Prediction of plant growth based on statistical methods and remote sensing data

Marwa Hachicha^{a,b,*}, Abdelaziz Kallel^{a,b} and Mahdi Louati^c

^aUniversity of Sfax, Digital Research Centre of Sfax (CRNS), Sfax, Tunisia

^bUniversity of Sfax, Laboratory of Signals, Systems,

Artificial Intelligence and Networks (SM@RT), Sfax, Tunisia

^cUniversity of Sfax, National School of Electronics and Telecommunications of Sfax (ENET'Com), Sfax, Tunisia

Abstract. Based on the data provided from the new satellite constellation, the crop phenology can be mapped accurately with a high spatial and temporal resolution. The potential of Sentinel-2 (high spatial and temporal resolution, 10 m and 5 days) is investigated in order to monitor and forecast the crop growth over different olive groves located in Sfax, Tunisia, from the beginning of 2016 to the end of 2020. The normalized difference vegetation index (NDVI) is used as an indicator of vegetation health due to its high correlation with crop growth health and productivity, and it is particularly forecasted to predict the trends. The prediction is done using two different statistical methods, autoregressive (AR) and Markov chain (MC), based on historical data derived from Sentinel-2 data. The prediction is applied for three different areas having various vegetation types and density. Moreover, for an accurate and precise prediction, our study areas are divided into different homogeneous clusters using the well-known Gaussian mixture model. Furthermore, the performances of our approach are assessed by means of the root mean squared error (RMSE) between the predicted results and the actual data. Globally, the obtained results for all the clusters of each area show that the forecast, using both methods, is accurate where the error is less than 5%. Nevertheless, the MC model displays the highest performance where the predicted NDVI curves of the different study areas are the closest to the actual observation. MC (resp. AR) precision is of 97% with $RMSE = 0.025$ (resp. 95% and $RMSE = 0.041$) for all the clusters of Jbeniana and Limaya, and 99% with $RMSE = 0.0073$ (resp. 98% and $RMSE = 0.0127$) for the four clusters of Chaâl. © The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JRS.15.042410](https://doi.org/10.1117/1.JRS.15.042410)]

Keywords: normalized difference vegetation index; Sentinel-2; statistical method; forecasting; Markov chain model.

Paper 210203SS received Apr. 15, 2021; accepted for publication Aug. 17, 2021; published online Sep. 3, 2021.

1 Introduction

The agricultural sector plays an important role in economic development strategy, where it is the pivotal sector to ensure food security. The undeveloped countries are marked by a low profitability of the agricultural sector and an inability to meet the needs of food security. In such situations, sustainable food production is required and can be assured by the development of effective crop management strategies.

Recently, adaptation of precision agriculture (PA) techniques has shown an improvement in profitability and has increased productivity in an effective manner.^{1,2} Specially, the prediction of plant growth is the main challenge for PA, as it leads to better planning of agricultural activities.³ Contrary to the traditional methods of mapping and vegetation tracking based on *in situ* monitoring, which are costly and time consuming, satellite remote sensing techniques are automatic

*Address all correspondence to Marwa Hachicha, hachichamarwa91@gmail.com

and powerful tools that provide an accurate information on plant growth with high spatial and temporal resolution.^{4,5} Indeed, the growth and yield of crops are well correlated with the vegetation indices. The latter are indicators determined from a mixture of different spectral bands of remote sensing data used, to monitor the vegetation properties for crops management,^{6–8} to detect drought,⁹ and to classify the vegetation.¹⁰ Numerous studies have shown particularly that the normalized difference vegetation index (NDVI)¹¹ is well correlated with crop growth and yield.^{12,13}

It is commonly used in agriculture to show the state of vegetation (healthy or unhealthy) and to monitor crop growth and quality.^{14–16}

The combination of NDVI time series observations gives more insight into vegetation monitoring.^{13,17} The NDVI has drawn considerable attention of researchers, and numerous related works have been realized in different aspects. For instance, in Ref. 18, NDVI has been used for land cover classification in China, and in Ref. 19 to identify the crop type and in Ref. 20 to estimate the vegetation seasonal variation.

Recently, several models were developed to predict NDVI time series. This allows us to investigate the state of vegetation cover. For example in Ref. 21, authors developed a vegetation greenness forecast model to predict the NDVI using Advanced Very High Resolution Radiometer data and climate data such as temperature and precipitation. Others used the combination between the cellular automata model's and the Markov model to predict and simulate NDVI distribution using MODIS data.²² In addition,²³ used time-delay neural network to predict NDVI for the arid and semi-arid grassland.

The main contribution of our article is to develop a new forecasting technique of NDVI time series based on autoregressive (AR) and Markov chain (MC) models. These models are applied over Sentinel-2 time series in order to take advantage of the high spatial and temporal resolutions.

The remainder of the paper is organized as follows. Section 2 presents the study area and the used dataset. In Sec. 3, the proposed prediction methods as well as the criteria for evaluating the performance are shown. In Sec. 4, we present the experimental results and we compare the performance and quality of the different statistical methods. This paper is closed by a conclusion and some perspectives.

2 Study Area and Remote Sensing Data

2.1 Study Area

Sfax is a city in southern Tunisia, which is located in North Africa. The cultivation of this city is characterized by various types of crops, such as vegetables, cereals, almond trees, and pistachio trees, but the mean part of the agriculture land is dedicated to olive grove. It is the first producer of olive oil in Tunisia where the percentage of production is average 40%.

Sfax is characterized by a semi-arid climate, with a hot and dry summer, and wet and cold winter. Additionally, the average annual temperature is 19°C. It is around 31°C in August and 8°C in December and the average annual rainfall is 212 mm.

Our study areas are “Jebeniana,” “Limaya,” and “Chaâl” located in Sfax (see Fig. 1). They are characterized by various types of vegetation. (A) Jebeniana is an agricultural area of a size 4×4 km² centered at 35°2'5.95"N and 10°54'26.10"E. It is a mixture of olive field, agriculture field, and grass. (B) Limaya is of a size 2.50×1.65 km², located around 35°0'23.95"N and 10°8'40.19"E. This area is an olive grove situated around a river and composed of two common variety of olive trees, namely “Arbequina” and “Arbanozona.” They are irrigated olive trees planted in the rows, which are closer within than between them. The spaces are 4 m between two rows and 2 m between two trees. (C) The last area is named Chaâl. Its size is 7.50×8.65 km². Its latitude and longitude are around 34°33'33.09"N and 10°20'11.32"E, respectively. It is an olive grove dominated by “Chlemléle” variety and characterized by a rain mode. Olive trees are placed 24 m apart.

Fig. 1 Map of different study areas located in Sfax, Tunisia. On the left is Limaya, in the bottom right is Chaâl and on the top right is Jbeniana.

2.2 NDVI Data

In this paper, NDVI²⁴ is among the most used spectral index to monitor the vegetation quality and greenness due to its sensitivity to photosynthetic activity.^{25–27} NDVI is derived in our study from Sentinel-2 images. Sentinel-2 mission is developed by the European Space Agency. It is composed of twin satellites, which have relatively wide field of view well adapted for vegetation cover monitoring. Their product is characterized by a high spatial and temporal resolutions (10 m, average 5 days revisit time).

Sentinel-2 images have 13 spectral bands ranging from visible to mid-infrared wavelengths. The red and near-infrared bands of the electromagnetic spectrum are used to calculate the NDVI values.²⁸ Theoretically, NDVI is computed as indicated in the following equation:

$$\text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_R}{\rho_{\text{NIR}} + \rho_R}, \quad (1)$$

where ρ_R and ρ_{NIR} denote, respectively, the red (665 nm) and the near-infrared (842 nm) spectral reflectance measurements.

From January 2016 to October 2020, 258 Sentinel-2 images were acquired from THEIA.²⁹ These data are free downloaded content and ready-to-use. Since the percentage of cloud cover of some data is high, a data pre-selection has been made where the cloudiest data were eliminated based on the mask of cloud provided by THEIA.

After the filtering step for each study area, 251 observations were obtained for Jbeniana, 242 for Limaya, and 246 for Chaâl. Table 1 summarizes the descriptive statistics of NDVI time series of each area.

According to Table 1, NDVI average value corresponding to the Limaya area is the highest. This is due to the high vegetation density since trees are too close in hyper-insensitive olive grove. Moreover, in rainy season, the NDVI value reaches its maximum equal to 0.353, 0.622, and 0.155. Conversely, in dry season, the NDVI values were at their minimum level. The region of Jbeniana, Limaya, and Chaâl have minimum NDVI value of 0.123, 0.218, and 0.102, respectively.

In addition, the skewness and kurtosis values are two indicator values that allow us to determine the shape of the NDVI distribution compared with the normal one. For instance, the NDVI

Table 1 Descriptive statistics of NDVI time series for each area.

Value	Region name		
	Jbeniana	Limaya	Chaâl
Observations	251	242	246
Mean	0.1995	0.3535	0.1233
Q1	0.1563	0.2894	0.1113
Median	0.1864	0.3506	0.1208
Q3	0.2441	0.405	0.1341
Minimum	0.1231	0.2185	0.1026
Maximum	0.3533	0.6219	0.1555
Variance	0.0026	0.0059	0.0002
Skewness	0.600	0.6709	0.5284
Kurtosis	2.3614	3.6214	2.2058

distribution of Limaya area is unapplied and not asymmetric since the skewness value was different to zero and the kurtosis was less than 3.

The mean NDVI reflectance of the red and infrared bands was determined for each image by averaging the pixel values. The mean reflectance curves of the red (RED) and near-infrared (NIR) bands and the NDVI curves of each region are shown in Fig. 2.

Fig. 2 Reflectance and NDVI time series from January 2016 to October 2020 of (a) Jbeniana, (b) Limaya, and (c) Chaâl areas.

According to Fig. 2, near-infrared reflectance is higher than the red one since the chlorophyll absorption is located in the red domain. Moreover, the leaf chlorophyll abundance is seasonal. Their highest values are reached during the rainy season (September to March), which is why the mean reflectance of the red curve between September and March is decreased. It is important to notice that the near-infrared reflectance decreases during the same period but the effect is less pronounced. It is explained by the decrease of the earth surface brightness due to the sun low elevation, which increases the shadow area.

Based on Fig. 2, for each region, we notice that the variation of NDVI time series and reflectance curves was in opposition, where over time, NDVI curves are subject to seasonal oscillations. These curves increase between September and March, which corresponds to the growing period, and decreases between April and August, which corresponds to the no-growing and senescence period. This variation is due to the anthropogenic influence and the effect of climate change, such as temperature, precipitation, and humidity.

In addition, the range of NDVI values has varied depending on the type of covered vegetation. For instance, Limaya, as it was an irrigated olive trees field, has the highest NDVI values ($\text{NDVI} \in [0.219; 0.622]$). While the NDVI of Chaâl area, which is characterized by a non-irrigated olive trees field ranges from 0.103 to 0.156. Finally, Jbeniana area is a mixture of olive and agriculture field, its NDVI values are between 0.123 and 0.353.

3 Methodology

The proposed process adapted to predict the plant growth for different regions of Sfax, Tunisia, is presented in Fig. 3.

Our methodology allows us to predict the NDVI time series from January 2016 to October 2020 for different regions of Sfax, Tunisia, using statistical methods based on satellite data. To do it, the mentioned process in Fig. 3 has been adopted. It takes the Sentinel-2 image time series as input and the predicted NDVI time series as output. This process is composed of four important steps. The first one is the pre-processing. It converts the data generated by Sentinel-2 to the standard monthly NDVI time series. Then, a classification step is executed to split the field into

Fig. 3 Illustration of the adopted methodology.

various clusters for a precise forecasting. This allows us to have a monthly NDVI time series for each cluster. In the third step, the forecasting is applied to the NDVI time series using both techniques AR and MC. The final step consists of evaluating the performance of the predicted measurement.

3.1 Pre-Processing

The dataset is filtered using the mask provided by THEIA in order to eliminate the cloudy data. Since the statistical study of NDVI is skewed due to the presence of extreme outliers, the dataset is collected by month to obtain more smoothed and regular NDVI distribution. To do so, the NDVI dataset is collected by month and averaged. About 52 NDVI values were obtained for each study area. The correlation between NDVI profiles and plant phenology, month by month is so significant that we can understand from it whether the cycles are early or late. Therefore, this new time series will be used in our prediction.

3.2 Classification

As the vegetation development is not the same for all the areas, the distinction between the types of vegetation covers of the same study area can increase the prediction accuracy. Therefore, each area was divided up into four homogeneous region. To do so, the Gaussian mixture model (GMM),³⁰ which is a clustering algorithm, is adopted. The different clusters of each study areas are identified in Fig. 4.

According to ground truth, in Jbeniana area, C4 and C2 represent vegetable a cereal fields, C3 represents the olive trees, and C1 represents tracks and urban areas. For Limaya area, C4, C3, and C2 correspond to large, medium, and small olive trees, respectively, whereas C1 corresponds to the tracks between fields. For Chaâl area, C2 corresponds to grass and olive trees, C4 and C3 correspond to large and medium no-irrigated olive trees, respectively, and C1 corresponds to the small olive trees.

For the different study areas, the monthly NDVI time series of each clusters were obtained during the study period. The variations of the monthly NDVI time series of each region are presented in Fig. 5.

Fig. 4 Image classified of (a) Jbeniana, (b) Chaâl, and (c) Limaya areas.

Fig. 5 The monthly NDVI time series variation as a function of time of each clusters for (a) Jbeniana, (b) Limaya, and (c) Chaâl areas.

According to Fig. 5, for each region, NDVI time series of the different clusters have the same fluctuation and follow a similar variation as a function of time. The cycle is clear and stable. In the middle period of September and March, which is a rainy period, the monthly NDVI curves are increasing and it reaches its maximum. From middle of April to August, which is a dry period, the NDVI curves are decreasing and they reach their minimum.

C4 of Jbeniana, Limaya, and Chaâl areas have the highest NDVI values, which correspond, according to Fig. 4, to the densest clusters. Likewise, C1 of Jbeniana, Limaya, and Chaâl areas, which represent the least dense, have the lowest NDVI values.

3.3 Forecasting Step: Prediction Methods

In this step, two different statistical methods, AR and MC, are designed to forecast the monthly NDVI time series based on historical NDVI data. These methods are described in the next subsection.

3.3.1 Autoregressive

This model³¹ consists of predicting current state based on past observations. This permits to obtain the future value. The prediction value is linked linearly to the past values, which are selected depending on the used k -order. Let $k \in \mathbf{N}^*$, a k -order of autoregressive $AR(k)$ is expressed as follows:

$$Y_t = p_1 Y_{t-1} + p_2 Y_{t-2} + \dots + p_k Y_{t-k} + \varepsilon, \quad (2)$$

where $p_1, p_2, \dots, p_k$ represent the parameters of the model, $Y_t, Y_{t-1}, \dots, Y_{t-k}$ are the mean of NDVI values of the month $t, t-1, \dots, t-k$, and ε is a random variable that represents the error.

Table 2 The k -order of AR corresponds to each month.

Order (k)	1	2	3	4	5	6
Months of first cycle	10	11	12	1	2	3
Months of second cycle	4	5	6	7	8	9

In our study, the distribution of the NDVI time series shows different fluctuations with extreme values, which could affect the results of the prediction since AR is considered as a linear estimator. The estimation is based on prior measurements for maximum and minimum values, which are either low or high measurements relative to the considered one, leading to either an under or over estimation. To overcome this issue, the distribution is divided up into two separate cycles based on its trend.

The first cycle corresponds to the increase in the NDVI distribution and the second cycle corresponds to its decrease, which correspond to the growing period during September to March and senescence period between March and September, respectively.

For each cycle, the month is predicted using the previous months of the same cycle. Since the increase in k -order provides in general more reliable prediction outcomes, for each month, the highest possible order has been chosen. Table 2 shows the k -order used for each month of cycle.

3.3.2 Markov model

An MC³² is a particular type of stochastic process that sequentially moves from one state to another in the state space $E = \{e_k; 1, 2, \dots, N\}$ (where N is the number of E states). It is characterized by a transition probability $\pi_{i,j}^t$. It is the probability that the MC is at the time $t + 1$ point in state e_i , given that it is at the current time t point in state e_j . It is defined as follows. For all $i, j \in \{1, \dots, N\}$,

$$\pi_{i,j}^t = P(y^{t+1} = e_i | y^t = e_j), \quad (3)$$

where y^t corresponds to the vector of NDVI values at time t , which represents the month.

The transition probability $\pi_{i,j}^t$ creates a square matrix $\Pi^{\text{ERROR: NO BASE FOR SCRIPT}_{t,t+1}}$. It contains all the possibilities for switching from one state to another for two successive months t and $t + 1$,

$$\Pi^{t,t+1} = [\pi_{i,j}^t]_{i,j \in \{1, \dots, N\}} = \begin{pmatrix} \pi_{1,1}^t & \pi_{1,2}^t & \dots & \pi_{1,j}^t & \dots & \pi_{1,N-1}^t & \pi_{1,N}^t \\ \pi_{2,1}^t & \pi_{2,2}^t & \dots & \pi_{2,j}^t & \dots & \pi_{2,N-1}^t & \pi_{2,N}^t \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots \\ \pi_{i,1}^t & \pi_{i,2}^t & \dots & \pi_{i,j}^t & \dots & \pi_{i,N-1}^t & \pi_{i,N}^t \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots \\ \pi_{N-1,1}^t & \pi_{N-1,2}^t & \dots & \pi_{N-1,j}^t & \dots & \pi_{N-1,N-1}^t & \pi_{N-1,N}^t \\ \pi_{N,1}^t & \pi_{N,2}^t & \dots & \pi_{N,j}^t & \dots & \pi_{N,N-1}^t & \pi_{N,N}^t \end{pmatrix}. \quad (4)$$

In our case, NDVI interval that is ranged from 0 to 1 is divided into N intervals of width 0.05, where each interval presents the state space $e_{i;i \in \{1, \dots, N\}}$, i.e., $E = \{0, 0.05, 0.10, \dots, 1\}$.

Using the transition matrix, the prediction of the next state is calculated based on the current state. More precisely, if $z^t = (P(y^t = e_i))_{i \in \{1, \dots, N\}}$ denotes the state vector probability distributions at time t , then

$$z^{t+1} = \bar{\Pi}^{t,t+1} \times z^t, \quad (5)$$

where

$$\bar{\Pi}^{t,t+1} = \frac{1}{N_y} \sum_{i=0}^{N_y} \Pi^{t+12,t+1+12i} \quad (6)$$

is the mean of the previous transition matrix $\Pi^{t,t+1}$ of the maximum possible years N_y for the same couple of months $(t, t+1)$. This transition probability varies depending on the month. Therefore, 12 transition probability matrices were obtained, each corresponding to two consecutive months.

Furthermore, as the dimension of the state vector z^t is N , this vector is averaged using the below equation to obtain the NDVI values, Y_t , at time t :

$$Y_t = \frac{1}{N} \sum_{i=1}^N z^t e_i. \quad (7)$$

3.4 Model Performance

For all methods, the root mean squared error (RMSE) is used to measure the quality of the prediction. It represents the square root of the differences between the predicted and the observed values. It is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N_m} \sum_{i=1}^{N_m} (Y_t - \hat{Y}_t)^2}, \quad (8)$$

where Y_t and $\hat{Y}_t$ denote, respectively, the vectors of observed and predicted NDVI values, and N_m is the number of months.

4 Experiments Results

In this section, for the three selected areas, the prediction of the monthly NDVI is done using the two different forecasting methods described in Sec. 3. For each method, the prediction is carried out with and without classification. Moreover, in order to evaluate the performance and goodness of this prediction methods, the RMSE is computed for each NDVI time series.

4.1 Without Classification

In this subsection, the three regions' NDVI time series are forecasted using the two prediction methods, which take into account all of the data from the entire region. The observed and forecasted NDVI curves are given for each region during the study period in Fig. 6. The black and red curves represent the forecasted NDVI time series using AR and MC methods, respectively, and the blue curve represents the observed NDVI time series.

According to Fig. 6, the predicted curves have the same form as the original curve and have the same fluctuation. But, using MC method, the predicted NDVI time series are more closer to the actual one than AR model. It is clear that using MC method, the error decreases as a function of the time. This is explained by the increase in database learning.

As explained by Eq. (8), the RMSE is calculated for each of two prediction methods. The obtained numerical results are shown in Table 3.

Using both methods, the obtained errors are less than 3% for Chaâl, 5% for Limaya and Jbeniana. The low error observed in Chaâl is due to the slow variation of the NDVI along the years. Indeed, this area is characterized by the smallest vegetation cover.

Fig. 6 The observed and forecasted NDVI time-series using AR and MC methods for (a) Jbeniana, (b) Limaya, and (c) Chaâl areas.

Table 3 Performance of the forecasting methods.

Methods	Area		
	Jbeniana	Limaya	Chaâl
AR	0.0303	0.0461	0.0115
MC	0.0225	0.0274	0.0059

4.2 With Classification

In this section, a classification method is adopted to split the area into four homogeneous clusters C_i with $i \in \{1, 2, 3, 4\}$. Each cluster's prediction is done using two ways based on the training data. First, the learning data are generated taking into account the entire region (full learning), and second, the learning data are built with only the data from the class itself (partial learning).

The NDVI time series of the three regions are predicted using AR and MC models based on full and partial learning. The prediction results of Jbeniana, Limaya, and Chaâl are illustrated in Figs. 7, 8 and 9, respectively.

Table 4 gives the numerical values of the RMSE corresponding to each method and to each learning process.

Table 4 discloses that the two forecasting learnings show accurate prediction. However, the predictions based on the partial learning are more accurate than the prediction based on the full learning. This is due to the separation of different vegetation types that evolve in different ways and must be predicted separately. For that reason, subsequently, we will adopt the partial learning.

Fig. 7 Comparison of the two approaches for each prediction methods of Jbeniana area.

4.2.1 Cluster forecasting

For each cluster of each area, the forecasting is done using the AR and MC models. The obtained results are illustrated in Fig. 10.

The blue curve represents the observed NDVI time series collected by month, and the black and red curves represent the predicted NDVI time series using AR and MC models, respectively.

The distributions of the observed and predicted NDVI time series have the same oscillations over time. In Limaya and Chaâl areas, the predictions have more stable cycle than Jbeniana area. This is due to the change in fundamental culture of Jbeniana area every year, which depends on

Fig. 8 Comparison of the two approaches for each prediction methods of Limaya area.

the season and the farmer's choice of crop. This explains the not perfectly smoothing of the measured and expected curves. Moreover, the cycle of C3 and C4 of Limaya area are clear and stable as they are dense olive trees with well-known periodicity. Therefore, the prediction is good. Specifically, using MC model, the performance is very high where the RMSE result is less than 0.03. In C1 and C2 of Limaya area, the variation from a month to another is low. For this reason, with the AR model, the predicted curves approach the ones using MC.

Fig. 9 Comparison of the two approaches for each prediction methods of Chaâl area.

Table 5 gives the numerical values of the RMSE for different statistical methods.

Table 5 highlights that the prediction for different prediction methods are lower than 5%. Moreover, for all NDVI time series, the difference in performance between the AR and MC model is clear where the MC outperforms the former as it takes advantage from all the dataset.

Moreover, among the three study areas, Chaâl's RMSE findings, independent of the cluster, have the smallest error compared to the other areas. The RMSE mean that corresponds to this

Table 4 RMSE of different clusters for each study areas.

Methods	Area					
	Jbeniana		Limaya		Chaâl	
	AR	MC	AR	MC	AR	MC
<i>C1</i>						
Full learning	0.0182	0.0176	0.0354	0.0209	0.0113	0.0069
Partial learning	0.0176	0.0152	0.0352	0.0206	0.0109	0.0067
<i>C2</i>						
Full learning	0.0341	0.0243	0.0511	0.0319	0.0094	0.0056
Partial learning	0.0332	0.0227	0.0504	0.0301	0.0086	0.0050
<i>C3</i>						
Full learning	0.0462	0.0356	0.0584	0.0360	0.0098	0.0076
Partial learning	0.0451	0.0304	0.0483	0.0254	0.0089	0.0069
<i>C4</i>						
Full learning	0.0464	0.0371	0.0572	0.0437	0.0232	0.0113
Partial learning	0.0462	0.0312	0.0456	0.0292	0.0225	0.0109

area is of order 0.006. This is due to the minor difference in variation of the distribution from January 2016 to October 2020. Nevertheless, the Jbeniana's RMSE results have the highest error due to the variation of various vegetation. The mean of RMSE is about 0.025.

5 Discussion

In this paper, we developed two statistical methods, AR and MC, to predict the NDVI time series based on the past NDVI dataset derived from Sentinel-2 observations from January 2016 to October 2020. To check the reliability of our approach and to ensure its success regardless of the vegetation type, it has tried and proven in three different areas characterized by various types of vegetation with different densities. Based on the obtained results, we made the following observation. First, both statistical methods are able to predict the NDVI time series with a high performance (lower RMSE <5%). Furthermore, the predicted NDVI time series curves using AR and MC seem to follow the same trend as the actual observation and with an acceptable accuracy (RMSE <0.05). Nevertheless, the MC model produces a predicted curve closer to the actual one than the AR model, which shows a disturbance at the extreme outlines. In fact, the AR model is not able to detect abrupt changes in the seasonal component of the simulated NDVI time series due to its linearity.

Contrariwise, the MC model is able to detect the seasonal changes (with good precision) as well as any abrupt changes in the NDVI time series. This is confirmed in terms of RMSE, which is around 0.003 and 0.005 for MC and AR models, respectively. Second, whatever the nature of vegetation, the obtained results show that the accuracy, adaptability, and efficiency of MC model in predicting the time series. This allows us to conclude that the MC model gives a precise and accurate predicted NDVI time series.

In comparison to prior studies, for example in Refs. 33 and 34, the authors forecast NDVI using different approaches based on data provided by MODIS sensors, where the obtained

Fig. 10 The observed and forecasted NDVI curves using AR and MC models of the three regions.

RMSE is around 9% and 7%, respectively. Moreover, in Ref. 35, long-term convolutional memory (ConvLSTM), a deep learning architecture based on RNNs, is presented to perform much more comprehensive and detailed NDVI forecasts. The RMSE of the ConvLSTM is 0.08. Even though we use different satellites with different spatial and temporal resolutions, when comparing the state-of-the-art technique to our MC model, we can say that our model with RMSE less than 0.03 is competitive and successful in predicting the NDVI time series.

By adopting the statistical method described in the paper, it is possible to accurately predict vegetation cover, density, and health. Our approach can be used in the future to identify and diagnose diseases at an early stage that helps to take proactive measures to protect and improve the crop yield. Indeed, visualizing the next stage of vegetation helps to indicate the presence of abnormal behavior if a plot differs in a way from what is expected. More precisely, if a plot or a part of it behaves differently compared to the last years or the neighbors, it is likely a sign of a stress.

Table 5 Performance of the different statistics methods.

	C N°				
Methods	1	2	3	4	All pixels
Jbeniana area					
AR	0.0176	0.0332	0.0451	0.0462	0.0303
MC	0.0152	0.0227	0.0304	0.0312	0.0225
Limaya area					
AR	0.0352	0.0504	0.0483	0.0456	0.0461
MC	0.0206	0.0301	0.0254	0.0292	0.0274
Chaâl area					
AR	0.0109	0.0086	0.0089	0.0225	0.0115
MC	0.0067	0.0050	0.0069	0.0109	0.0059

6 Conclusion and Future Work

To improve the productivity and maintain the crop health, it is important to forecast their growth along the year to be assured about its behavior. In this context, this paper presents two statistical methods, AR and MC models, to forecast the plant growth through NDVI values using remote sensing data. The study was conducted for three agriculture areas located in Sfax, Tunisia. They are composed mainly of olive groves with different types and densities in order to assess and ensure that our methodology is profitable and reliable in various cases. We used monthly NDVI time series derived from the red and near-infrared bands of the Sentinel-2 images for the period from January 2016 to October 2020. To ensure the reliability of our prediction approach, it should be applied on homogeneous areas for that we need to distinguish between the different types of vegetation cover as well as between the same vegetation kind but with different growth stages. To do it, a GMM model was applied to decompose each study site into different homogeneous areas. The obtained results show that this classification increases the forecast accuracy. Moreover, the predicted NDVI time series curves of the three study sites revealed that they have the same oscillation trends in time and show two main and stable cycles: A high cycle characterized by the highest NDVI values during the growing season (September to March) and a low cycle characterized by the lowest NDVI values during the dry season (April to August). The performance of the prediction was quantitatively tested using the RMSE between the actual and the predicted NDVI values. Prediction results based on full and partial training (i.e., using respectively the entire study site and the homogeneous areas separately) show high accuracy. Nevertheless, based on the partial training, the predicted curves are closer to the actual one and the obtained error is the smallest for all the cases. This confirms the decomposition into homogeneous areas usefulness. In addition, with the MC model, the predicted NDVI curves of the different study areas are the closest to the real ones where the RMSE is the smallest. This confirms that the prediction using MC outperforms the AR. This is due to the fact that MC takes advantage from all the dataset without any assumption about the temporal variation whereas the AR method assumes a linear relationship.

As a future scope of research, we would like to predict the NDVI time series by making in consideration other explanatory variables, such as temperature, humidity, and precipitation, since these features have an implicit impact on vegetation growth and health. We propose also to enhance our approach to predict the NDVI time series using for instance the multiple linear regression and the hidden MC.

References

1. R. Bongiovanni and J. Lowenberg-DeBoer, "Precision agriculture and sustainability," *Precis. Agric.* **5**(4), 359–387 (2004).
2. D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: key advances and remaining knowledge gaps," *Biosyst. Eng.* **114**(4), 358–371 (2013).
3. J. L. Hatfield et al., "Application of spectral remote sensing for agronomic decisions," *Agron. J.* **100**, S-117 (2008).
4. S. Chakraborty et al., "Time-varying modeling of land cover change dynamics due to forest fires," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **11**(6), 1769–1776 (2018).
5. J. Li et al., "Response of net primary production to land use and land cover change in mainland china since the late 1980s," *Sci. Tot. Environ.* **639**, 237–247 (2018).
6. C. Dineshkumar et al., "Phenological monitoring of paddy crop using time series MODIS data," *Multidiscipl. Digital Publ. Inst. Proc.* **24**(1), 19 (2019).
7. I. Rousta et al., "Impacts of drought on vegetation assessed by vegetation indices and meteorological factors in Afghanistan," *Remote Sens.* **12**(15), 2433 (2020).
8. P. J. Pinter, Jr. et al., "Remote sensing for crop management," *Photogramm. Eng. Remote Sens.* **69**(6), 647–664 (2003).
9. Z. Wan, P. Wang, and X. Li, "Using MODIS land surface temperature and normalized difference vegetation index products for monitoring drought in the Southern Great Plains, USA," *Int. J. Remote Sens.* **25**(1), 61–72 (2004).
10. D. Lloyd, "A phenological classification of terrestrial vegetation cover using shortwave vegetation index imagery," *Remote Sens.* **11**(12), 2269–2279 (1990).
11. C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sens. Environ.* **8**(2), 127–150 (1979).
12. M. Mkhabela et al., "Crop yield forecasting on the Canadian Prairies using MODIS NDVI data," *Agric. For. Meteorol.* **151**(3), 385–393 (2011).
13. M. Labus et al., "Wheat yield estimates using multi-temporal NDVI satellite imagery," *Int. J. Remote Sens.* **23**(20), 4169–4180 (2002).
14. R. Fensholt, I. Sandholt, and M. S. Rasmussen, "Evaluation of MODIS LAI, fAPAR and the relation between fAPAR and NDVI in a semi-arid environment using *in situ* measurements," *Remote Sens. Environ.* **91**(3–4), 490–507 (2004).
15. J. Huang and D. Han, "Meta-analysis of influential factors on crop yield estimation by remote sensing," *Int. J. Remote Sens.* **35**(6), 2267–2295 (2014).
16. F. Baret and S. Buis, "Estimating canopy characteristics from remote sensing observations: review of methods and associated problems," in *Advances in Land Remote Sensing*, S. Lang, Eds., pp. 173–201, Springer (2008).
17. R. Benedetti and P. Rossini, "On the use of NDVI profiles as a tool for agricultural statistics: the case study of wheat yield estimate and forecast in Emilia Romagna," *Remote Sens. Environ.* **45**(3), 311–326 (1993).
18. J. Liu et al., "Land-cover classification of China: integrated analysis of AVHRR imagery and geophysical data," *Int. J. Remote Sens.* **24**(12), 2485–2500 (2003).
19. B. Zheng et al., "A support vector machine to identify irrigated crop types using time-series landsat NDVI data," *Int. J. Appl. Earth Obs. Geoinf.* **34**, 103–112 (2015).
20. S. Huang et al., "A case study on a combination NDVI forecasting model based on the entropy weight method," *Water Resour. Manage.* **31**(11), 3667–3681 (2017).
21. L. Ji and A.-J. Peters, "Forecasting vegetation greenness with satellite and climate data," *IEEE Geosci. Remote Sens. Lett.* **1**(1), 3–6 (2004).
22. L. Wang et al., "Study on NDVI changes in Weihe watershed based on CA-Markov model," *Geol. J.* **53**, 435–441 (2018).
23. T. Wu et al., "A new approach to predict normalized difference vegetation index using time-delay neural network in the arid and semi-arid grassland," *Int. J. Remote Sens.* **40**(23), 9050–9063 (2019).
24. C. J. Tucker et al., "Satellite remote sensing of total herbaceous biomass production in the Senegalese Sahel: 1980–1984," *Remote Sens. Environ.* **17**(3), 233–249 (1985).

25. A. Ben Abbes et al., “Comparative study of three satellite image time-series decomposition methods for vegetation change detection,” *Eur. J. Remote Sens.* **51**(1), 607–615 (2018).
26. B. Martínez et al., “Characterizing land condition variability in Ferlo, Senegal (2001–2009) using multi-temporal 1-km Apparent Green Cover (AGC) SPOT Vegetation data,” *Global Planet. Change* **76**(3–4), 152–165 (2011).
27. R. de Jong et al., “Trend changes in global greening and browning: contribution of short-term trends to longer-term change,” *Global Change Biol.* **18**(2), 642–655 (2012).
28. J.-W. Rouse, Jr., et al., “Monitoring vegetation systems in the great plains with ERTS,” in *Third Earth Resources Tech. Satellite-1 Symp.*, E. P. Mercanti and M. A. Becker, Eds., NASA, Paper A-20, pp. 309–318 (1974).
29. THEIA (telescopic high-definition earth imaging apparatus), <https://theia.cnes.fr/>, Toulouse, France (2012).
30. C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, New York (2004).
31. J. D. Hamilton, *Time Series Analysis*, Vol. 2, Princeton University Press, Princeton, New Jersey (1994).
32. E. Cinlar, *Introduction to Stochastic Processes*, p. 420, Prentice-Hall, Englewood Cliffs, New Jersey (1975).
33. A. Berger et al., “Predicting the normalized difference vegetation index (NDVI) by training a crop growth model with historical data,” *Comput. Electron. Agric.* **161**, 305–311 (2019).
34. R. Ahmad et al., “A machine-learning based convLSTM architecture for NDVI forecasting,” *Int. Trans. Oper. Res.*, 1–24 (2020).
35. T. Wu et al., “A spatio-temporal prediction of NDVI based on precipitation: an application for grazing management in the arid and semi-arid grasslands,” *Int. J. Remote Sens.* **41**(6), 2359–2373 (2020).

Marwa Hachicha received her engineering degree in electronic systems and communication from the National School of Electronics and Telecommunications of Sfax, Sfax, Tunisia, in 2016. She is currently working toward her PhD in the University of Sfax, Sfax, Tunisia. She is currently a student researcher in Digital Research Centre of Sfax, Sfax, Tunisia. Her research interests include optical remote sensing data processing (satellite remote sensing), vegetation monitoring, statistic, and image processing.

Abdelaziz Kallel received his MS degree in telecommunications from the Higher School of Communication of Tunis (Sup’Com), Tunis, Tunisia, in 2003 and his PhD in physics from the Paris-Sud University, Orsay, France, in 2007. He was a postdoctoral scientist at the Laboratoire des Sciences du Climat et de l’Environnement, France, and the Tartu Observatory, Tartu, Estonia, in 2008 and 2009, respectively. Currently, he is an associate professor of signal processing at the Institut Supérieur d’Electronique et de Communication de Sfax, Sfax, Tunisia. His research interests concern optical and thermal remote sensing data processing and inversion, canopy reflectance modeling based on radiative transfer theory, data fusion using evidence theory, and image processing.

Mahdi Louati graduated from the University of Sfax. He received his PhD in applied mathematics from the Faculty of Sciences of Sfax in 2009 and his HDR degree in probability and statistics from the Faculty of Sciences of Sfax in 2015. He is a scientist at the Laboratory of Probability and Statistics of Sfax, Tunisia, since 2004. Currently, he is an associate professor of probability and statistics at the National School of Electronics and Telecommunications of Sfax, Tunisia. He is a data scientist and his research interests focus on probabilistic modeling, statistical learning, stochastic processes, random matrices, and estimation theory.