Cyprus University of Technology

杭州電子科技大學 HANGZHOU DIANZI UNIVERSITY

Master's Thesis

Prediction Study of Vegetation Dynamics in the Troodos Mountains Based on NDVI Time Series

Ziheng Huang

Limassol, May and 2025 of thesis submission

MSc in Electronics Science and Technology

CYPRUS UNIVERSITY OF TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER ENGINEERING AND INFORMATICS

Master's Thesis

Prediction Study of Vegetation Dynamics in the Troodos Mountains Based on NDVI Time Series

Ziheng Huang

Supervisor

Professor Takis Kasparis

Limassol, May and 2025 of thesis submission

Approval Form

Master's Thesis

Prediction Study of Vegetation Dynamics in the Troodos Mountains Based on NDVI Time Series

Presented by

Ziheng Huang

Supervisor: Takis Kasparis

Member of the committee: Michael Mavrovouniotis

Member of the committee: Stelios Neophytidis

Cyprus University of Technology

Limassol, May and 2025 of thesis submission

Copyrights

Acknowledgements:

I would like to express my heartfelt gratitude to Professor Takis Kasparis, Dr. Michael Mavrovouniotis, and Researcher Stelios Neophytidis for their invaluable academic support and guidance throughout this research. I am also deeply grateful to all the staff members at The ERATOSTHENES Centre of Excellence for their generous help and support during my time at the institute. Special thanks go to both Hangzhou Dianzi University and the Cyprus University of Technology for providing the academic environment and resources that made this work possible.

ABSTRACT

This study generates and analyzes time series based on the Normalized Difference Vegetation Index (NDVI) to predict vegetation dynamics in the Troodos Mountains. Renowned for its ecological diversity, the Troodos Mountains represent a typical Mediterranean mountain region. However, significant vegetation changes due to climate change and human activities highlight the necessity of understanding and forecasting ecological shifts.

Using multi-year NDVI data, the study examines temporal changes in vegetation cover, focusing on seasonal fluctuations and long-term trends. Advanced statistical and machine learning methods, including time series forecasting and regression models, are employed to predict future vegetation changes.

This research provides critical insights for ecological management, conservation, and sustainable land-use planning, addressing the challenges posed by climate change and environmental pressures.

Keywords: NDVI; Time Series Analysis; Remote Sensing; Machine Learning; Google Earth Engine

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GIS:	Geographic Information System
RS:	Attention Deficit-Hyperactivity Disorder
BMI:	Body Mass Index
NDVI:	Normalized Difference Vegetation Index
GEE:	Google Earth Engine
MSE:	Mean Squared Error
RMSE:	Root Mean Squared Error
Dropout:	Dropout Regularization
DBSCAN:	Density-Based Spatial Clustering of Applications with Noise
IQR Filter:	Interquartile Range Filter
LOF:	Local Outlier Factor

1 Introduction

The dual influence of global climate change and human activities has significantly exacerbated the vulnerability of the ecological environment, especially in mountain ecosystems. As a key repository of global biodiversity, mountain ecosystems play an irreplaceable role in climate regulation, maintenance of hydrological cycle, and formation of carbon sinks. As the core component of mountain ecosystems, vegetation is not only an important carrier of the carbon cycle but also the cornerstone for maintaining ecological service functions. Therefore, monitoring the changes of mountain vegetation has significant ecological value and economic significance.[1].

In recent years, global climate change and the increase in human activities have triggered profound changes in mountain vegetation, leading to a decline in ecosystem stability and the deterioration of its service functions. This evolution process is mainly driven by factors such as rising temperatures and alterations in precipitation patterns. It not only endangers the ecological sustainability of mountainous regions but also weakens their environmental regulatory functions on a larger scale.[2].

The Troodos Mountains, located in the eastern part of the Mediterranean Sea in Cyprus, constitute a typical mountain ecosystem with complex climatic conditions, diverse topography and rich vegetation types. This unique combination makes this region an ideal place for studying the dynamic changes of mountain vegetation. However, this ecosystem is currently facing increasing pressures from climate warming, changes in precipitation patterns, agricultural expansion and urbanization brought about by human activities. Addressing these challenges requires a deep understanding of how these factors interact with each other and how they affect mountain vegetation and ecosystem services.

Changes in vegetation cover, particularly issues like drought, forest degradation, and soil erosion, have critically undermined the ecological security of the Troodos Mountains.[3]. Accurate monitoring of the changing trend of vegetation coverage and identification of its driving factors are of vital importance for formulating effective strategies for ecological protection and sustainable development.

Normalized Difference Vegetation Index (NDVI) is a widely recognized remote sensing tool that quantifies the reflection of red light and near-infrared light to provide an effective means for monitoring vegetation dynamics and is used to assess the growth status and

coverage of vegetation. Despite its wide application, the complex terrain and climatic conditions of the Troodos Mountains pose significant challenges to the application of NDVI data.[4, 5]. Vegetation distribution and growth in this region are shaped by diverse factors, including temperature, precipitation, and topography, resulting in spatial variability and temporal dynamics in NDVI patterns.

To address these challenges, advanced methods are needed to enhance NDVI data analysis. Integrating innovative technologies, such as deep learning, provides a way to achieve higher accuracy in vegetation trend research, thereby facilitating the understanding of mountain ecosystem changes and supporting targeted conservation efforts.

$$\boldsymbol{NDVI} = \frac{\boldsymbol{NIR-Red}}{\boldsymbol{NIR+Red}} \tag{1.1}$$

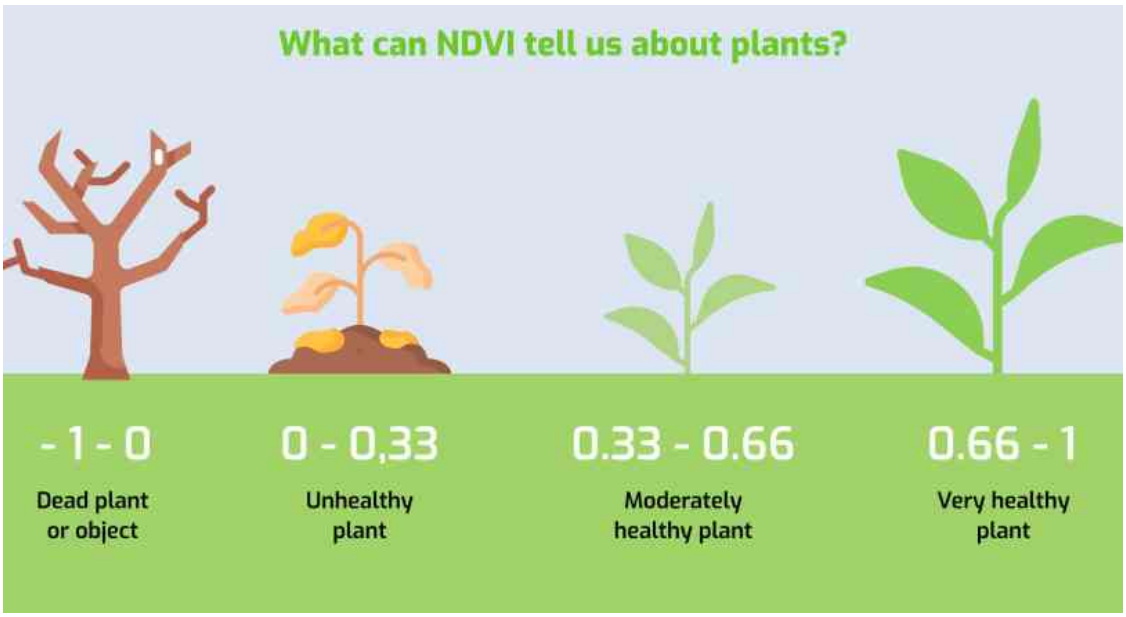

Figure 1: Vegetation growth reflected by NDVI

This study utilized NDVI data derived from remote sensing images, combined with advanced deep learning techniques, to systematically investigate the spatial-temporal characteristics and influencing factors of vegetation coverage in the Troodos Mountains. Moreover, the study aimed to predict the future trends of vegetation dynamics under different environmental and human activities conditions.

By constructing a powerful and integrated model that integrates climate variables, human activity indicators and topographic features, this study aims to reveal the patterns and

mechanisms of vegetation coverage changes. Specifically, the research will focus on analyzing the vegetation coverage situation over the past few decades, identifying temporal and spatial trends and variation magnitudes, determining the key periods and regions of significant changes, and predicting potential scenarios of future vegetation dynamics.

The core objective of this study is to enhance the accuracy of vegetation change prediction using NDVI data through deep learning techniques. The key research questions are:

1、 How can effective data preprocessing be conducted to manage high-dimensional, noisy, and incomplete data, ensuring robust time series?
2、 How can a deep learning model be designed to accommodate the unique seasonal variations in the Troodos Mountains for accurate predictions?
3、 Based on NDVI data analysis results, what strategies can be proposed for vegetation conservation and ecological restoration in the Troodos Mountains?

To address these questions, this research will pursue the following specific objectives:

1、 **Data Preprocessing:** Develop and optimize NDVI data preprocessing techniques, including noise removal, anomaly detection, and data imputation, to enhance input data quality[6]. Techniques such as LOF (Local Outlier Factor) and 3-sigma will be explored to reconstruct time series while avoiding over-smoothing.

2、 **Deep Learning Modeling:** Design deep learning models that are adapted to the vegetation distribution characteristics of the Troodos Mountains. These models will extract vegetation change patterns from NDVI data to achieve high-precision predictions. Traditional statistical methods (such as linear regression) often fail to effectively handle high-dimensional, nonlinear, noisy and incomplete remote sensing data [4]. Moreover, conventional models may overlook valuable parameters and texture features from RGB images [5]. Compared with traditional models (such as XGBoost), this study will attempt to use advanced models, such as TimeGPT, for training and result comparison. The key to predicting vegetation coverage series over long time periods lies in how to fully utilize the rich seasonal patterns and sequential relationships through time series to complete the classification task. Wang Haoyu and Zhao Xiang et al. effectively capture temporal correlations using recurrent neural networks (RNNs), especially long short-term memory networks (LSTMs) [7].

3、 **Data Acquisition and Processing:** The NDVI time series data will be obtained through the Google Earth Engine platform using the Landsat dataset, and 100 random points will be selected within the specified Troodos Mountain area for analysis. Data preprocessing will include spatial-temporal registration, outlier removal (for example, using the LOF method), and multimodal data fusion.

This research mainly takes the time series prediction of NDVI as the starting point, introduces the specific research area and acquires data after introducing and reviewing the literature. The research methods include data preprocessing (removing outliers and repairing missing values), as well as time prediction models based on deep learning. Subsequently, the results are analyzed and discussed, and the reliability and accuracy of the model are evaluated through specific indicators. Finally, in the conclusion, we summarize and look forward to the future development trends.

2 Backround

In recent years, China has made significant progress in NDVI-related research, covering a wide range of applications, including vegetation dynamics monitoring, ecosystem health assessment, and climate change response. The following are key directions and case studies:

• Deepening the Application of Remote Sensing Data:

Chinese scholars have utilized data from Landsat, MODIS, and domestically produced high-resolution satellites (GF-1, ZY-3) to monitor vegetation cover changes over long time spans. For instance, MODIS NDVI data were used to study the spatial-temporal evolution of vegetation cover on the Tibetan Plateau, revealing the sensitivity of different climatic zones to temperature and precipitation. Liu Qionghuan used the DBEST method to detect the characteristics of four different vegetation change trends[7]. The RF method was then applied to identify the driving factors for vegetation changes in each direction based on climate change, water source replenishment (i.e., proximity to rivers, lakes, glaciers), human interference, and climate features[8].

Yang Yuanhe and Park Shilong analyzed NDVI data for grassland vegetation on the Tibetan Plateau using data from the Global Inventory Monitoring and Modeling Studies (GIMMS) research group, ultimately drawing conclusions on the rate of change and identifying seasonal NDVI trends across different grassland types on the Plateau [9].

	*NDVI*变化量 *NDVI* changing magnitude (a^{-1})				*NDVI*变化率 *NDVI* changing ratio (a^{-1})			
	生长季 Growing season	春季 Spring	夏季 Summer	秋季 Autumn	生长季 (%) Growing season	春季 (%) Spring	夏季 (%) Summer	秋季 (%) Autumn
研究区域Study area	0.001 0	0.001 4	0.001 0	0.000 4	0.41	0.92	0.37	0.15
高寒草甸Alpine meadow	0.001 3	0.001 9	0.001 5	0.000 4	0.39	1.10	0.41	0.11
高寒草原Alpine steppe	0.000 7	0.000 9	0.000 7	0.000 4	0.41	0.72	0.39	0.22
温性草原 Temperate steppe	0.001 0	0.001 5	0.000 9	0.000 8	0.42	0.92	0.35	0.31

Figure 2: NDVI changing magnitude and changing ratio in different seasons during 1982-1999

• Introduction of New Algorithms:

The widespread application of deep learning algorithms has significantly improved the precision of NDVI analysis. For example, Zhang Peng et al. (2022) used convolutional neural networks (CNN) to classify NDVI data from degraded grassland areas, achieving an accuracy rate of over 95%, providing technical support for grassland management[10]. However, it is necessary to analyze the specific context, as more complex crop models do not necessarily yield more accurate yield estimates than simpler regression models.

International Research Status:

Over the past two decades, multispectral remote sensing data has become the primary data source for agricultural analysis[11]. International research has focused on multi-source data fusion, complex system modeling, and the prediction of global ecological changes:

- Breakthroughs in Multi-Source Data Fusion:

International scholars have combined LiDAR data, hyperspectral imaging, and meteorological observations to enhance the spatial-temporal resolution of NDVI calculations. For example, Feng et al. (2020) studied urban vegetation changes and validated model predictions using drone-based remote sensing data, demonstrating high prediction accuracy.

- Advances and Innovations in Image Data and Time Series Processing：

Before conducting precise analysis and prediction of all data, it is essential to filter and reconstruct the relevant datasets. On the existing foundation, it becomes necessary to remove outliers and supplement the gaps with valid data. Medium spatial-resolution NDVI time series data with long temporal coverage are particularly significant in this context, and Landsat, which provides publicly available data at a 30-meter resolution, has greatly facilitated research efforts [12].

However, the benefits come with limitations. The 16-day revisit cycle of the Earth Resources Satellites, cloud contamination, and sensor imaging issues in Landsat 7 can severely impact results[13].

To address these challenges, the reconstruction methods mentioned earlier have shown their effectiveness. For instance, Cao et al. applied a newly developed ARRC

algorithm to process reference images, achieving significant removal of thick cloud contamination[14]. Multi-scale data fusion between high spatial and high temporal resolution imagery offers a promising solution for reconstructing image sequences, filling gaps caused by cloud contamination, and predicting missing data [15].

Cuizhen Wang successfully employed a hybrid approach, integrating data from Landsat, AVHRR GIMMS, and climate reanalysis datasets at different spatial scales to reconstruct NDVI data [16]. However, challenges persist in integrating spatial, temporal, long-term series, and multi-sensor information, which limit the feasibility of precision analysis and prediction.

On the other hand, Haitao Lv addressed the issue using gap-filling or spatial-similar pixel-filling methods, but the results were far inferior to fitting-based methods[17]. Consequently, Peng Qin, Huabing Huang, and others proposed ReCoff, a deep learning spatiotemporal fusion method with residual constraints, which achieved remarkable success in addressing NDVI time series gaps[18]. This demonstrates that modern time series reconstruction is increasingly intertwined with deep learning techniques.

- Optimization of Time Series Analysis Models:

Deep learning models, such as LSTM and Transformer, have been widely used for NDVI time series prediction. Saygin Abdikan collected in-situ crop height measurements during the data collection period and used regression methods such as SLR, MLR, ANN, XGBoost, and CNN for crop height estimation [11].

Harmonic Analysis of Time Series (HANTS) is a sequence reconstruction method based on harmonic analysis, one of the oldest methods for handling satellite observation time series affected by atmospheric conditions or snow contamination [11].

- Global Research and Regional Characteristics:

Kingsley Kanjin and Bhuiyan Monwar Alam used supervised classification of Landsat images to examine land cover changes in the Sundarbans mangrove forest from 1973 to 2023 [19]. Remus Prăvălie studied forest land in Romania from 1987 to 2018, based on NDVI, to analyze forest vegetation ecological dynamics and their relationship with climate change [20], uncovering the connection between forest vegetation and climate.

Research Development Trends in Domestic and International Contexts:

The future development of NDVI-related research is expected to focus on the following aspects:

1. Multi-Scale Spatial-Temporal Fusion Analysis:

Both domestic and international scholars are increasingly attempting to integrate satellite data with ground-based observation data to improve the explanatory power of vegetation dynamics. For example, the application of China's high-resolution satellite data provides new opportunities for refining regional vegetation studies, while multi-satellite joint monitoring technologies have become mainstream internationally[21].

Marianna Belgiu employed more sophisticated deep learning methods to process data[22] and highlighted the need for further research to account for temporal changes occurring during the observation period when predicting spectral reflectance values at fine spatial and temporal scales, such as using deep learning approaches[23].

Additionally, the introduction of new models and data fusion techniques can reduce interference from unknown factors such as cloud cover in the results.

3. Model Optimization and Intelligence:

Building on deep learning, reinforcement learning (RL) and transfer learning (TL) are gradually being introduced into ecological models to enhance their adaptability and generalization capabilities[24]. Deep learning (DL) techniques have received attention due to their scalability and active learning capability. In comparison to traditional ML techniques[25], DL models can self-learn salient features from raw data to recognize object types. Optimizing the model's predictive ability, while reinforcement learning, through a reward-punishment mechanism, improves decision-making processes, thus optimizing ecological environmental prediction models based on remote sensing data[26].

For instance, in ecological restoration, reinforcement learning can continuously optimize land management plans by simulating the effects of different restoration strategies[27]. Moreover, transfer learning techniques can effectively transfer knowledge from well-studied regions to areas with sparse or highly variable data, improving the prediction accuracy and efficiency of models in new environments[28]. Combining these

two learning methods with traditional deep learning models presents new opportunities for intelligent ecological monitoring and prediction.

4.Cross-Domain Data Integration:

Integrating socio-economic data, land use data, and climate prediction information to construct a multi-dimensional ecological evaluation system has become an important direction in recent ecological research[29].

With the development of remote sensing technologies, researchers are increasingly integrating additional cross-domain data sources, such as climate model predictions[30] and socio-economic activity data, into comprehensive assessments. For example, Jakeman et al. (2017) combined climate models with NDVI predictions[31],using data fusion to reveal the joint impact of human activities and climate change on African savanna ecosystems, and proposed more comprehensive ecological restoration and conservation strategies.

By integrating multiple data sources[32], a better understanding of the complexity of ecosystems and the impact of human activities and climate change on ecological services can be achieved.

Evaluation of Regional Ecosystem Services:

In ecological research, there has been an increasing focus on how to quantitatively evaluate ecosystem services[33], with NDVI being widely used as a quantitative indicator for ecosystem services. Carbon gain dynamics can be readily characterized from vegetation spectral indices strongly associated with the (spatio-temporal) patterns of primary production, such as the Normalized Difference Vegetation Index (NDVI)[34].

NDVI not only reflects vegetation cover but also reveals changes in vegetation health and productivity[35]. Enrica Nestola, Carlo Calfapietra and colleagues[36]investigated the seasonal productivity of grasslands at Mattheis Ranch in Alberta, Canada, by utilizing various NDVI derivatives.

Furthermore, NDVI plays a significant role in assessing biodiversity conservation and soil conservation effects [37], providing data support for ecosystem managers to optimize regional ecosystem service protection strategies.

3 Research Methodology

3.1 Study Area

The study area is located at the Troodos Mountains of Cyprus (Figure **3**). The Troodos Mountains are situated in the eastern Mediterranean, spanning the southern part of Cyprus, with geographic coordinates approximately between 34°40′N to 35°20′N and 32°40′E to 33°10′E.

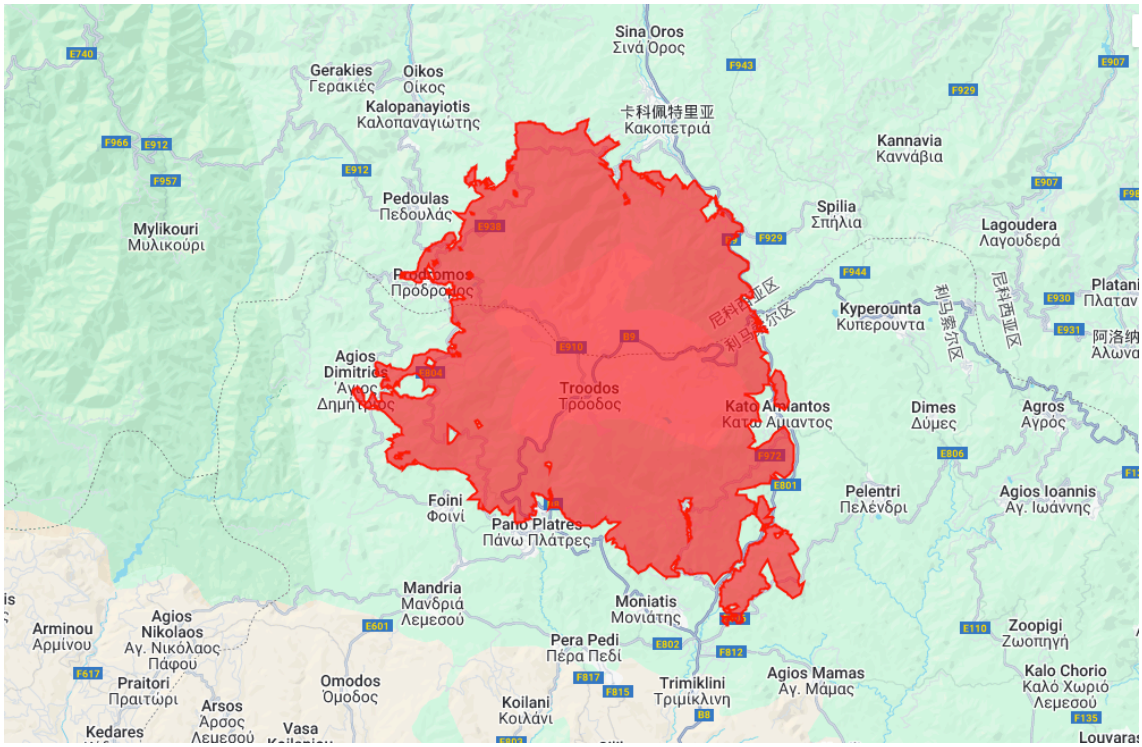

Figure 3: The range of areas to be studied

This area features complex topography and diverse vegetation types, making it a typical mountainous ecosystem. The main vegetation types in the Troodos Mountains are highly diverse and include the following:

1. Low-altitude areas are dominated by the Cyprus pine (Pinus brutia), reaching up to 1200 meters, with forests extending as high as 1500 meters in the southern parts.
2. Near rivers, dense vegetation of oriental plane trees (Platanus orientalis), alder trees (Alnus orientalis), and myrtle trees (Myrtus communis) is found, adding to the region's diversity and providing ideal refuges for wildlife.

3. In lower altitude areas, wild olive trees (Olea europaea) are found up to 1000 meters, while arbutus trees (Arbutus andrachne), with their striking color changes, extend from 600 meters to 1500 meters. Additionally, the sumac (Rhus coriaria) and the endemic Lagoia oak (Quercus alnifolia) are present from 600 meters up to 1650 meters.
4. In the higher altitude areas of the forest, the black pine tree forest (Pinus nigra) extends up to the Chionistra area, coexisting with psychotropic shrubs such as junipers (Juniperus foetidissima), wild apple trees (Sorbus aria), wild quince trees (Cotoneaster racemiflorus), barberry (Berberis cretica), the endemic raisin bush (Genista sphacelata ssp. crudelis), and others.

Vegetation cover in the Troodos Mountains is influenced by various factors, such as temperature, precipitation, soil types, and elevation. The mountain ecosystem in this region is highly vulnerable[38], especially under the dual pressures of climate change and human activities, which have led to significant vegetation changes.

In recent decades, the vegetation cover of the Troodos Mountains has undergone varying degrees of change, primarily manifested in forest degradation, grassland desertification, and soil erosion. The remote sensing characteristics of different vegetation types exhibit notable spatial and temporal heterogeneity, complicating vegetation monitoring and change analysis.

3.2 Data Processing

In this study, to construct a time-series dataset spanning from 1985 to 2020, we systematically selected and processed Landsat series satellite imagery using the Google Earth Engine (GEE) platform combined with Python scripts. The core objective of this step was to extract remote sensing imagery that met the research criteria for the Troodos Mountains region, conduct necessary preprocessing and filtering, and ultimately generate a time-series dataset stored as a CSV file. This provides foundational data for subsequent analyses. All the steps of this study are shown in the Figure 4 below.

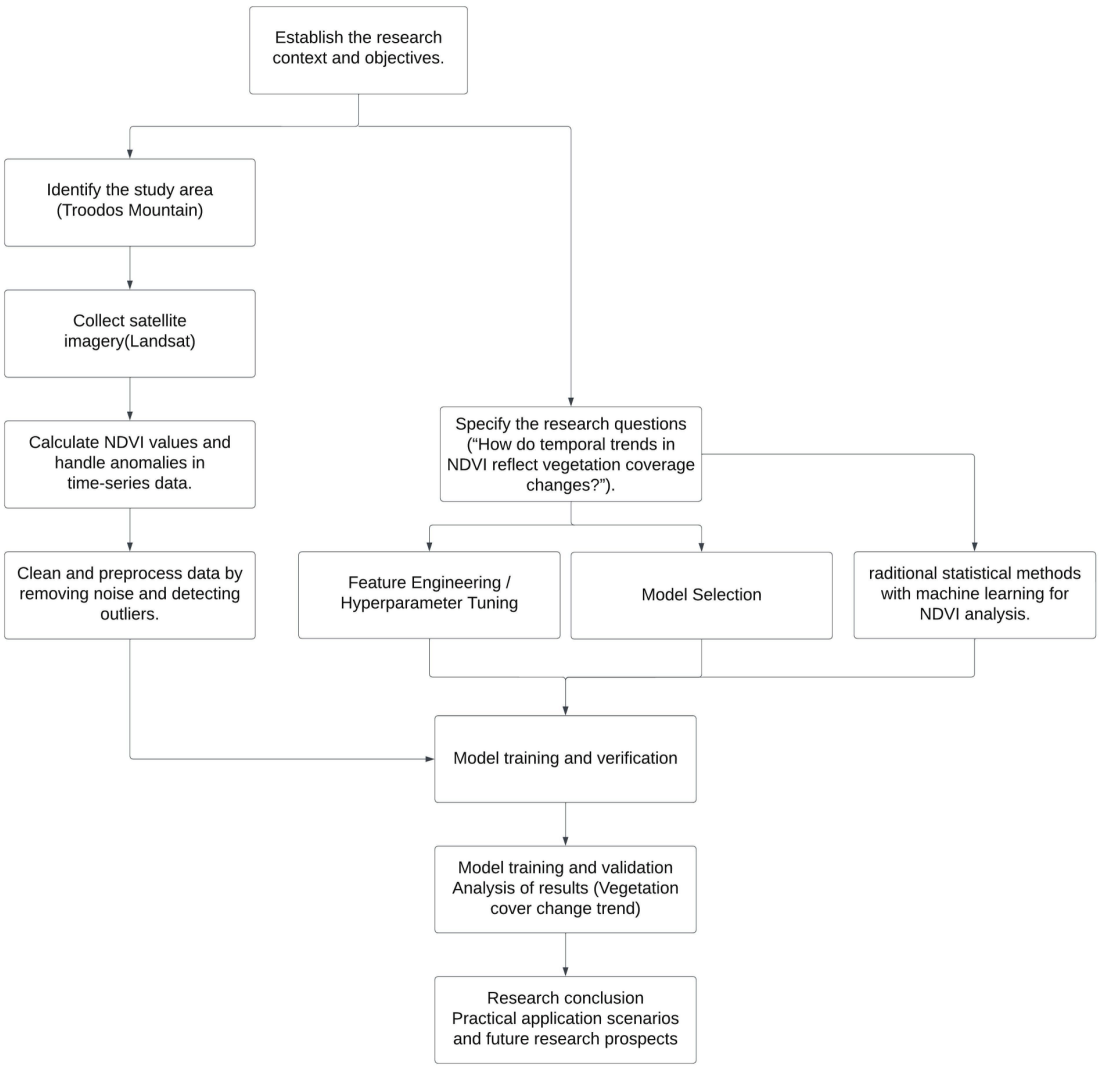

Figure 4: All processes

3.2.1 Data source and image screening

In this study, in order to construct a time series dataset covering the period from 1985 to 2020, we systematically selected and processed Landsat series satellite images by integrating the Google Earth Engine (GEE) platform and Python scripts. The core objective of this step was to extract remote sensing images that met the research standards for the Troodos Mountain region, conduct necessary preprocessing and filtering, and ultimately generate a time series data set stored as a CSV file. This provided the basic data for subsequent analyses.

We utilized three primary Landsat data sources:

1. **Landsat 5 (LANDSAT/LT05/C02/T1_L2)**: Medium-resolution remote sensing data covering the period from 1984 to 2012.
2. **Landsat 7 (LANDSAT/LE07/C02/T1_L2)**: Available from 1999 onwards, although data collected after 2003 suffer from partial striping issues due to the failure of the Scan Line Corrector (SLC).
3. **Landsat 8 (LANDSAT/LC08/C02/T1_L2)**: Provides higher-quality imagery, available from 2013 to the present.

By using the image set interface of GEE, we filtered the images in these three datasets by specifying the time range (from January 1, 1985 to December 31, 2020), and spatially limited the image area to ensure that only the Troodos Mountain region was covered.

To further enhance the data quality, we have implemented the cloud masking function (mask Clouds), which can remove the pixels contaminated by clouds. This function reads the Quality Assessment (QA_PIXEL) band and applies logical operations to mask the cloud pixels. This step significantly improves the clarity and usability of the images and ensures higher accuracy of the analysis results.

3.2.2 Image collection merging and timestamp extraction

After screening and preprocessing, we merged the image collections from three Landsat sources to form a comprehensive dataset covering the entire time period. We used the GEE interface to extract the timestamps of each image and converted them into a readable date format (YYYY-MM-DD) for visualization and verification of the time distribution.

Finally, we recorded the total number of all available images during the research period and stored the date information of each image in a time series CSV file. This file includes all available image dates and is organized in a clear structure for subsequent time series analysis and cross-validation. The key steps are as follows:

- **Generate time-series data**: Store the dates of all eligible images in a list.
- **Export to CSV**: Using Python's csv module, export the data to a CSV file named available_image_dates.csv, including a header labeled "Date" for clarity.

This process established a comprehensive dataset of Landsat imagery for the Troodos Mountains from 1985 to 2020. It provided a solid foundation for NDVI (Normalized Difference Vegetation Index) calculations and time-series modeling. The cloud masking and standardized data formatting significantly enhanced data quality and usability. The resulting time-series file not only supports this research but also serves as a reusable dataset for other ecological and environmental studies based on Landsat data.

3.2.3 NDVI calculation at random study sites

Based on the preprocessed data and Troodos Mountain region information available through GEE, we randomly generated 100 study points (Figure 5). Using these points, we extracted remote sensing image data at each time step and calculated the corresponding NDVI values. For each study point, we computed NDVI values based on data from each time step, ultimately creating a dataset of NDVI time-series data for 100 study points. The goal was to generate sufficient time-series data to support subsequent vegetation dynamics analysis and model validation.

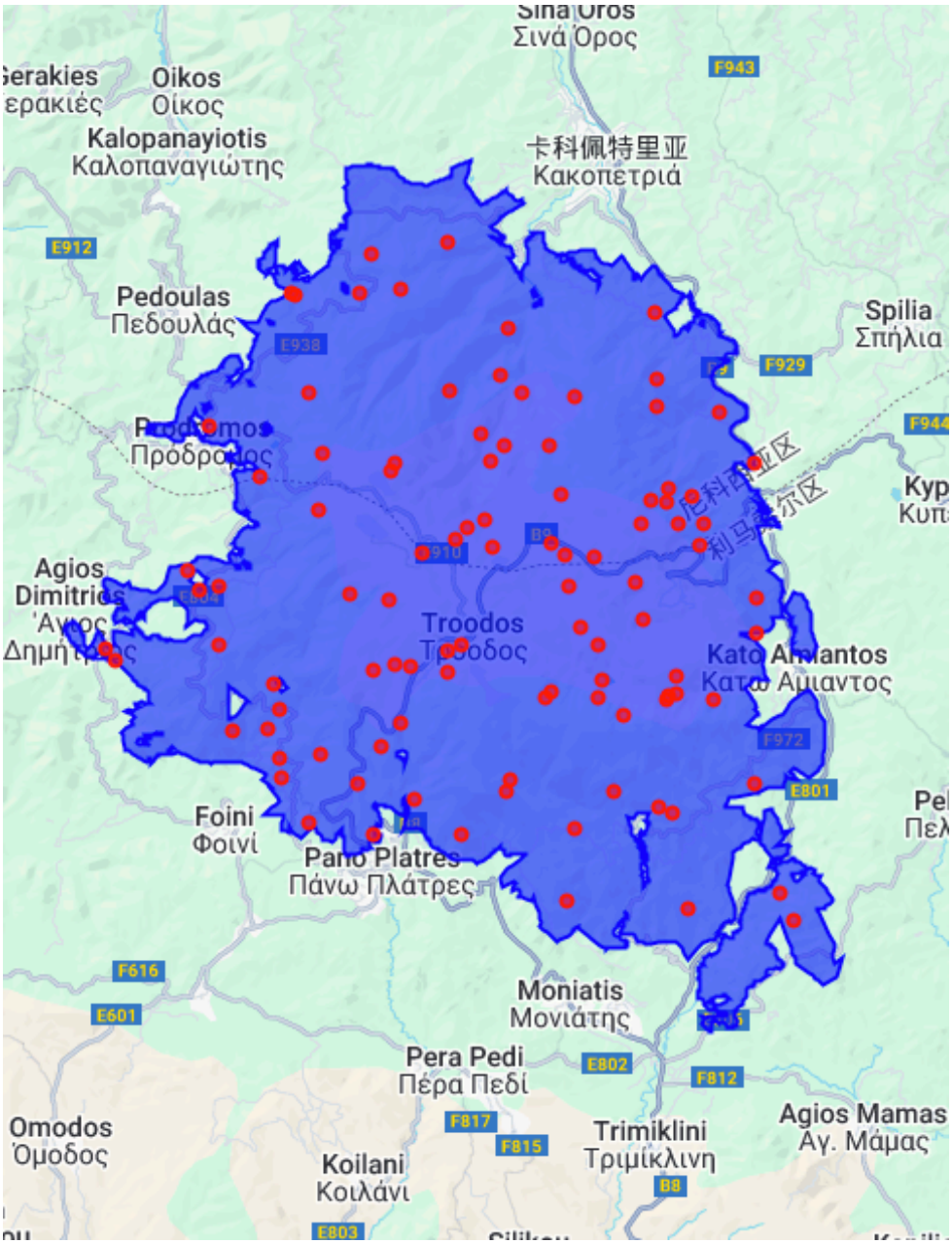

Figure 5: The distribution of random points generated in the study area

3.2.4 Analysis and selection of outlier processing results

In the outlier detection phase, we applied several methods, including 3Sigma, DBSCAN, IQR, LOF, and LOF_Withoutmissingpoint, to analyze the processed time-series data. The performance of these methods varied, as detailed below:

1. **3Sigma**

   The 3Sigma cluster scheduling system uses job runtime histories in a new way[39]. The core idea of this method assumes that the data follows a normal distribution, using the mean (μ) and standard deviation (σ) to identify outliers that deviate significantly from the normal range. The acceptable value range is defined as: [μ − 3σ, μ + 3σ], as shown in Formula 2. Data points falling outside this interval are classified as anomalies.

   The primary advantage of this method is its computational simplicity, making it highly effective for processing approximately normally distributed data. However, it is prone to misclassification and should be applied with caution.

   The judgment criteria for outliers:

$$x_i \text{ is an outlier} \Leftrightarrow x_i \notin [\mu - 3\sigma, \mu + 3\sigma] \quad (3.1)$$

2. **DBSCAN**

   DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies outliers based on density parameters. DBSCAN algorithm has good clustering results in the application, which is a typical representative of density algorithm[40]. The core idea is: "Clusters are high-density regions, while anomalies are isolated points in low-density areas." However, this method performs poorly in data with significant density variations, such as highly dynamic time series. If the data contains substantial fluctuations, it may fail to reliably detect anomalies.

3. **IQR**

   The IQR (Interquartile Range) method detects outliers by identifying extreme values outside the interquartile range. While it effectively removes clear outliers, it also tends to remove many data points during periods of high variability, potentially reducing representativeness. Careful threshold setting is essential to avoid excessive data loss.

4. **LOF**

LOF (Local Outliers Factor) algorithm is a very classic anomaly detection algorithm. In order to detect the outliers more accurately, avoid too much testing error[41]. In this paper, we elaborate on this method, which begins by defining a local neighborhood for each point based on its k-nearest neighbors (k-NN) using distance metrics such as Euclidean distance.

$$\textbf{k-distance}(\boldsymbol{p}) = \textbf{distance to its } \boldsymbol{k^{th}}\textbf{nearest neighbor} \quad \textbf{(3.2)}$$

Reachability distance between and:

$$\textbf{reach-dist}_{\boldsymbol{k}}(\boldsymbol{p}, \boldsymbol{o}) = \boldsymbol{max}\{\textbf{k-distance}(\boldsymbol{o}), \boldsymbol{d}(\boldsymbol{p}, \boldsymbol{o})\} \quad \textbf{(3.3}\boldsymbol{)}$$

where d(p,o) is the Euclidean distance.

Local reachability density (LRD) of p:

$$\textbf{LRD}_{\boldsymbol{k}}(\boldsymbol{p}) = \frac{|\boldsymbol{N_k(p)}|}{\sum_{\boldsymbol{o}\in \boldsymbol{N_k(p)}} \textbf{reach-dist}_{\boldsymbol{k}}(\boldsymbol{p},\boldsymbol{o})} \quad \textbf{(3.4)}$$

Nk(p) denotes the set of k-nearest neighbors of p.

LOF score:

$$\textbf{LOF}_{\boldsymbol{k}}(\boldsymbol{p}) = \frac{\boldsymbol{1}}{|\boldsymbol{N_k(p)}|}\sum_{\boldsymbol{o}\in \boldsymbol{N_k(p)}} \frac{\textbf{LRD}_{\boldsymbol{k}}(\boldsymbol{o})}{\textbf{LRD}_{\boldsymbol{k}}(\boldsymbol{p})} \quad \textbf{(3.5)}$$

- LOF ≈ 1: Similar density to neighbors (inlier)
- LOF ≫ 1: Lower density than neighbors (outlier)

5. **LOF_Withoutmissingpoint**

This is a variation of the LOF algorithm where we specifically remove missing values and then filter outliers, this variation of LOF specifically addresses missing values and exhibited the best performance in balancing anomaly removal and data integrity. It accurately detected noise while preserving data trends, making it the most effective method in this study.

6. **Quantiles**

Quantiles divide a dataset into equal-sized intervals, helping to identify and analyze the distribution of data by pinpointing its spread and potential outliers. The quantiles

method identifies outliers using data quartile information. This method is particularly effective for non-normally distributed data and performs better than the 3Sigma method under diverse distributions. However, the situations of different distributions are often difficult to be properly controlled in time series.

3.2.5 NDVI Time Series Interpolation Methods

After addressing outliers and missing values, the next step is to interpolate the time series data to fill the missing NDVI values. Since remote sensing data often exhibits complex spatiotemporal variations, selecting an appropriate interpolation method is crucial. Despite our attempts to adopt many different approaches, in this study we merely introduce the two methods that have shown relatively better results:

1. **Kalman Interpolation**

   The filling process of the Kalman method mainly consists of two steps. The first is the prediction step: based on the control input and the state of the previous moment, calculations are made. Then, data prediction for the next time point is carried out.

   The second is the update step: the current measurement value is obtained, and the error (residual) is calculated with the predicted value. According to the residual, the current state estimation is adjusted. At the same time, the weighted average (Kalman gain) is calculated based on the calculation to combine the predicted and measurement results, obtaining a more accurate state estimation. In summary, the key of Kalman filtering lies in the continuous iterative optimization to minimize the error (covariance), thereby effectively suppressing the noise influence in the dynamic changing environment and providing accurate system state estimation[42].

2. **Spline Interpolation**

   Spline Interpolation fits smooth polynomial functions between data points for interpolation. It provides excellent fitting results for time series data with smooth variations and is particularly suitable for regions were vegetation cover changes gradually.

3.2.6 NDVI Time Series Fitting Methods

We used multiple fitting methodology, mainly including Savitzky-Golay (SG), Whittaker, Median Vegetation Index (MVI) filter, and Double Logistic[43] Different methodology perform best for different time series, with each method's advantages depending on the data's specific characteristics.

1. **Savitzky-Golay Smoothing**

   The Savitzky-Golay method is a polynomial smoothing technique that reduces noise while preserving underlying data trends[44].The core principle uses a moving window with least-squares fitting of 2nd/3rd-order polynomials, continuously smoothing the central data points throughout the time series[45].This experiment applied double SG filtering to reduce high-frequency noise and refine data, significantly improving fitting results[46].

2. **Whittaker Smoothing**

   Whittaker smoothing minimizes the roughness of the fitted curve by applying penalized least squares. It balances the fit by penalizing large changes in slope or curvature, ensuring a smooth curve that retains the underlying trend[47] This method is particularly useful for dealing with noisy, spatiotemporal data like NDVI time series[48]where irregular fluctuations or outliers may exist.

3. **Median Vegetation Index (MVI)**

   The Multi-temporal Vegetation Index (MVI) mitigates seasonal fluctuations and short-term noise by computing the median vegetation index (e.g., NDVI) from multi-temporal remote sensing data (captured at different time points)[49]. This approach provides more stable vegetation health monitoring results, effectively smoothing out seasonal variations.

4. **Double Logistic**

   The double logistic function consists of two distinct sigmoidal growth curves, each representing a different phase. The first curve typically depicts a rapid growth stage, while the second reflects a gradual slowdown in growth rate, eventually approaching saturation[50]. By adjusting its parameters (e.g., maximum value, growth rate, inflection points), the double logistic function can flexibly fit time-series data. Through this two-phase modeling, it effectively captures the S-shaped growth patterns observed in many real-world processes.

3.2.7 Time Series Prediction Methods

The five types of time series prediction methods to be introduced next cover both traditional statistical models and cutting-edge machine learning, large time series models, etc.

1. **XGBoost (eXtreme Gradient Boosting)**

   XGBoost is an ensemble learning method that enhances model accuracy by iteratively constructing multiple decision trees to progressively minimize prediction errors. As a boosting algorithm, XGBoost combines the strengths of various base learners to achieve superior predictive performance compared to any individual constituent algorithm, demonstrating exceptional results across numerous domains[51].

   Moreover, XGBoost incorporates a regularization term into its objective function, which enhances the generalization capability of individual trees and reduces model complexity. In essence, XGBoost has garnered significant attention from researchers due to its computational efficiency[52]exceptional classification performance, and flexibility in supporting user-defined loss functions.

2. **LightGBM (Light Gradient Boosting Machine)**

   LightGBM, proposed by Microsoft in 2017, is a gradient boosting framework based on GBDT (Gradient Boosting Decision Trees). Like other boosting algorithms, GBDT combines multiple weak learners to form a strong learner[53]. LightGBM is a lightweight gradient boosting framework. Traditional GBDT algorithms often spend a significant amount of computation time on decision tree construction, which requires finding the optimal split points. The common approach involves sorting feature values and enumerating all possible split points, which is time-consuming and memory intensive.

   LightGBM uses an improved histogram-based algorithm, which discretizes continuous feature values into k bins and selects split points from these k values[54]. This not only reduces computation but also has a regularization effect, helping to prevent overfitting. Compared to traditional GBDT algorithms, LightGBM performs better in terms of training efficiency, memory usage, and handling large-scale data[55]. It also supports direct input of categorical features, avoiding the overhead

of one-hot encoding. In time series forecasting, it often relies on feature engineering techniques such as sliding windows.

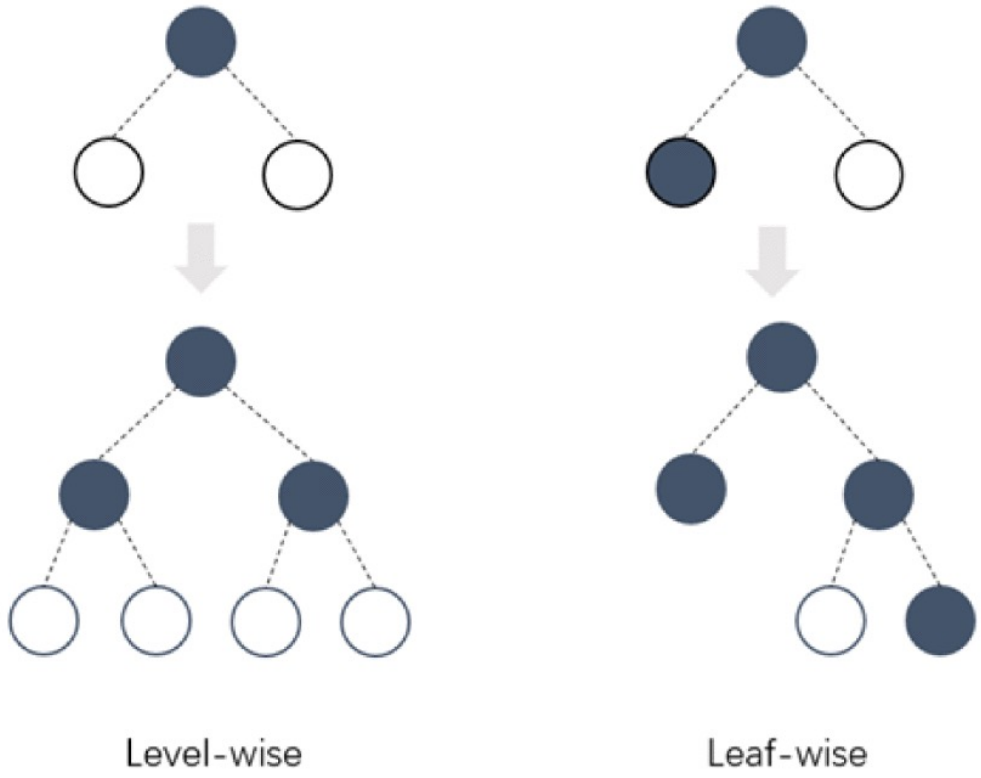

Figure 6: The working mode of LGBM

3. **N-HiTS (Neural Hierarchical Interpolation for Time Series)**

   N-HiTS is a deep learning model specifically designed for time series forecasting. It models sequence features at different temporal scales through a hierarchical structure and uses an interpolation mechanism to generate predictions[56]. The model can automatically identify patterns such as trends and seasonality, making it particularly suitable for multi-scale and non-stationary time series, such as NDVI data, which exhibit long-term changes and seasonal fluctuations.

4. **TimeGPT (Time Series Generative Pre-Trained Transformer)**

   TimeGPT is a Transformer-based pre-trained time series generation model, inspired by the pretraining paradigm of language models. It is trained on large-scale time series data, giving it strong generalization and transfer capabilities. Operating in a univariate channel setup, it is specifically designed for detection and forecasting tasks. As the largest foundation model in the time series domain based on Transformer architecture, TimeGPT has been pre-trained on over 100 billion data points.

Although it claims to have collected the largest time series repository from public sources, TimeGPT has not publicly disclosed the details of its repository or the data used in training[57].

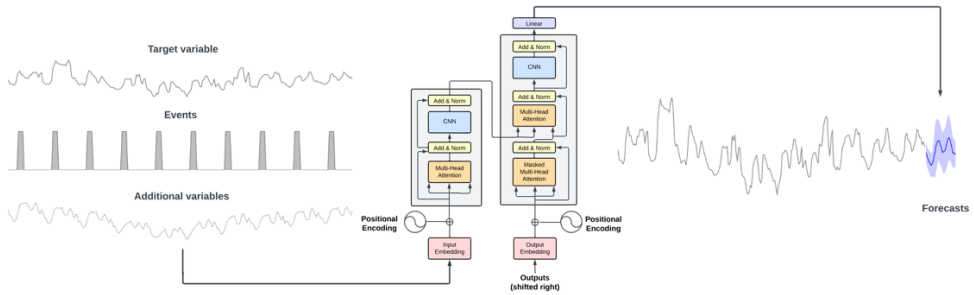

Figure 7: The prediction method of TimeGPT

5. **LongTimeGPT (Long Time Series Generative Pre-Trained Transformer)**

LongTimeGPT is an upgraded version of TimeGPT, designed to handle ultra-long time series data. By optimizing attention mechanisms and related techniques, it addresses the efficiency and memory bottlenecks commonly encountered in long-sequence modeling. This enables it to capture long-term trends and structural changes, making it well-suited for modeling and forecasting tasks involving long-term data, such as climate change and environmental monitoring.

3.2.8 Metrics for evaluating the results

This section evaluates system stability using the following metrics: Mean Absolute Percentage Error (MAPE), Root Mean Square Deviation (RMSE), Nash-Sutcliffe Efficiency (NSE), and Mean Absolute Error (MAE).

$$\mathbf{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{3.6}$$

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3.7}$$

$$\mathbf{NSE} = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{3.8}$$

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{3.9}$$

MAPE represents the percentage of prediction error relative to the actual value and penalizes both underestimation and overestimation. RMSE emphasizes larger errors by squaring the deviations, which amplifies the impact of outliers. NSE measures the goodness of fit relative to the mean of the observed data, but it is highly sensitive to extreme values. MAE, on the other hand, is more intuitive, as it calculates the average magnitude of errors using absolute values, providing a straightforward measure of how much the predictions deviate from the observations on average.

4 Results and Discussion

In this chapter, the study further clarifies its research objectives by specifically analyzing the effects of outlier handling and time series fitting. First, it compares different outlier filtering methods and evaluates their performance in reconstructing NDVI data. Subsequently, the imputed and fitted results from each method are systematically compared to determine the optimal data processing strategy, thereby providing a solid data foundation for the development of subsequent forecasting models.

In the time series analysis section, this study develops several forecasting and estimation models based on NDVI data derived from satellite remote sensing. Considering that the mean is easily influenced by outliers, the median is uniformly adopted for NDVI processing to obtain a more robust time series. During the data preprocessing stage, observations affected by cloud cover were rigorously removed—for example, by filtering out cloud-masked pixels—to minimize the impact of atmospheric and remote sensing conditions on NDVI values and to ensure the reliability of data quality.

Finally, based on the cleaned, high-quality NDVI time series, an ensemble forecasting model is designed to estimate vegetation dynamics for the coming years. It is important to note that this model uses only the NDVI sequence itself as the input variable, without incorporating any exogenous factors. Therefore, its predictive capability relies entirely on the temporal characteristics of NDVI. This also means the study focuses on evaluating how well different models can capture and fit the temporal patterns of NDVI, rather than analyzing external driving factors.

4.1 Analysis and selection of outlier processing results

As shown in Figure 9, DBSCAN demonstrates limited effectiveness in filtering outliers; in this study, it failed to remove many abnormal data points, indicating that it performs poorly in low-density or uniformly distributed datasets. Therefore, DBSCAN may not be suitable for the characteristics of this dataset. In contrast, the IQR method (Figure 10) applies overly aggressive filtering, removing a substantial portion of the data.

Using LOF directly without removing missing values results in the incorrect elimination of some important data points, revealing clear shortcomings in this approach. The 3Sigma method identifies outliers as those falling beyond three standard deviations from the

historical mean. While this method has low sensitivity to small fluctuations, it can preserve subtle anomalies (Figure 8). Under stable data conditions, it is acceptable for retaining non-significant variations.

The LOF method performs as expected. As seen in Figure 11, its high sensitivity to noise can lead to the misclassification of normal fluctuations as outliers. However, after removing missing values, the LOF_Withoutmissingpoint approach (Figure 12) clearly demonstrates strong filtering performance, offering a more accurate representation of the underlying NDVI patterns.

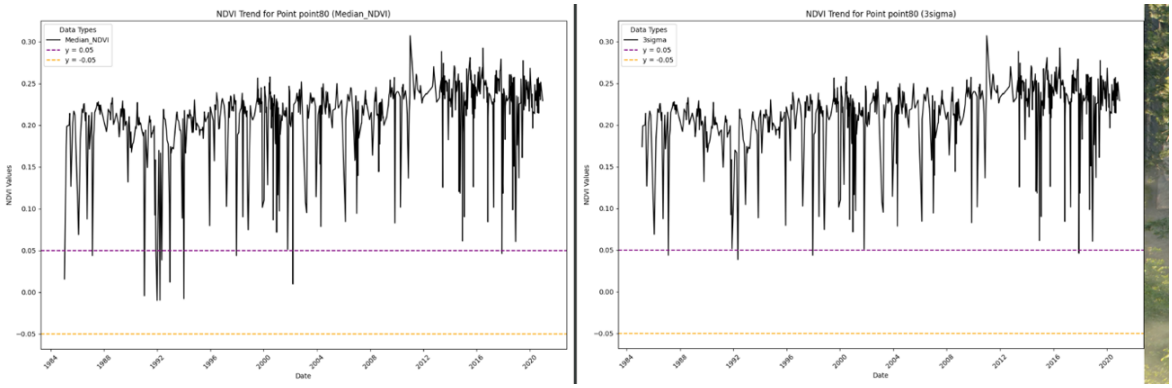

Figure 8: The effect of outlier removal using the 3Sigma method

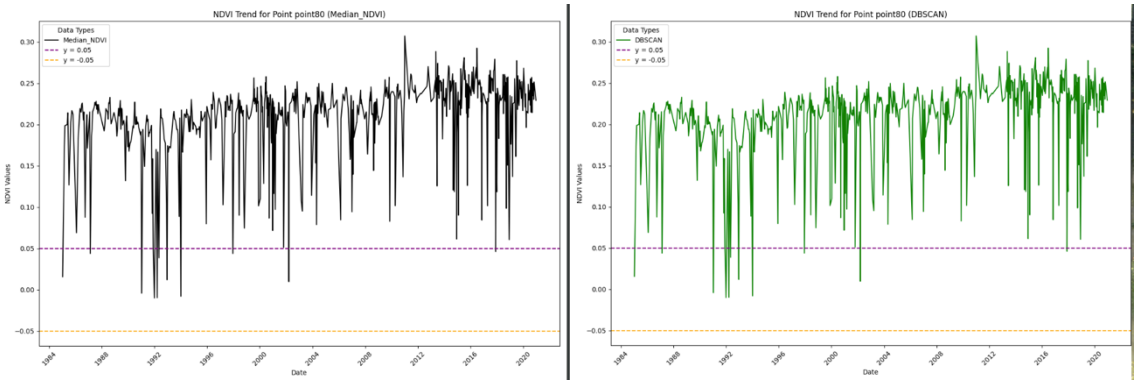

Figure 9: The effect of outlier removal using the DBSCAN method.

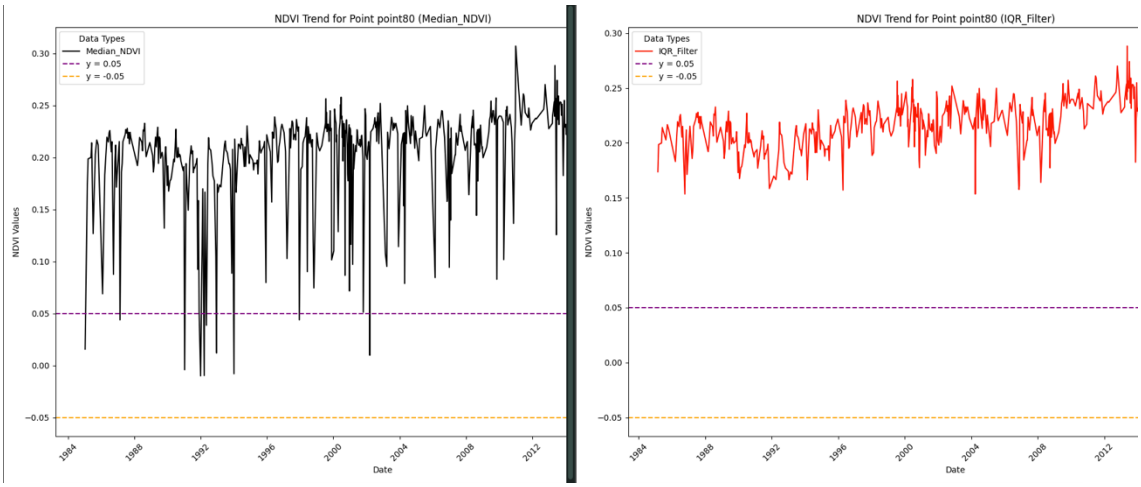

Figure 10: The effect of outlier removal using the IQR method.

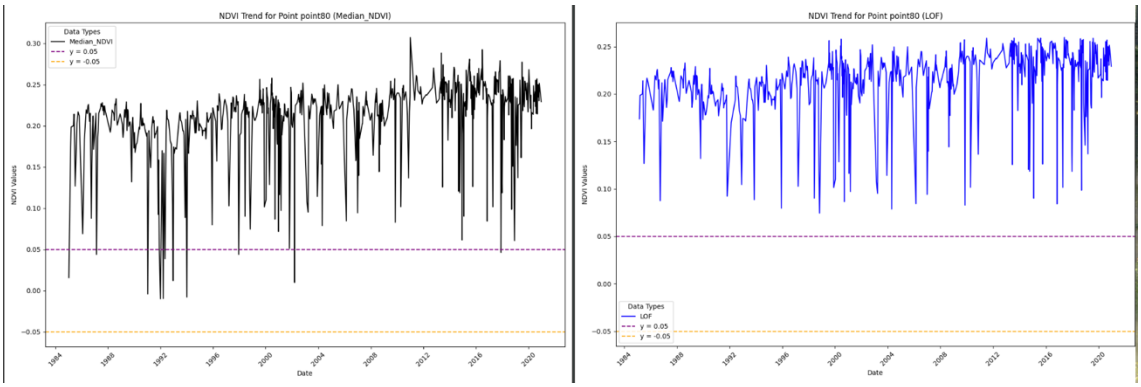

Figure 11: The effect of outlier removal using the LOF method.

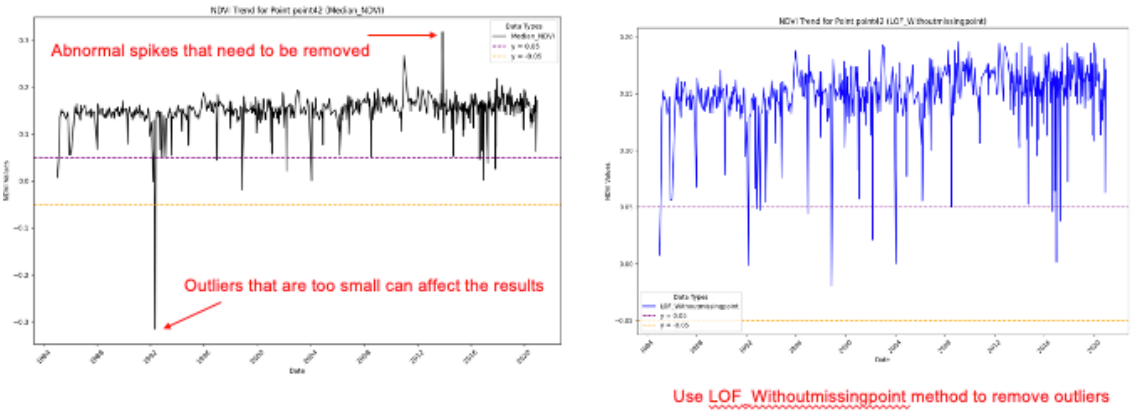

Figure 12: The effect of outlier removal using the LOF_Withoutmissingpoint method

In summary, while the LOF_Withoutmissingpoint method demonstrated the best overall performance, other methodology, such as 3Sigma and the quantile method, also produced acceptable results. Specifically, detailed results and summaries can be seen in Table 1.

Table 1. Method introduction

Method	Result Description	Detect the number of outliers
3Sigma	Suitable for scenarios with small data fluctuations and where outliers are not particularly severe	402 (0.61%)
DBSCAN	best suited for data with large density differences	102(0.15%)
IQR	Suitable for cases where outliers are extremely obvious, but caution is needed to avoid excessive data removal	4433(6.71%)
LOF	Data sets that fit with few missing values are affected by null values	6636(10.51%)
LOF_Withoutmissingpoint	Best-performing method	2023(3.06%)
Quantiles	Overprocessing data	2738(4.15%)

In summary, while the LOF_Withoutmissingpoint method demonstrated the best overall performance, other methods, such as 3Sigma and the quantile method, also produced acceptable results. Specifically detailed results and summaries can be seen in Table 1.

Based on this analysis, we selected the LOF_Withoutmissingpoint method as the primary outlier detection strategy, retaining the results of the 3Sigma and quantile methodology for flexible use in different scenarios to ensure dataset quality and integrity.

4.2 Time Series Reconstruction and Smoothing

In the following section, the final processing results are presented. To demonstrate the effectiveness of the fitting methodology used, a random point from the study area was selected, and various processing techniques were applied to its NDVI time series. The images (Figure 13) after fitting the random point are displayed, highlighting the results of different smoothing and fitting methodology. Comparing different smoothing and fitting methodology, the following conclusions can be drawn:

1. **MVI Spline and Quadratic SG Filtering with Parameter 5**

   Both methodologies have obvious flaws in practical applications, with minimal smoothing effect. The data still shows significant fluctuations and does not effectively mitigate short-term fluctuations. These methodologies have weak fitting performance and are not suitable for smoothing current data.

2. **ARMD3 and Single-pass SG Filtering (Underfitting)**

   Both methodologies suffer from underfitting, failing to capture the overall trend of the data adequately. The single-pass SG filtering method shows many peaks in the fitting curve, indicating that this methodology may be too simplistic and fail to account for the complexity of the time series.

3. **Quadratic SG Filtering with Parameter 30 (Overfitting)**

   Although this method follows the data well, it clearly suffers from overfitting, fitting too much of the data, including information that should not be fitted, leading to the removal of too much data. Overfitting reduces the model's ability to generalize to new data, so parameter selection needs to be more cautious.

4. Best Method: Whittaker and Quadratic SG Filtering with Parameter 15

   This methodology performs well in both smoothing and trend fitting. They effectively reduce short-term anomalies while preserving the long-term trend. The Whittaker method performs well in noise suppression, while the quadratic SG filtering method with parameter 15 achieves good smoothing without overfitting. Therefore, these two methodologies are optimal for the current dataset.

   In conclusion, MVI Spline and SG filtering with parameter 5 failed to effectively filter noise, ARMD3 and single-pass SG filtering suffered from underfitting, and SG filtering with parameter 30 had overfitting issues. Overall, Whittaker filtering and SG filtering with parameter 15 strike the best balance between signal retention and smoothing, making them the most suitable for current NDVI data processing.

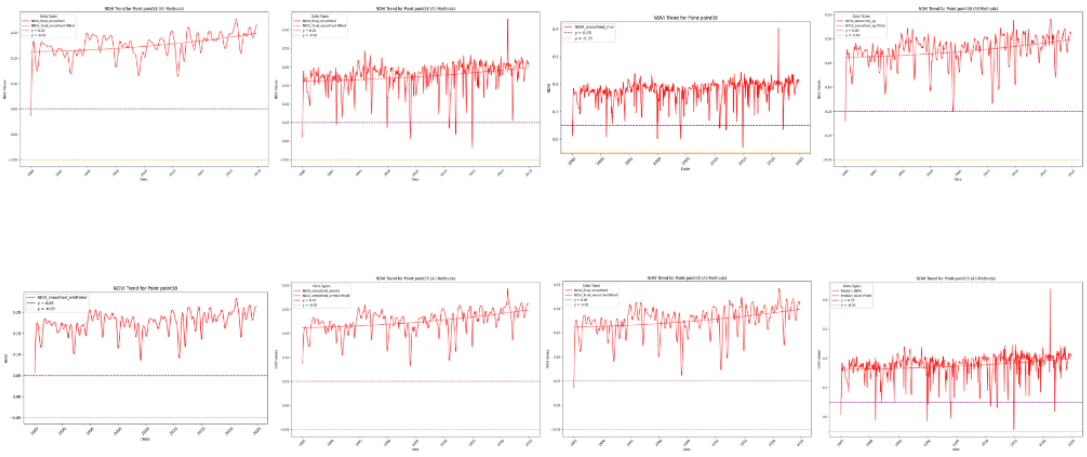

Figure 13. The effect of all fitting methodology

To quantitatively evaluate the performance of different smoothing methodology, we calculated key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These metrics are used to assess the fitting performance of each method on the time series data and evaluate their effectiveness in noise reduction and trend preservation. Table 2 summarizes the results of applying these methodologies to a randomly selected data point.

Table 2. Data Results Summary

Method	MSE	RMSE	MAE
Double SG_5	0.000819895	0.028633562	0.019775822
Double SG_30	0.000748731	0.027362938	0.018563044
ARMD3	0.000539939	0.023234644	0.015212056
Savitzky-Golay	0.000466357	0.021595299	0.014017021
Whittaker	0.000605135	0.024599485	0.01602796

The Savitzky-Golay method performs best in terms of smoothing effects, with the lowest MSE, RMSE, and MAE. It effectively reduces noise while preserving data trends, making it the optimal smoothing method in this study. The ARMD3 method performs well, with slightly higher MSE and RMSE than Savitzky-Golay, but it still strikes a good balance between noise reduction and trend preservation. However, its error control is slightly less effective. The Whittaker method performs well in smoothing and trend preservation, but with slightly higher RMSE and MAE compared to ARMD3. While it is suitable for time series with clear trends, it shows some limitations in noise reduction. The Double SG_5 method performs poorly, with higher MSE, RMSE, and MAE, weaker smoothing effects, and a risk of overfitting. Even with parameter adjustments, its performance does not surpass other methodology and remains unsatisfactory.

Considering all error metrics, the Savitzky-Golay filtering method performs the best in this study, effectively reducing noise while preserving data trends. The ARMD3 method ranks second, followed by the Whittaker method with moderate performance. The Double SG method performs poorly, with insufficient noise reduction at a window size of 5 and a risk of overfitting at a window size of 30. It is recommended to select the appropriate method based on actual needs.

4.3 Prediction results comparison

NDVI (Normalized Difference Vegetation Index) is closely related to the physical properties of vegetation. Research has shown that long-term changes and fluctuations in vegetation can be effectively monitored by detecting NDVI values.[58].After a rigorous data reconstruction process—including outlier removal, spatiotemporal interpolation, and smoothing—we obtained a high-quality NDVI time series spanning from January 10, 1985, to November 11, 2024. Considering global climate change and the accelerated updates of radar satellite data, we divided the dataset into three time periods: 1985-2011 as the training set (covering the complete vegetation growth cycle and typical climate fluctuations), 2012-2020 as the validation set (characterized by frequent extreme climate events), and 2021-2024 as the test set (used to evaluate the model's predictive ability for recent vegetation changes).

In terms of model selection, we considered both traditional machine learning methods and cutting-edge time series forecasting techniques. We used XGBoost and LightGBM as representatives of gradient boosting frameworks, which excel at capturing feature interactions. N-HiTS processes long-term dependencies through multi-scale decomposition, while TimeGPT and our self-developed LongTimeGPT (which employs an improved attention mechanism and long-term memory module) explore the transferability of large language models in time series forecasting. This combination allows for a systematic evaluation of the applicability boundaries of different algorithms in predicting vegetation dynamics.

The evaluation system uses multiple metrics: RMSE (to measure absolute error), MAPE (to reflect relative error proportions), NSE (to assess the overall goodness of fit of the model), and MAE (to provide a robust estimate of prediction bias). Specifically, for the NDVI values ranging from [0,1], we set an outlier filtering threshold for MAPE calculation to avoid percentage distortion in the lower value range.

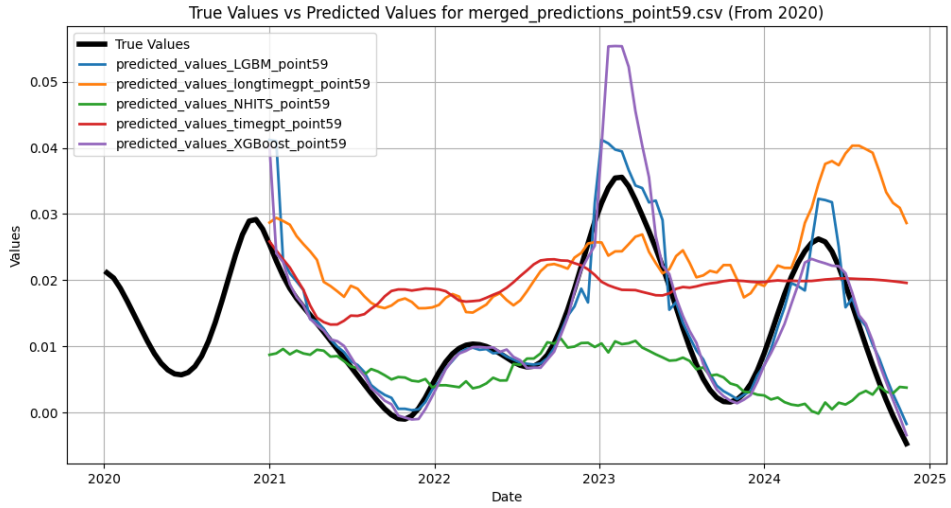

Figure 14: Various methods for predicting the result of point 59

As shown in **Figure 14**, using the randomly selected Point 59 as an example, the actual NDVI observations for this point are compared with the results of five different prediction methods. In the figure, the red solid line represents the true values, and the blue dashed line indicates the predicted values. The gap between the two is shaded in yellow to visually highlight the prediction error. This figure effectively compares the performance of the models from 2020 to 2024.

From the Figure 15 and Figure 16 LGBM and XGBoost models have significantly smaller yellow error regions, demonstrating excellent performance in NDVI time series forecasting. Among them, LGBM not only better aligns with the overall trend of the true data but also provides a smoother output curve, reflecting stronger time structure modeling ability and prediction stability.

On the flip side, the other three methods have their own shortcomings. N-HiTS and LongTimeGPT struggle with stability when it comes to capturing trends in the time series, leading to a noticeable gap between their predicted curves and the actual ones. Meanwhile, TimeGPT tends to underestimate how much things will change, which makes it too cautious when trying to fit the dynamic changes in NDVI, so it misses out on capturing the real ups and downs.

One possible reason for these problems could be that these models rely heavily on the amount of training data they get. Methods like LongTimeGPT and TimeGPT, which are

based on large model architectures, usually shine when they have access to big, diverse time series datasets. While our NDVI time series covers a long period, the number of samples we used for training in this study is still limited, which might not be enough for these models to learn effectively.

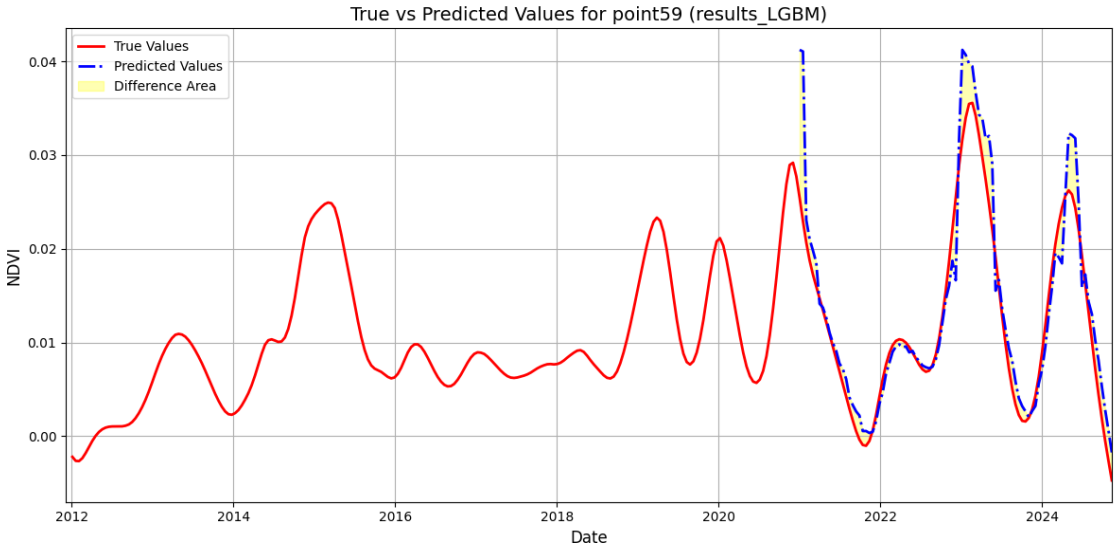

Figure 15: : LGBM methods for predicting the result of point 59

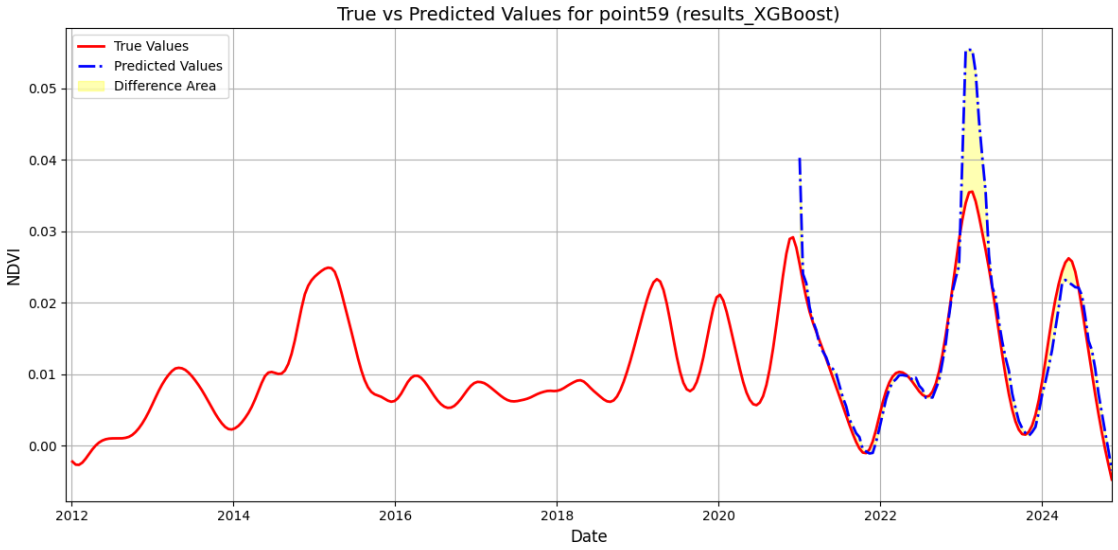

Figure 16: XGBoost methods for predicting the result of point 59

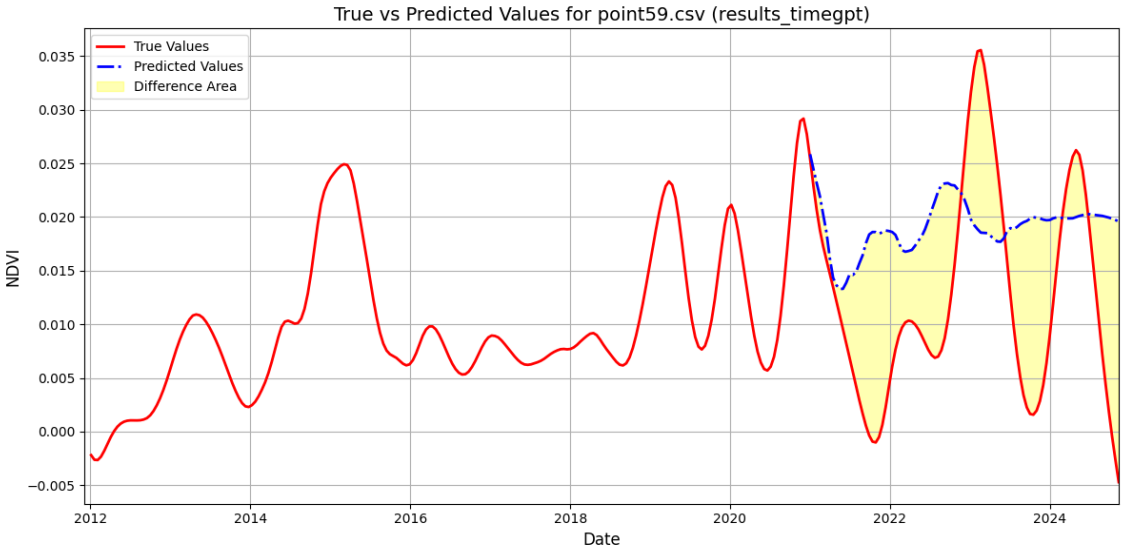

Figure 17:TimeGPT methods for predicting the result of point 59

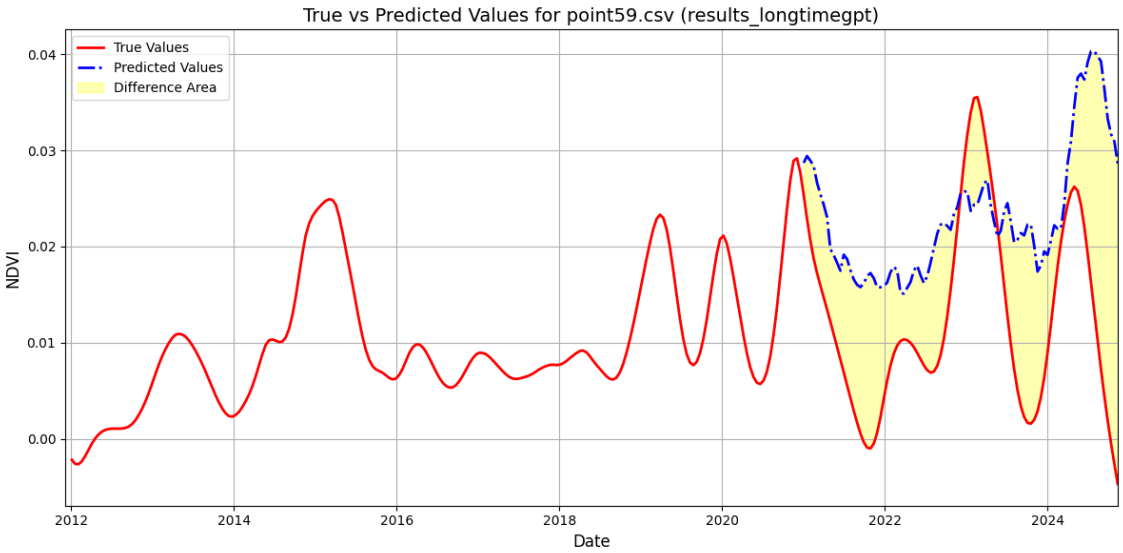

Figure 18: LongTimeGPT methods for predicting the result of point 59

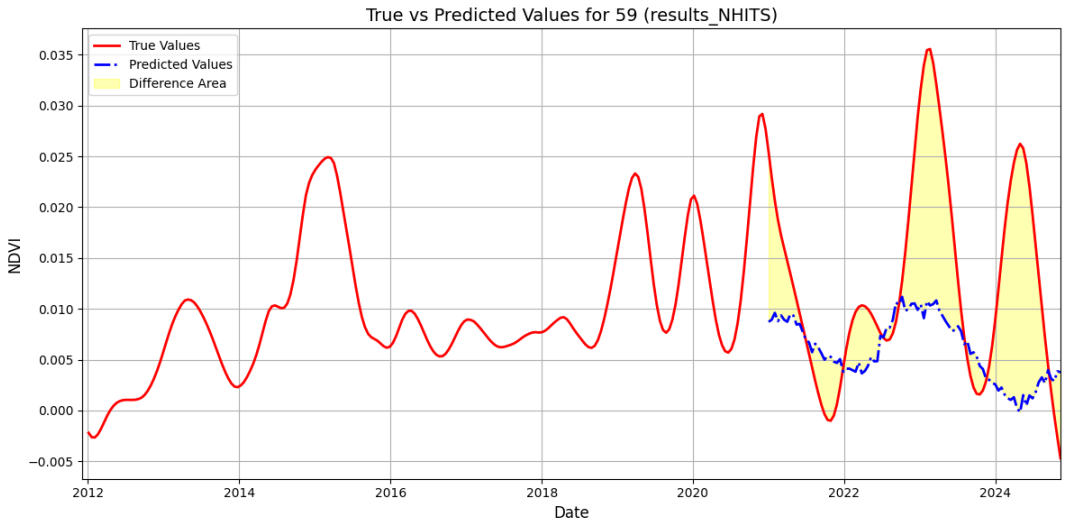

Figure 19: NHITS methods for predicting the result of point 59

To comprehensively assess the actual predictive capabilities of the five forecasting models (XGBoost, LightGBM, N-HiTS, TimeGPT, and LongTimeGPT), this study calculates five mainstream error metrics for all the points in the test set: MAE, RMSE, NSE, SMAPE, and MAPE. Each metric characterizes the deviation between predicted and true values from different perspectives:

Table 3. The predicted results of different methods

Method	MAE	RMSE	NSE	SMAPE	MAPE
LGBM	**0.00567**	**0.00741**	**0.79987**	**5.70154**	**4.64404**
NHITS	**0.02559**	**0.03107**	**-2.10849**	**21.41465**	**19.23557**
XGBoost	**0.00529**	**0.00688**	**0.83296**	**5.10141**	**4.18068**
LongTimeGPT	**0.01881**	**0.02366**	**-0.35426**	**14.59826**	**18.87607**
TimeGPT	**0.01876**	**0.02394**	**-0.28320**	**14.25929**	**18.67929**

Based on the overall evaluation metrics, the XGBoost model performs the best across all dimensions. Its MAE is 0.00529, RMSE is 0.00688, NSE is 0.83296, SMAPE is 5.10%, and MAPE is 4.18%, showing excellent performance in error control and trend fitting. Following closely is LightGBM, with an MAE of 0.00567, RMSE of 0.00741, and NSE of 0.79987, demonstrating stable overall performance.

On the other hand, the N-HiTS model performed the worst, with an MAE of 0.02559 and an NSE as low as -2.10849, failing to effectively capture the NDVI trend and showing poor predictive ability. LongTimeGPT and TimeGPT also performed significantly worse than XGBoost and LGBM in terms of MAE (0.01881 and 0.01876, respectively) and NSE (-0.35426 and -0.28320, respectively), indicating that large models did not show the expected advantages with the dataset used in this study.

Overall, the other three models each have their own drawbacks. N-HiTS and LongTimeGPT are unstable in capturing time series trends, leading to a significant gap between their predicted curves and the actual ones. Meanwhile, TimeGPT tends to underestimate the magnitude of predictions, making it too conservative when fitting the dynamic changes in NDVI, which makes it hard to reflect the real fluctuations.

One possible reason for these issues could be the models' reliance on the scale of the training data. Models like LongTimeGPT and TimeGPT, which are based on large architectures, usually perform best when they have access to large, diverse time series datasets. While we constructed a long NDVI time series, the number of training samples is still relatively limited, which may not be enough to meet the data requirements for parameter learning in these models.

In summary, it is recommended to prioritize using the XGBoost model for NDVI time series forecasting, with LightGBM as an alternative. The performance of N-HiTS, LongTimeGPT, and TimeGPT is poor, making them suitable only for experimental use in specific scenarios.

4.4 Discussion

The main goal of this study is to see how different models perform when forecasting NDVI (Normalized Difference Vegetation Index) time series, and to find the best model for real-world applications. We analyzed NDVI data from 1985 to 2024 using several predictive models, including XGBoost, LightGBM, N-HiTS, TimeGPT, and LongTimeGPT. The results show that XGBoost stands out as the best performer across all evaluation metrics, with an MAE of 0.00529, RMSE of 0.00688, and an NSE of 0.83296. This indicates that it's highly accurate and stable in forecasting NDVI, even with a limited sample size.

Previous studies, like those by H. Łoś and G. Sousa Mendes, have noted that XGBoost and LightGBM do well in predicting Sentinel-2 data[55]. Rahul Gupta and Anil Kumar Yadav[59]have pointed out that XGBoost performs well in time series forecasting, and this study confirms that observation. However, our research goes further to highlight the advantages of XGBoost and LightGBM in capturing NDVI changes, especially during periods of frequent extreme climate events. At the same time, Mohit Apte and Yashodhara Haribhakta[60]have noted that N-HiTS tends to make accurate predictions in financial time series forecasting. However, in this study, which focuses on NDVI time series, the results are quite the opposite, suggesting that N-HiTS may not be as effective in capturing the dynamics of NDVI changes compared to its performance in other domains such as finance.

In summary, our results support our hypothesis that time series forecasting models can effectively capture the changing trends of NDVI. However, deep learning models like N-HiTS, TimeGPT, and LongTimeGPT didn't perform well in this study, indicating that they might not be suitable under certain data conditions. By comparing the performance of different models, we offer a fresh perspective for future research, recommending that XGBoost be prioritized for NDVI time series forecasting. This finding opens new avenues for future studies, especially in exploring how well these models work across different datasets and application scenarios.

Our observations also revealed significant non-seasonal abnormal fluctuations in the NDVI time series of certain points, especially in some years where the phase overlaps with the Landsat 7 SLC-off failure period[61], causing sharp and sudden changes in

NDVI values. We reasonably speculate that these fluctuations may be related to satellite sensor malfunctions during specific periods, potentially caused by the failure of the sensor's line scan corrector, leading to striping noise. Moreover, natural conditions such as cloud cover and other interference factors cannot be ignored, as they may have impacted the accuracy of the observations to some extent[62]. Therefore, in practical applications, further consideration of these interference factors, combined with monitoring data on sensor failures, can enhance the robustness and accuracy of forecasting models.

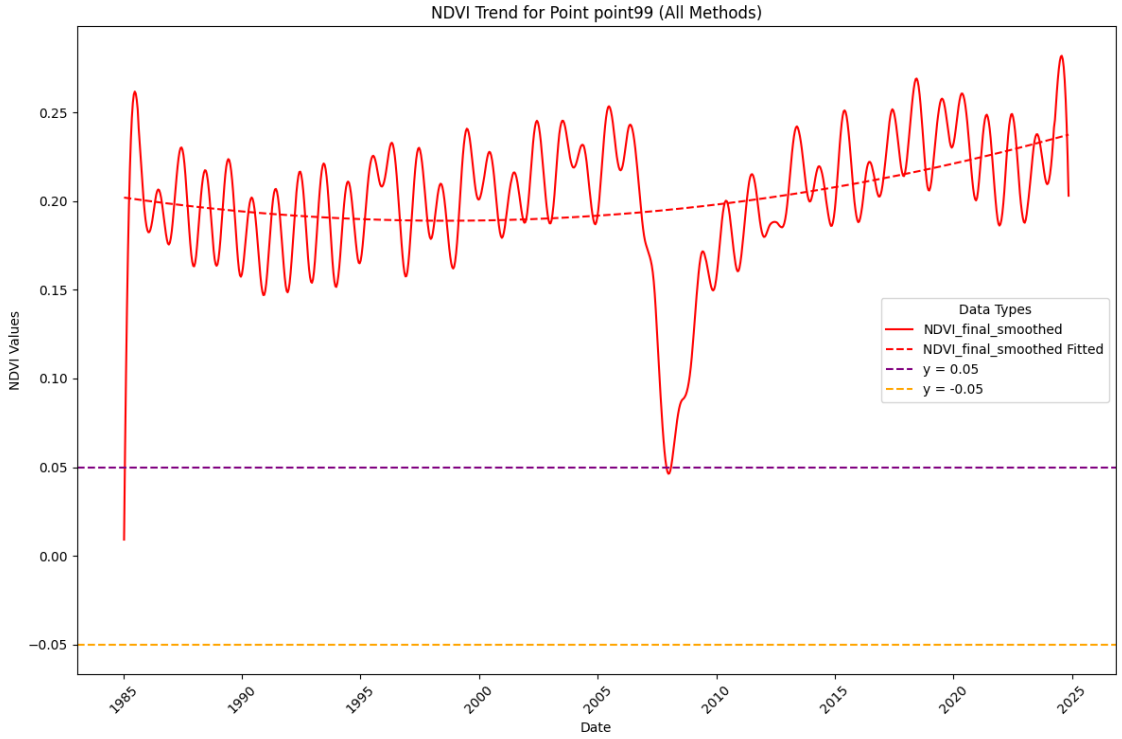

Figure 20: Extreme Case 1

On the other hand, in some other locations, the NDVI values showed a significant abrupt change during a specific period, after which they gradually stabilized and fluctuated around a new level without further obvious anomalies. Such changes are likely not coincidental, but rather the result of irreversible vegetation changes triggered by human activities, such as shifts in land use, construction of buildings, or intense agricultural interventions.

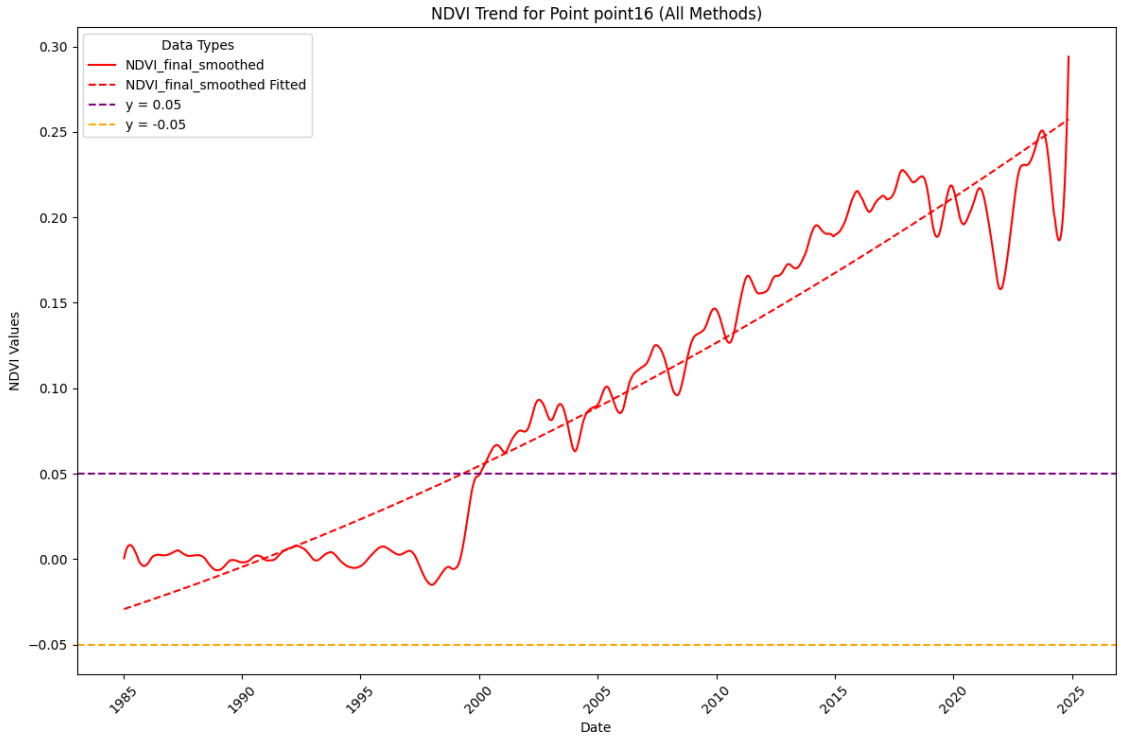

Figure 21: Extreme Case 2

5 Conclusions

This study is all about evaluating how different machine learning and time series forecasting models perform when predicting the NDVI (Normalized Difference Vegetation Index) time series, especially focusing on the vegetation distribution in the Troodos Mountains region. The Troodos Mountains are the main mountain range in Cyprus, known for their rich biodiversity and unique ecosystems. The vegetation here mainly consists of pine trees, oak trees, shrubs, and grasslands, and it shows significant changes over time due to climate change and human activities.

We looked at NDVI data from 1985 to 2024 and used several models for forecasting, including XGBoost, LightGBM, N-HiTS, TimeGPT, and LongTimeGPT. The results showed that XGBoost outperformed all the other models across the board (in terms of MAE, RMSE, and NSE), giving us the most accurate and stable predictions. On the other hand, the N-HiTS and GPT-based models didn't do so well, struggling to effectively capture the patterns of NDVI changes.

These findings are in line with previous research highlighting the importance of NDVI in tracking vegetation changes. However, we noticed that some models, especially the larger language models, performed poorly when the data was limited. Because of this, we recommend prioritizing XGBoost for NDVI forecasting or considering a mix of models to take advantage of their strengths and boost prediction accuracy.

Overall, this study offers new insights into NDVI time series forecasting in the Troodos Mountains and sets the stage for future research. There's a lot of potential for exploring how well these models work with different datasets and in various application scenarios. Future studies could also focus on optimizing the models that didn't perform as well or using ensemble learning methods to improve overall prediction capabilities. Plus, diving deeper into the changes in vegetation distribution in the Troodos Mountains will help us better understand how climate change impacts these ecosystems.

BIBLIOGRAPHY

[1] 穆. ·迪赫 *et al.*, "Empirical curvelet transform based deep DenseNet model to predict NDVI using RGB drone imagery data," *Computers and Electronics in Agriculture,* vol. 221, 2024/06/01, doi: 10.1016/j.compag.2024.108964.

[2] A. Pons and P. Quézel, "A propos de la mise en place du climat méditerranéen," *Comptes Rendus de l'Académie des Sciences-Series IIA-Earth and Planetary Science,* vol. 327, no. 11, pp. 755-760, 1998.

[3] P. L. Fall, "Modern vegetation, pollen and climate relationships on the Mediterranean island of Cyprus," *Review of Palaeobotany and Palynology,* vol. 185, pp. 79-92, 2012.

[4] Y. Liu, H. Liu, Y. Chen, C. Gang, and Y. Shen, "Quantifying the contributions of climate change and human activities to vegetation dynamic in China based on multiple indices," *Science of The Total Environment,* vol. 838, 2022/09/10, doi: 10.1016/j.scitotenv.2022.156553.

[5] 杨达 *et al.*, "青藏高原植被生长季 NDVI 时空变化与影响因素," (in chi), *Chinese Journal of Applied Ecology / Yingyong Shengtai Xuebao,* vol. 32, no. 4, pp. 1361-1372, 2021, doi: 10.13287/j.1001-9332.202104.014.

[6] D. Chen *et al.*, "Crop NDVI time series construction by fusing Sentinel-1, Sentinel-2, and environmental data with an ensemble-based framework," *Computers and Electronics in Agriculture,* vol. 215, 2023/12/01, doi: 10.1016/j.compag.2023.108388.

[7] Q. Liu *et al.*, "Identification of impact factors for differentiated patterns of NDVI change in the headwater source region of Brahmaputra and Indus, Southwestern Tibetan Plateau," *Ecological Indicators,* vol. 125, 2021/06/01, doi: 10.1016/j.ecolind.2021.107604.

[8] D. Viviroli *et al.*, "Climate change and mountain water resources: overview and recommendations for research, management and policy," *Hydrology and Earth System Sciences,* vol. 15, no. 2, 2011/02/04, doi: 10.5194/hess-15-471-2011.

[9] 杨元合 and 朴世龙, "青藏高原草地植被覆盖变化及其与气候因子的关系," *植物生态学报,* vol. 30, no. 1, pp. 1-8, 2006, doi: 10.17521/cjpe.2006.0001.

[10] D. V. Gaso, A. G. Berger, and V. S. Ciganda, "Predicting wheat grain yield and spatial variability at field scale using a simple regression or a crop model in conjunction with Landsat images," *Computers and Electronics in Agriculture,* vol. 159, 2019/04/01, doi: 10.1016/j.compag.2019.02.026.

[11] S. Abdikan, A. Sekertekin, O. G. Narin, A. Delen, and F. B. Sanli, "A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height estimation of sunflower using Sentinel-1 and Sentinel-2," *Advances in Space Research,* vol. 71, no. 7, 2023/04/01, doi: 10.1016/j.asr.2022.11.046.

[12] Y. Chen, R. Cao, J. Chen, L. Liu, and B. Matsushita, "A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 180, 2021/10/01, doi: 10.1016/j.isprsjprs.2021.08.015.

[13] A. Toosi, F. D. Javan, F. Samadzadegan, S. Mehravar, A. Kurban, and H. Azadi, "Citrus orchard mapping in Juybar, Iran: Analysis of NDVI time series and feature fusion of multi-source satellite imageries," *Ecological Informatics,* vol. 70, 2022/09/01, doi: 10.1016/j.ecoinf.2022.101733.

[14] 曹如茵, 杨. 晨, 陈瑾, X. Zhu, and M. Shen, "Thick cloud removal in Landsat images based on autoregression of Landsat time-series data," *Remote Sensing of Environment,* vol. 249, 2020/11/01, doi: 10.1016/j.rse.2020.112001.

[15] T. Hilker *et al.*, "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sensing of Environment,* vol. 113, no. 8, 2009/08/01, doi: 10.1016/j.rse.2009.03.007.

[16] C. Wang, A. Wang, D. Guo, H. Li, and S. Zang, "Off-peak NDVI correction to reconstruct Landsat time series for post-fire recovery in high-latitude forests," *International Journal of Applied Earth Observation and Geoinformation,* vol. 107, 2022/03/01, doi: 10.1016/j.jag.2022.102704.

[17] H. Lv, Y. Wang, and Y. Shen, "An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands," *Remote Sensing of Environment,* vol. 179, 2016/06/15, doi: 10.1016/j.rse.2016.03.034.

[18] S. Mohanasundaram, T. Baghel, V. Thakur, P. Udmale, and S. Shrestha, "Reconstructing NDVI and land surface temperature for cloud cover pixels of Landsat-8 images for assessing vegetation health index in the Northeast region of Thailand," *Environ Monit Assess,* vol. 195, no. 1, p. 211, Dec 19 2022, doi: 10.1007/s10661-022-10802-5.

[19] K. Kanjin and B. M. Alam, "Assessing changes in land cover, NDVI, and LST in the Sundarbans mangrove forest in Bangladesh and India: A GIS and remote sensing approach," *Remote Sensing Applications: Society and Environment,* vol. 36, 2024/11/01, doi: 10.1016/j.rsase.2024.101289.

[20] R. Prăvălie *et al.*, "NDVI-based ecological dynamics of forest vegetation and its relationship to climate change in Romania during 1987–2018," *Ecological Indicators,* vol. 136, 2022/03/01, doi: 10.1016/j.ecolind.2022.108629.

[21] V. B. Verhoeven and I. C. Dedoussi, "Annual satellite-based NDVI-derived land cover of Europe for 2001–2019," *Journal of Environmental Management,* vol. 302, 2022/01/15, doi: 10.1016/j.jenvman.2021.113917.

[22] M. Belgiu, A. Stein, M. Belgiu, and A. Stein, "Spatiotemporal Image Fusion in Remote Sensing," *Remote Sensing 2019, Vol. 11, Page 818,* vol. 11, no. 7, 2019-04-04, doi: 10.3390/rs11070818.

[23] S. Anand, R. Sharma, S. Anand, and R. Sharma, "Pansharpening and spatiotemporal image fusion method for remote sensing," *Engineering Research Express,* vol. 6, no. 2, 2024-04-11, doi: 10.1088/2631-8695/ad3a34.

[24] H. Hassani, R. Razavi-Far, M. Saif, and L. Lin, "Towards Sample-Efficiency and Generalization of Transfer and Inverse Reinforcement Learning: A Comprehensive Literature Review," 2024/11/15, doi: 10.48550/arXiv.2411.10268.

[25] W. Li and C.-Y. Hsu, "Automated terrain feature identification from remote sensing imagery: a deep learning approach," *International Journal of Geographical Information Science,* vol. 34, no. 4, 2020-4-2, doi: 10.1080/13658816.2018.1542697.

[26] L. F. Fuentes-Cortés, A. Flores-Tlacuahuac, and K. D. P. Nigam, "Machine Learning Algorithms Used in PSE Environments: A Didactic Approach and Critical Perspective," *Industrial & Engineering Chemistry Research,* vol. 61, no. 25, May 24, 2022, doi: 10.1021/acs.iecr.2c00335.

[27] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *nature,* vol. 518, no. 7540, pp. 529-533, 2015.

[28] S. J. Pan and Q. Yang, "A Survey on Transfer Learning | IEEE Journals & Magazine | IEEE Xplore," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, 2010, doi: 10.1109/TKDE.2009.191.

[29] E. Kriegler *et al.*, "The need for and use of socio-economic scenarios for climate change analysis: A new approach based on shared socio-economic pathways," *Global Environmental Change,* vol. 22, no. 4, 2012/10/01, doi: 10.1016/j.gloenvcha.2012.05.005.

[30] S. Elsawah *et al.*, "Eight grand challenges in socio-environmental systems modeling," *Socio-Environmental Systems Modelling,* vol. 2, 2020/01/01, doi: 10.18174/sesmo.2020a16226.

[31] A. J. Jakeman, S. El Sawah, S. Cuddy, B. Robson, N. McIntyre, and F. Cook, "QWMN Good Modelling Practice Principles," 2018.

[32] F. Villa *et al.*, "A Methodology for Adaptable and Robust Ecosystem Services Assessment," *PLOS ONE,* vol. 9, no. 3, 2014 年 3 月 13 日, doi: 10.1371/journal.pone.0091001.

[33] I. Soubry *et al.*, "A Systematic Review on the Integration of Remote Sensing and GIS to Forest and Grassland Ecosystem Health Attributes, Indicators, and Measures," *Remote Sensing 2021, Vol. 13, Page 3262,* vol. 13, no. 16, 2021-08-18, doi: 10.3390/rs13163262.

[34] J. Cabello *et al.*, "The ecosystem functioning dimension in conservation: insights from remote sensing," *Biodiversity and Conservation 2012 21:13,* vol. 21, no. 13, 2012-09-23, doi: 10.1007/s10531-012-0370-7.

[35] J. M. Kattimani and T. R. Prasad, "Normalized Difference Vegetation Index (NDVI) Applications in Part of South-Eastern Dry Agro-Climatic Zones of Karnataka Using Remote Sensing and GIS," *International Journal,* vol. 3, no. 12, pp. 1593-1596, 2015.

[36] E. Nestola *et al.*, "Monitoring Grassland Seasonal Carbon Dynamics, by Integrating MODIS NDVI, Proximal Optical Sampling, and Eddy Covariance Measurements," *Remote Sensing 2016, Vol. 8, Page 260,* vol. 8, no. 3, 2016-03-19, doi: 10.3390/rs8030260.

[37] N. Fernández, "Earth observation for species diversity assessment and monitoring," *Earth observation of ecosystem services,* pp. 151-177, 2013.

[38] C. R. Robins, "Spatial analysis of soil depth variability and pedogenesis along toposequences in the Troodos Mountains, Cyprus," 2004.

[39] "3Sigma | Proceedings of the Thirteenth EuroSys Conference," *Proceedings of the Thirteenth EuroSys Conference,* 2018, doi: 10.1145/3190508.3190515.

[40] D. Deng, "DBSCAN Clustering Algorithm Based on Density | IEEE Conference Publication | IEEE Xplore," *2020 7th International Forum on Electrical Engineering and Automation (IFEEA),* 2020, doi: 10.1109/IFEEA51475.2020.00199.

[41] S. Xu, H. Liu, L. Duan, and W. Wu, "An Improved LOF Outlier Detection Algorithm | IEEE Conference Publication | IEEE Xplore," doi: 10.1109/ICAICA52286.2021.9498181.

[42] M. Khodarahmi, V. Maihami, M. Khodarahmi, and V. Maihami, "A Review on Kalman Filter Models," *Archives of Computational Methods in Engineering 2022 30:1,* vol. 30, no. 1, 2022-10-01, doi: 10.1007/s11831-022-09815-7.

[43] Z. Zhitao, Y. Lan, W. Pute, and H. Wenting, "Model of soybean NDVI change based on time series," *International Journal of Agricultural and Biological Engineering,* vol. 7, no. 5, 2014/10/30, doi: 10.25165/ijabe.v7i5.1061.

[44] R. Cao *et al.*, "A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter," *Remote Sensing of Environment,* vol. 217, 2018/11/01, doi: 10.1016/j.rse.2018.08.022.

[45] S. R. Krishnan and C. S. Seelamantula, "On the Selection of Optimum Savitzky-Golay Filters | IEEE Journals & Magazine | IEEE Xplore," *IEEE Transactions on Signal Processing,* vol. 61, no. 2, 2013, doi: 10.1109/TSP.2012.2225055.

[46] M. Lepot, J.-B. Aubin, F. H. L. R. Clemens, M. Lepot, J.-B. Aubin, and F. H. L. R. Clemens, "Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment," *Water 2017, Vol. 9, Page 796,* vol. 9, no. 10, 2017-10-17, doi: 10.3390/w9100796.

[47] Y. Julien and J. A. Sobrino, "Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data," *International Journal of Applied Earth Observation and Geoinformation,* vol. 76, 2019/04/01, doi: 10.1016/j.jag.2018.11.008.

[48] J. Gu, X. Li, C. Huang, and G. S. Okin, "A simplified data assimilation method for reconstructing time-series MODIS NDVI data," *Advances in Space Research,* vol. 44, no. 4, 2009/08/17, doi: 10.1016/j.asr.2009.05.009.

[49] A. B. Baloloy, A. C. Blanco, R. R. C. S. Ana, and K. Nadaoka, "Development and application of a new mangrove vegetation index (MVI) for rapid and accurate mangrove mapping," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 166, 2020/08/01, doi: 10.1016/j.isprsjprs.2020.06.001.

[50] S. Lipovetsky, "Double logistic curve in regression modeling," *Journal of Applied Statistics,* 2010-11-1, doi: 10.1080/02664760903093633.

[51] C.-X. Lv *et al.*, "Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model," *BMC Infectious*

Diseases 2021 21:1, vol. 21, no. 1, 2021-08-19, doi: 10.1186/s12879-021-06503-y.

[52] 罗俊玲, 张忠良, 姚. 福, and 冯. 饶, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results in Physics,* vol. 27, 2021/08/01, doi: 10.1016/j.rinp.2021.104462.

[53] "A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting," *IEEE Access,* vol. 7, 2019, doi: 10.1109/ACCESS.2019.2901920.

[54] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu-Rub, "A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting," *Energy,* vol. 214, 2021/01/01, doi: 10.1016/j.energy.2020.118874.

[55] "Evaluation of Xgboost and Lgbm Performance in Tree Species Classification with Sentinel-2 Data," *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS,* 2021, doi: 10.1109/IGARSS47720.2021.9553031.

[56] "Integrating the Focusing Neuron Model with N-BEATS and N-HiTS," *2024 9th International Conference on Computer Science and Engineering (UBMK),* 2024, doi: 10.1109/UBMK63289.2024.10773495.

[57] C. Shyalika, H. K. Bagga, A. Bhatt, R. Prasad, A. A. Ghazo, and A. Sheth, "Time Series Foundational Models: Their Role in Anomaly Detection and Prediction," 2024/12/26, doi: 10.48550/arXiv.2412.19286.

[58] D. Mao, Z. Wang, L. Luo, and C. Ren, "Integrating AVHRR and MODIS data to monitor NDVI changes and their relationships with climatic parameters in Northeast China," *International Journal of Applied Earth Observation and Geoinformation,* vol. 18, 2012/08/01, doi: 10.1016/j.jag.2011.10.007.

[59] "LTBoost: Boosted Hybrids of Ensemble Linear and Gradient Algorithms for the Long-term Time Series Forecasting | Proceedings of the 33rd ACM International Conference on Information and Knowledge Management," *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management,* vol. 38, 2024, doi: 10.1145/3627673.3679527.

[60] M. Apte and Y. Haribhakta, "Advancing Financial Forecasting: A Comparative Analysis of Neural Forecasting Models N-HiTS and N-BEATS," 2024/08/31, doi: 10.48550/arXiv.2409.00480.

[61] F. Chen, L. Tang, C. Wang, and Q. Qiu, "Recovering of the thermal band of Landsat 7 SLC-off ETM+ image using CBERS as auxiliary data," *Advances in Space Research,* vol. 48, no. 6, 2011/09/15, doi: 10.1016/j.asr.2011.05.012.

[62] Y. M. Asare, E. K. Forkuo, G. Forkuor, and M. Thiel, "Evaluation of gap-filling methods for Landsat 7 ETM+ SLC-off image for LULC classification in a heterogeneous landscape of West Africa," *International Journal of Remote Sensing,* vol. 41, no. 7, 2020-4-2, doi: 10.1080/01431161.2019.1693076.