Development of a simplified technique for gap filling of Normalize Difference Vegetation Index (NDVI) time series data

M. Faisal
Drainage Research Institute, National Water Research Center, Egypt

R. S. Makar
Soils and Water Use Department, Agricultural and Biological Research Institute, National Research Centre, Dokki, Cairo, Egypt

*Corresponding author Email: randa_sgmm@yahoo.com

Abstract
The presence of gaps or missing values in time series prevents the practical use of such data. The current research aims at developing a simplified, straightforward technique for gap-filling the time series data of the Normalize Difference Vegetation Index (NDVI) generated using Moderate Resolution Imaging Spectroradiometer (MODIS). This research assumes that a relationship exists between the pixel location, date of acquisition and its NDVI value within a defined timeline. Therefore, two relatively simple methods were tested: the Multiple Linear Regression (MLR) analysis and the Artificial Neural Networks (ANN) to fill the NDVI missing values. While MLR is a well-known simple statistical method, the ANN has been successfully applied for the analysis of various scientific data, including the gap-filling of time series data. Nevertheless, ANN proved its supremacy in such approach. The accuracy of estimation utilizing the developed ANN model reached an average of R² of 0.8, while the average accuracy of MLR was about 0.3. Nevertheless, the developed model could only be applied within the same timeframe of the images used for developing the model. Otherwise, the accuracy of determination was reduced significantly. The results showed that according to its performance, ANN are promising for filling missing data of NDVI time series and could be applied to any other vegetation indices as well.

Keywords: Artificial neural networks, Gap filling, Multiple linear regression, MODIS-NDVI, Time series

INTRODUCTION
Remote sensing has become an important tool for the continuous monitoring of environment (Sarafanov et al., 2020). Regional and global environmental monitoring using remote sensing time series data is becoming important, especially for temporal vegetation patterns (Colditz et al., 2008). Moreover, changes in vegetation cover is crucial in climatic change studies, estimating water budget and for setting up environmental conservation strategies (Reddy and Prasad, 2018). The Normalized Difference Vegetation Index (NDVI) is considered one of the most popular indices used for vegetation assessment. Such popularity is related to the ease of calculation from any multispectral sensor with a visible and a near-IR band (Huang et al., 2021). The NDVI is highly useful in land cover classification, which is extremely beneficial in resource and environmental decision-making (Gandhi et al., 2015). Moreover, time series NDVI is commonly used to study vegetation type and its dynamics (Lan and Dong, 2022). However, these time series are sensitive to atmospheric conditions, in particular the presence of cloud cover, which results in the loss of significant data (Sarafanov et al., 2020). Although several methods have been proposed to reconstruct continuous NDVI time-series data, some challenges remain in the existing reconstruction methods (Chen et al., 2021). Linear interpolation and maximum value composite are the earlier classical time-series reconstruction methods. However, the first replaces the noise with linear estimation without considering vegetation growth, while the second ignores many details of the time-series (Liu et al., 2022). Other methods have various difficulties, such as the need for detailed climate data with the climate incorporated gap-filling (CGF) method (Yu et al., 2021) and the selection of a suitable maximum frequency with the Harmonic Analysis of Time Series (HANTS) (Zhou et al., 2022).
Over the past decade, the use of ANN increased considerably and compared with traditional statistical methods, ANN model is capable to model highly complex non-linear patterns of ecological processes (Atkinson and Tatnall, 1997). Researchers have used ANN for gap filling of vegetation index data series (NDVI, Kang et al., 2016 and the enhanced vegetation index (EVI), Nay et al., 2017). On the contrary, the pixel-based MLR approaches for NDVI gap filling is limited in the existing literature (Mohanasundaram et al., 2022). Moreover, one of the profound limitations of utilizing the gap in remote sensing data series was extracting more features about time series trends in a limited number of datasets. The present research compares between two simple, easy-to-use, repeatable methods to fill in the gaps in NDVI data within a designated timeframe. The proposed approach utilized location (latitude and longitude) of the pixel and the date of image acquisition to predict the missing NDVI values.

MATERIALS AND METHODS

Study area

The study area is located between coordinates 31° 12’ 36.7” N, 32° 02’ 02.4” E and 31° 8’ 38.0” N, 32° 07’ 28.9” E and covers an area of about 64.3 Km² (Fig. 1).

The selected study area could be classified into six land use/land cover classes, including: aquatic vegetation, vegetation, Lake El-Manzalah water, fallow soils, fishponds and urban area. The lake is located at the northeastern edge of the Nile Delta. The water quality of the lake is subjected to inflows from industrial, domestic and agricultural sources. Furthermore, overgrowth of aquatic vegetation is observed in the lake and large parts of the lake are converted into agricultural areas (Morgan and Faisal, 2018).

Remote sensing data processing

The Moderate Resolution Imaging Spectroradiometer (MODIS) is a high temporal frequency satellite sensor and is considered a valuable tool for mapping vegetation and monitoring seasonal dynamics (Colditz et al., 2008). It has provided observations for monitoring the earth’s surface since 1999 at almost daily complete global coverage (Chen et al., 2021). It acquires data in 36 spectral bands with spatial resolutions of 250 m (2 bands), 500 m (5 bands), and 1000 m (29 bands) (Savtchenko et al., 2004). The first two 250 m resolution bands were used in this study. These bands included one in the visible red range between 620 and 670 nm (Band 1) and the other in the near-infrared range between 841 and 876 nm (Band 2) (Moreno-Madrinan et al., 2010). MODIS data were downloaded from https://earthexplorer.usgs.gov. The data is in MOD09GQ Version 6.1 product providing surface spectral reflectance and is corrected for atmospheric conditions (Vermote, 2015 a and b). Six MODIS images were used in this research. These images were acquired on 30/04/2021, 08/06/2021, 04/07/2021, 29/07/2021, 01/09/2021 and 14/10/2021 and the Julian formats of these dates were 120, 159, 185, 210, 244 and 287 respectively. Using the QGIS’s Semi-Automatic Classification Plug-in (SCP), these data were imported and converted into Tiff format as described by Makar and Faisal (2021) and the two bands (Band 1 and 2) of each date were stacked together into one image. The raster calculator was used to produce the NDVI values for the selected dates. The NDVI was calculated according to Rouse et al. (1974) as:

\[
NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}}
\]

Where, \(\rho_{NIR}\) is the near-infrared reflectance and \(\rho_{RED}\) refers to the red reflectance.

NDVI values will be high and positive with green vegetation, which has high near infrared reflectance and high visible light absorption. On the other hand, other features such as water, soil and dry vegetation will have lower NDVI values because of high absorption in the near infrared wavelength (Myneni et al., 1995). Based on the field survey, the selected study area could be classified into six land use/land cover classes including the following: fallow soils, urban area, vegetation, aquatic vegetation, water and fishponds. For land use/land cover (LU/CL) classification Sentinel-2 (S2) and Landsat-8 data are used. While S2 data was acquired on the 8th May, 2021, Landsat-8 data were acquired on the 5th May, 2021. For S2, the blue, red, infrared and shortwave infrared bands were used, while for the Landsat-8 data the thermal bands. The images were pre-processed and processed according to Morgan and Faisal (2018), who used an integrated method of decision trees (DT) and ANN to classify the LU/CL. They recommended a series of data processing on these data and included a simple ratio between blue and first shortwave infrared bands was used to differentiate the water and fish ponds from the other classes and between each other (Fig. 2). Furthermore, the soil adjusted vegetation index (SAVI) developed by Huete (1988) (Eq. 2) was used to differentiate the vegetation and aquatic vegetation from other classes.

\[
SAVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED} + L} * (1 + L)
\]

Where, \(\rho_{NIR}\) is the near-infrared reflectance and \(\rho_{RED}\) refers to the red reflectance and \(L\) is the soil adjustment factor set to 0.5 as recommended by Huete (1988).

The first PCA of the thermal bands of Landsat 8 was also used to differentiate between the two types of veg-
etation. On the other hand, a specific PCA of the shortwave infrared bands of S2 was included to differentiate between the urban area & fallow soils classes. Thereafter, they produced a four-layer image from these images and applied the ANN on this image for LULC classification.

**Filling the gaps of the NDVI time series**

In this research, simple approaches were tested to predict the missing values of the NDVI. For each pixel, a relationship was assumed between the pixel location, date of acquisition and NDVI value within a defined timeline. The NDVI images were exported to MATLAB software and examined using Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN). MLR is considered the most common form of linear regression, which practically utilizes multiple variables (Zhang et al., 2022). The MLR formula (Eq. 3) is as follows (Wei et al., 2022):

\[
\text{NDVI} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \quad \text{Eq. 3}
\]

In the above, \(y\) is the NDVI value; \(a_0, a_1, \ldots, a_n\) are the model-fitting parameters; \(x_1, \ldots, x_n\) are the pixel variables; and \(n\) is the number of variables.

On the other hand, ANN was used to take advantage of its nonlinear modeling capability. The feed-forward backpropagation, which is a very popular ANN, was applied. The information flow in this ANN is from input layer to output layer via the hidden layers in one direction (Dada et al., 2021) (Fig. 2). The major strength of these networks is their ease of implementation and management and therefore, they are suitable for approximating any type of input and output (Hornik et al., 1989). When designing the ANN, 70% of the data were used for training, 15% for testing and 15% for validation. The ANN structure was manually adjusted until the highest correlation was achieved using a trial-and-error process (Gummadi, 2013 and Morgan et al., 2017). The ANN's performance was assessed in terms of the coefficient of determination \(r^2\).

**RESULTS AND DISCUSSION**

**Land use/ Land cover classification**

As mentioned in methodology, according to Morgan and Faisal (2018), four images were developed from processing S2 and Landsat-8 data. These four images included the SAVI, the simple ratio of blue to shortwave infrared band, the first PCA of the S2 shortwave infrared bands and the first PCA of the thermal bands of Landsat-8. These bands were stacked into a four-layer image and were used to classify the land use/land cover of the studied area. A total of 15 ground truth data locations representing the different classes were selected from the study area and polygons representing these locations were delineated on the four-layer image as region of interests (ROI). These ROI were used to collect the spectral signatures for different classes. Unlike Morgan and Faisal (2018) who used ANN to classify the image, the simple Maximum Likelihood classifier (MLC) was used to classify the LULC in the study area. The overall accuracy was 97.2% and therefore the results were accepted (Fig. 3). The agricultural area covered by vegetation and fallow soils covered 16.3% of the studied area while the aquatic vegetation covered 32.5% of the studied area. On the other hand, the Lake El-Manzalah water and fish ponds covered 45.9% while the urban area covered only 5.3% of the studied area.

**Filling the gaps of NDVI time series**

The NDVI images, which were used to develop the proposed approach to fill in the gaps, are shown in Fig.
4 (A, B, C, D, E, F representing Julian date of 120, 159, 185, 210, 244 and 287 respectively). In addition to these images, the image acquired on 09/05/2021 (129 Julian format) and was covered with clouds was used to validate the proposed method shown in Fig. 5A. Furthermore, examples of the NDVI values of the different LU/LC classes for the six dates as well as the cloudy image are shown in Fig. 5B. In this figure, the effect of cloud coverage on the NDVI values could be seen in NDVI value.

To process these images in MATLAB, each NDVI image should be imported into MATLAB and merged into a single file for all these dates. Nevertheless, that was impossible due to the large data size. Therefore, each image was divided into 28 sub-images using the VRT option of the QGIS software and thereafter, imported into MATLAB in a tabular format. Each sub-image file

The MLR was applied to each of the 28 files and the accuracy in the form of $r^2$ are shown in Fig. 6. Accordingly, the average of this method throughout the images was $r^2$ of 0.28. Nevertheless, it was observed that $r^2$ slightly exceeded 0.5 in sub-images 8, 9, 10 and 11 but that only represented less than 15% of the studied area and therefore, this approach was not accepted in gap filling of NDVI time series.

Thereafter, the MATLAB’s Neural Network Toolbox was used to design the ANN to fill the gaps in the NDVI time series.

The pixel's latitude, longitude and the date of acquisition expressed in Julian format were used as the ANN input nodes. The ANN structure was manually alternated using a trial-and-error process until the best performance was achieved. The designed network included a three-node input layer, one-node output layer representing the NDVI values and a hidden layer with 10 nodes (Fig.7) and the average accuracy of estimation reached $r^2$ of 0.80 and the accuracies of the 28 sub-images ranged from 0.69 and to 0.90 (Fig.8). The cloudy MODIS-NDVI image acquired on 09/05/2021 (129 in Julian format) was subsetted for the study area (Fig.9A) to test the method validation. The image was divided into 28 sub-images, as previously mentioned. The ANN designed for each sub-image was applied to predict the missing NDVI values. Thereafter, the 28 sub-image was imported into QGIS and merged into one image (Fig.9B).

Furthermore, to test the accuracy of the network outside the timeline of the designed network, the ANN was applied on the NDVI images acquired on 15/4/2021 and 7/11/2021, which stand for 105 and 311 in Julian format, respectively. The results revealed that the prediction accuracy outside the timeframe declined to reach $r^2$ of less than 0.55 ($r^2 = 0.50$ for 105 and $r^2 = 0.53$ for 311).

**Conclusion**

This research compared MLR and ANN, two simple, straightforward pixel-based methods, for gap-filling time series of MODIS-NDVI data that were incomplete due to cloud coverage over the study area. The results revealed that ANN is more accurate than MLR. The proposed ANN approach presented in this research is...
easy and repeatable and could be applied to a single pixel or an entire missing image. The method is applicable within the timeframe of the designed ANN only but has the advantage of requiring few images to design the artificial neural network. In the present research, only one image was used per month. Even so, the method was restricted to NDVI-MODIS products and is applicable to a wide variety of remotely sensed time-series regardless of the cause of the missing data.

Conflict of interest
The authors declare that they have no conflict of interest.

REFERENCES

6. Gummadi, J.A. (2013). Comparison of various interpolation techniques for modeling and estimation of radon concentrations in Ohio, Master of Science Degree in Engineering, The University of Toledo, Toledo, USA.
15. Morgan, R.S., Abd El-Hady, M., Rahim, I.S., Silva, J. &


1506