results when using the same methodology.
- For tomato the use or not of a pre-selection of time series provides no major differences with the exception of using k-means where using a pre-selection of the time series with at least a Pearson correlation with the theoretical K_c curve provides better results. Using mean instead of k-means provides, in most cases, better results when using the same methodology.
- For potato the use or not of a pre-selection of time series provides no major differences with the exception of using k-means where using a pre-selection of the time series with the 10% highest values at the global maximum of all the time series for that crop provides worst results. Using k-means instead of mean provides, in most cases, better results when using the same methodology.
- For sunflower the use or not of a pre-selection of time series with at least a Pearson correlation with the theoretical K_c curve provides better results

compared to the other methods. Using mean instead of k-means provides, in most cases, better results when using the same methodology.

CHAPTER 5

Conclusion

This chapter summarizes the dissertation conclusions and presents some proposals for future work.

5.1. Conclusions

This work has studied different pre-selection techniques using means and k-means to generate time series. A literature review has been presented in Chapter 2, which addresses the fundamental concepts related to this work: remote sensing, evapotranspiration, crop coefficient, and different methodologies to determine evapotranspiration and vegetation indexes.

In Chapter 3, the methodology used in this work is described. It started by explaining how data was acquired and what contained. After, a description of how each time series period was selected using Pearson Correlation, followed by an explanation of all the pre-selection approaches used before using the mean and k-means to generate the time series. Finally, an explanation of the algorithms (linear and polynomial regression) used to fit the generated time series and the metrics used to evaluate the models are provided.

In Chapter 4, the models created are evaluated and compared for each one of the crops. All models seem to be able to capture the K_c curve relatively well. However, the models created for the sunflower struggle a little more by the end of the K_c curve which may be due to the number of the time series available being less than for the other crops. Another critical aspect to take into account is the fact that the use of pre-selection of the time series didn't provide a significant difference compared with the absence of the pre-selection although some of the best models are the ones where a pre-selection occurred. A pattern in common in all crops was that the polynomial regression always provided better results than the linear regression when the same methodology was followed.
As a final remark, the results obtained in this work confirm that using a pre-selection of the time series, mean, and k-means for these crops helps to capture the crop coefficient curve. Choosing the best methodologies depends on each crop although there isn't one that is overall better than the others. Finally, the methodologies presented show promising results that can be seen as potential methods to determine crop coefficients better, and the best models for each crop are adequate for their use, at least in the region of this study. The code to reproduce the results it´s available at https://github.com/GuilhermeDuarte-30/tesemestrado.

5.2. Future work

Regarding the results obtained in this dissertation, some future work suggestions can be considered:

- Test the same methodologies in different crops and see if the quality of the results still holds.
- Combine different methodologies, for example, first use Pearson correlation to choose the time series, and finally, use only the time series with the highest values at the global maximum of all the time series for that crop and assess the performance.
- Use different vegetation indexes and compare them to the ones used in this study.
- Study the impact of smoothing the data on max values of NDVI

APPENDIX A

Models Obtained

Table 6.1. Models obtained for Maize using mean

MAIZE						
Mean						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=1.496x-0.05	y=1.449x-0.030	y=1.377x-0.018	$y=-104.2\ x^5+263.6\ x^4-252.5\ x^3+114.6\ x^2-23.35\ x+2.016$	$y=-89.06\ x^5+225.9\ x^4-216.2\ x^3+97.64\ x^2-19.57\ x+1.702$	$y=-45.51\ x^5+120.4\ x^4-118.3\ x^3+53.89\ x^2-10.29\ x+0.9664$

Table 6.2. Models obtained for Maize using k-means

MAIZE						
K-Means						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=1.399x-0.03	y=1.374x-0.019	y=1.328x-0.018	$y=-87.28\ x^5+228.9\ x^4-225.6\ x^3+104.3\ x^2-21.5\ x+1.888$	$y=-82.82\ x^5+218\ x^4-215.2\ x^3+99.38\ x^2-20.35\ x+1.789$	$y=-80.59\ x^5+218.1\ x^4-219.4\ x^3+102.1\ x^2-20.88\ x+1.823$

Table 6.3. Models obtained for Tomato using mean

TOMATO						
Mean						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=0.939x+0.453	y=0.930x+0.457	y=0.857x+0.472	$y=-112.7\ x^5+261.3\ x^4-227.9\ x^3+91.75\ x^2-15.7\ x+1.527$	$y=-108.5\ x^5+251.8\ x^4-219.5\ x^3+88.23\ x^2-15.01\ x+1.479$	$y=-68.36\ x^5+163.9\ x^4-146.4\ x^3+59.56\ x^2-9.863\ x+1.147$

Table 6.4. Models obtained for Tomato using k-means

TOMATO						
K-Means						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=0.865x+0.463	y=0.859x+0.464	y=0862x+0496	y=-213.2 x^5+ 451.4 x^4- 361.7 x^3 + 134.5 x^2 - 21.44 x + 1.771	y=-77.16 x^5+ 188.2 x^4- 171.8 x^3 + 71.95 x^2- 12.64 x + 1.364	y=-74.26 x^5+ 167.7 x^4- 138.8 x^3 + 50.48 x^2- 6.645 x + 0.8292

Table 6.5. Models obtained for Potato using mean

POTATO						
Mean						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=1.316x+0.237	y=1.188x+0.273	y=1.091x+0.275	y=-202.6 x^5+ 463.7 x^4- 408.2 x^3 + 171.1 x^2- 32.37 x + 2.704	y=-145.5 x^5+ 340 x^4- 302.5 x^3 + 126.8 x^2- 23.56 x + 2.058	y=-115.7 x^5+ 284 x^4- 263.1 x^3 + 113.9 x^2- 21.79 x + 1.985

Table 6.6. Models obtained for Potato using k-means

POTATO						
K-Means						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=1.121x+0.249	y=1.055x+0.279	y=1.017x+0.279	y=-124.1 x^5+ 320.5 x^4- 315.9 x^3 + 147.3 x^2- 31.1 x + 2.885	y=-99.71 x^5+ 256.3 x^4- 249.4 x^3 + 113.6 x^2- 22.98 x + 2.164	y=-94.86 x^5 + 248.5 x^4- 244.8 x^3 + 112.1 x^2- 22.73 x + 2.15

Table 6.7. Models obtained for Sunflower using mean

SUNFLOWER						
Mean						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=1.556x-0.027	y=1.448x+0.019	y=1.388x+0.076	$y=108\ x^5-294.3\ x^4+298.8\ x^3-140.1\ x^2+31.69\ x-2.403$	$y=2.522\ x^5-22.16\ x^4+30.67\ x^3-14.82\ x^2+3.958\ x-0.07412$	$y=83.62\ x^5-222.2\ x^4+217.2\ x^3-96.22\ x^2+20.36\ x-1.251$

Table 6.8. Models obtained for Sunflower using k-means

SUNFLOWER						
K-Means						
	Linear			Polynomial		
	No Pre-Selection	≥0,9 P. Corr.	10% higher	No Pre-Selection	≥0,9 P. Corr.	10% higher
Equation	y=1.377x-0.015	y=1.333x+0.005	y=1.429x+0.111	$y=6.924\ x^5-53.77\ x^4+88.26\ x^3-56.12\ x^2+16.1\ x-1.323$	$y=0.8838\ x^5-21.02\ x^4+37.86\ x^3-23.57\ x^2+6.859\ x-0.3819$	$y=-113.4\ x^5+254.9\ x^4-223.3\ x^3+94.07\ x^2-17.16\ x+1.453$

References

AQUASTAT - FAO's Global Information System on Water and Agriculture. (n.d.). Retrieved November 30, 2021, from https://www.fao.org/aquastat/en/databases/

Aquafarm database. (n.d.). Aquafarm. Retrieved November 22, 2021, from https://aquafarm.hidromod.com/

Homepage | Copernicus. (n.d.). Retrieved November 30, 2021, from https://www.copernicus.eu/en

Savitzky, Abraham., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. In Analytical Chemistry (Vol. 36, Issue 8, pp. 1627–1639). American Chemical Society (ACS). https://doi.org/10.1021/ac60214a047

2.3. Clustering. (n.d.). Scikit-learn. Retrieved March 24, 2022, from https://scikit-learn.org/stable/modules/clustering.html

Ahmed, A. A. M., Deo, R. C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., & Yang, L. (2021). Hybrid deep learning method for a week-ahead evapotranspiration forecasting. *Stochastic Environmental Research and Risk Assessment*. https://doi.org/10.1007/s00477-021-02078-x

Alam, M. S., Lamb, D. W., & Rahman, M. M. (2018). A refined method for rapidly determining the relationship between canopy NDVI and the pasture evapotranspiration coefficient. *Computers and Electronics in Agriculture*, *147*(January), 12–17. https://doi.org/10.1016/j.compag.2018.02.008

Alataway, A., Al-Ghobari, H., Mohammad, F., & Dewidar, A. (2019). Lysimeter-based water use and crop coefficient of drip-irrigated potato in an arid environment. *Agronomy*, *9*(11), 1–11. https://doi.org/10.3390/agronomy9110756

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). *FAO Irrigation and Drainage Paper No. 56 - Crop Evapotranspiration*. *January 1998*.

Allen, Richard G., Pereira, L. S., Howell, T. A., & Jensen, M. E. (2011a). Evapotranspiration information reporting: I. Factors governing measurement accuracy. *Agricultural Water Management*, *98*(6), 899–920. https://doi.org/10.1016/j.agwat.2010.12.015

Allen, Richard G., Pereira, L. S., Howell, T. A., & Jensen, M. E. (2011b). Evapotranspiration information reporting: II. Recommended documentation. *Agricultural Water Management*, *98*(6), 921–929. https://doi.org/10.1016/j.agwat.2010.12.016

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, *07-09-Janu*, 1027–1035.

Bausch, W. C., & Neale, C. M. U. (1987). Crop Coefficients Derived From Reflected Canopy Radiation: a Concept. *Transactions of the American Society of Agricultural Engineers*, *30*(3), 703–709. https://doi.org/10.13031/2013.30463

Beeri, O., Pelta, R., Shilo, T., Mey-Tal, S., & Tanny, J. (2019). Accuracy of crop coefficient estimation methods based on satellite imagery. *Precision Agriculture 2019 - Papers Presented at the 12th European Conference on Precision Agriculture, ECPA 2019*, 437–444. https://doi.org/10.3920/978-90-8686-888-9_54

Bhattarai, N., Shaw, S. B., Quackenbush, L. J., Im, J., & Niraula, R. (2016). Evaluating five remote sensing based single-source surface energy balance models for estimating daily evapotranspiration in a humid subtropical climate. *International Journal of Applied Earth Observation and Geoinformation*, *49*, 75–86. https://doi.org/10.1016/j.jag.2016.01.010

Calera, A., Campos, I., Osann, A., D'Urso, G., & Menenti, M. (2017). Remote sensing for crop water management: From ET modelling to services for the end users. *Sensors (Switzerland)*, *17*(5). https://doi.org/10.3390/s17051104

Chen, J., Jönsson, P., Tamura, M., Gu, Z., Matsushita, B., & Eklundh, L. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. *Remote Sensing of Environment*, *91*(3–4), 332–344. https://doi.org/10.1016/j.rse.2004.03.014

Cheng, M., Jiao, X., Jin, X., Li, B., Liu, K., & Shi, L. (2021). Satellite time series data reveal interannual

and seasonal spatiotemporal evapotranspiration patterns in China in response to effect factors. *Agricultural Water Management*, *255*(July), 107046. https://doi.org/10.1016/j.agwat.2021.107046

Choudhury, B. J., Ahmed, N. U., Idso, S. B., Reginato, R. J., & Daughtry, C. S. T. (1994). Relations between evaporation coefficients and vegetation indices studied by model simulations. *Remote Sensing of Environment*, *50*(1), 1–17. https://doi.org/10.1016/0034-4257(94)90090-6

Cracknell, A. P. (2018). The development of remote sensing in the last 40 years. *International Journal of Remote Sensing*, *39*(23), 8387–8427. https://doi.org/10.1080/01431161.2018.1550919

Drexler, J. Z., Snyder, R. L., Spano, D., & Paw U, K. T. (2004). A review of models and micrometeorological methods used to estimate wetland evapotranspiration. *Hydrological Processes*, *18*(11), 2071–2101. https://doi.org/10.1002/hyp.1462

Glenn, E. P. E. P. E. P., Neale, C. M. U. C. M. U., Hunsaker, D. J. D. J., & Nagler, P. L. P. L. P. L. (2011). Vegetation index-based crop coefficients to estimate evapotranspiration by remote sensing in agricultural and natural ecosystems. *Hydrological Processes*, *25*(26), 4050–4062. https://doi.org/10.1002/hyp.8392

Goward, S. N., Markham, B., Dye, D. G., Dulaney, W., & Yang, J. (1991). Normalized difference vegetation index measurements from the advanced very high resolution radiometer. *Remote Sensing of Environment*, *35*(2–3), 257–277. https://doi.org/10.1016/0034-4257(91)90017-Z

Han, H., Bai, J., Ma, G., & Yan, J. (2020). Vegetation phenological changes in multiple landforms and responses to climate change. *ISPRS International Journal of Geo-Information*, *9*(2). https://doi.org/10.3390/ijgi9020111

Hastie, T., Tibshirani, R., James, G., & Witten, D. (2021). An introduction to statistical learning (2nd ed.). *Springer Texts*, *102*, 618.

Hong, M., Zeng, W., Ma, T., Lei, G., Zha, Y., Fang, Y., Wu, J., & Huang, J. (2017). Determination of growth stage-specific crop coefficients (Kc) of sunflowers (Helianthus annuus L.) under salt stress. *Water (Switzerland)*, *9*(3). https://doi.org/10.3390/w9030215

Hunsaker, D. J., Pinter, P. J., Barnes, E. M., & Kimball, B. A. (2003). Estimating cotton evapotranspiration crop coefficients with a multispectral vegetation index. *Irrigation Science*, *22*(2), 95–104. https://doi.org/10.1007/s00271-003-0074-6

Ihuoma, S. O., Madramootoo, C. A., & Kalacska, M. (2021). Integration of satellite imagery and in situ soil moisture data for estimating irrigation water requirements. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102396. https://doi.org/10.1016/j.jag.2021.102396

Johnson, L. F., & Trout, T. J. (2012). Satellite NDVI assisted monitoring of vegetable crop evapotranspiration in california's san Joaquin Valley. *Remote Sensing*, *4*(2), 439–455. https://doi.org/10.3390/rs4020439

Jurečka, F., Fischer, M., Hlavinka, P., Balek, J., Semerádová, D., Bláhová, M., Anderson, M. C. M. C. M. C., Hain, C., Žalud, Z., & Trnka, M. (2021). Potential of water balance and remote sensing-based evapotranspiration models to predict yields of spring barley and winter wheat in the Czech Republic. *Agricultural Water Management*, *256*(July). https://doi.org/10.1016/j.agwat.2021.107064

Kazemi, M. H., Majnooni-Heris, A., Kisi, O., & Shiri, J. (2021). Generalized gene expression programming models for estimating reference evapotranspiration through cross-station assessment and exogenous data supply. *Environmental Science and Pollution Research*, *28*(6), 6520–6532. https://doi.org/10.1007/s11356-020-10916-8

Kharrou, M. H., Simonneaux, V., Er-raki, S., Page, M. L., Khabba, S., & Chehbouni, A. (2021). Assessing irrigation water use with remote sensing-based soil water balance at an irrigation scheme level in a semi-arid region of morocco. *Remote Sensing*, *13*(6). https://doi.org/10.3390/rs13061133

Kukal, M., Irmak, S., & Kilic, A. (2017). Long-Term Spatial and Temporal Maize and Soybean Evapotranspiration Trends Derived from Ground-Based and Satellite-Based Datasets over the Great Plains. *Journal of Irrigation and Drainage Engineering*, *143*(9), 1–18. https://doi.org/10.1061/(asce)ir.1943-4774.0001212

Malachy, N., Zadak, I., & Rozenstein, O. (2022). Comparing Methods to Extract Crop Height and Estimate Crop Coefficient from UAV Imagery Using Structure from Motion. *Remote Sensing*, *14*(4). https://doi.org/10.3390/rs14040810

Murray, R. S., Nagler, P. L., Morino, K., & Glenn, E. P. (2009). An empirical algorithm for estimating agricultural and riparian evapotranspiration using MODIS enhanced vegetation index and ground measurements of ET. II. application to the lower Colorado river, U.S. *Remote Sensing*, *1*(4), 1125–1138. https://doi.org/10.3390/rs1041125

Nagler, P. L., Scott, R. L., Westenburg, C., Cleverly, J. R., Glenn, E. P., & Huete, A. R. (2005). Evapotranspiration on western U.S. rivers estimated using the Enhanced Vegetation Index from MODIS and data from eddy covariance and Bowen ratio flux towers. *Remote Sensing of Environment*, *97*(3), 337–351. https://doi.org/10.1016/j.rse.2005.05.011

Piticar, A., Mihăilă, D., Lazurca, L. G., Bistricean, P. I., Puţuntică, A., & Briciu, A. E. (2016). Spatiotemporal distribution of reference evapotranspiration in the Republic of Moldova. *Theoretical and Applied Climatology*, *124*(3–4), 1133–1144. https://doi.org/10.1007/s00704-015-1490-2

Pôças, I., Calera, A., Campos, I., & Cunha, M. (2020). Remote sensing for estimating and mapping single and basal crop coefficientes: A review on spectral vegetation indices approaches. *Agricultural Water Management*, *233*(February), 106081. https://doi.org/10.1016/j.agwat.2020.106081

Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2018). Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. *Soil and Tillage Research*, *181*(February), 63–81. https://doi.org/10.1016/j.still.2018.03.021

Rafn, E. B., Contor, B., & Ames, D. P. (2008). Evaluation of a Method for Estimating Irrigated Crop-Evapotranspiration Coefficients from Remotely Sensed Data in Idaho. *Journal of Irrigation and Drainage Engineering*, *134*(6), 722–729. https://doi.org/10.1061/(asce)0733-9437(2008)134:6(722)

Rana, G., & Katerji, N. (2000). Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: A review. *European Journal of Agronomy*, *13*(2–3), 125–153. https://doi.org/10.1016/S1161-0301(00)00070-8

Ratanamahatana, C. A., & Keogh, E. (2004). Making time-series classification more accurate using learned constraints. *SIAM Proceedings Series*, 11–22. https://doi.org/10.1137/1.9781611972740.2

Saboori, M., Mokhtari, A., Afrasiabian, Y., Daccache, A., Alaghmand, S., & Mousivand, Y. (2021). Automatically selecting hot and cold pixels for satellite actual evapotranspiration estimation under different topographic and climatic conditions. *Agricultural Water Management*, *248*(January), 106763. https://doi.org/10.1016/j.agwat.2021.106763

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Stanhill, G. (2005). Evapotranspiration. *Encyclopedia of Soils in the Environment*, *4*, 502–506. https://doi.org/10.1016/B0-12-348530-4/00359-3

Tikhamarine, Y., Malik, A., Souag-Gamane, D., & Kisi, O. (2020). Artificial intelligence models versus empirical equations for modeling monthly reference evapotranspiration. *Environmental Science and Pollution Research*, *27*(24), 30001–30019. https://doi.org/10.1007/s11356-020-08792-3

Toureiro, C., Serralheiro, R., Shahidian, S., & Sousa, A. (2017). Irrigation management with remote sensing: Evaluating irrigation requirement for maize under Mediterranean climate condition. *Agricultural Water Management*, *184*, 211–220. https://doi.org/10.1016/j.agwat.2016.02.010

Transon, J., d'Andrimont, R., Maugnard, A., & Defourny, P. (2018). Survey of hyperspectral Earth Observation applications from space in the Sentinel-2 context. *Remote Sensing*, *10*(2), 1–32. https://doi.org/10.3390/rs10020157

Wagle, P., Bhattarai, N., Gowda, P. H., & Kakani, V. G. (2017). Performance of five surface energy

balance models for estimating daily evapotranspiration in high biomass sorghum. *ISPRS Journal of Photogrammetry and Remote Sensing*, *128*, 192–203. https://doi.org/10.1016/j.isprsjprs.2017.03.022

Wang, T., Melton, F. S., Pôças, I., Johnson, L. F., Thao, T., Post, K., & Cassel-Sharma, F. (2021). Evaluation of crop coefficient and evapotranspiration data for sugar beets from landsat surface reflectances using micrometeorological measurements and weighing lysimetry. *Agricultural Water Management*, *244*. https://doi.org/10.1016/j.agwat.2020.106533

Zhao, J., Chen, X., Zhang, J., Zhao, H., & Song, Y. (2019). Higher temporal evapotranspiration estimation with improved SEBS model from geostationary meteorological satellite data. *Scientific Reports*, *9*(1), 1–15. https://doi.org/10.1038/s41598-019-50724-w

Zhu, R., Zheng, H., Wang, E., & Zhao, W. (2013). Multi-model ensemble simulation of flood events using Bayesian model averaging. *Proceedings - 20th International Congress on Modelling and Simulation, MODSIM 2013*, *December*, 455–461. https://doi.org/10.36334/modsim.2013.a10.zhu

Zhu, Z., Wulder, M. A., Roy, D. P., Woodcock, C. E., Hansen, M. C., Radeloff, V. C., Healey, S. P., Schaaf, C., Hostert, P., Strobl, P., Pekel, J. F., Lymburner, L., Pahlevan, N., & Scambos, T. A. (2019). Benefits of the free and open Landsat data policy. *Remote Sensing of Environment*, *224*(February), 382–385. https://doi.org/10.1016/j.rse.2019.02.016