International Journal of Innovative Research and Scientific Studies, 8(2) 2025, pages: 2453-2459

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Statistical approach to gap free time series data generation for remote sense images using Google earth engine

Megharani B. Mayani[1*], Rajeshwari L. Itagi[2]

[1]*Government Polytechnic Belgaum, Electronics and Communication Engineering, Department of Collegiate and Technical Education, Government of Karnataka, and Visvesvaraya Technological University, Belagavi 590018, India.*
[2]*KLE Institute of Technology, Electronics and Communication Department, Hubballi – Karnataka, and Visvesvaraya Technological University, Belagavi 590018, India.*

Corresponding author: Megharani B. Mayani (*Email: megha.mayani@gmail.com*)

Abstract

Remote sensing is gaining popularity in agriculture systems due to the advancements in technology and the precise spatial and temporal data sensed by Earth Observation satellites, aid in land and water monitoring. But to align with Sustainable Development Goal, SDG 2.0 (Zero Hunger), accurate crop yield forecast is necessary to handle the crop shortages and crop surpluses accordingly. As crop yield prediction depends on efficient crop maps, reliable crop maps (of late mostly done by Machine learning/Deep learning models) require continuous time series data of croplands. Atmospheric attenuation has a big role to play in Optical satellites. Snow / cloud cover / aerosols degrade the onboard recorded values considerably. Cloudy pixel values make the analysis less accurate. Since optical satellites fail to deliver gap free time series data, a technique to recover/predict the missing pixel data becomes essential. To handle such pixel values, a technique to reconstruct or predict the missing values is much needed as Machine learning / Deep learning models are efficient when fed with gap free data. Observations from the conducted study by employing statistical methods over cloud computing platform have revealed an acceptable range of Root Mean Square Error and F1 score. In this paper, investigations done to get a best estimate of the missing pixel values, using a statistical approach is discussed. Rolling statistics (Moving Average), Gap filling, SG (Savitzky Golay) filtering, Interpolation are performed with Normalized Difference Vegetation Index (NDVI) values, over a period of four years at a sugarcane farmland region. Using GEE (Google Earth Engine) as a cloud computing service and satellite dataset provider, NDVI values obtained are compared with the actual ground values. SG filtering gave better approximations of missing values compared to other statistical methods. Our study demonstrates an effective technique in generating gap free data that improvises the performance of crop yield prediction models.

Keywords: Interpolation, Moving Average, NDVI, GEE, Savitzky Golay filter, Sentinel, Small holding.

DOI: 10.53894/ijirss.v8i2.5671
Funding: This study received no specific financial support.
History: Received: 13 February 2025 / **Revised:** 14 March 2025 / **Accepted:** 20 March 2025 / **Published**: 26 March 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Climate change and soil degradation are the key parameters affecting crop growth and hence, agriculture-based products. Since Indian economy depends on Agriculture and its products, unpreparedness in handling uncertainties about crop yields, crop quality, etc. will result in a food crisis [1]. To analyse the trend and understand the pattern of farming and predict yield before harvest, it requires timely monitoring of crop lands and farming practices. Timely monitoring of croplands spread across geographical areas, requires time and logistics management which is cumbersome when it is done manually. Among modern tools, satellite remote sense images are used [2]. Aerial photography data acquisition using drones is costly in comparison with satellite images. Analysing satellite images over a period will give a clue about vegetation patterns, land cover change details, disaster alarms, etc. Remotely sensed satellite images and ML (Machine Learning) / DL (Deep Learning) will be a unique set of solutions to periodically monitor agriculture areas; for crop identification, crop map generation, yield forecasting, to help strategy planning for crop shortage or surplus accordingly.

Analysing satellite images involves huge data and the vegetative indices used for above mentioned applications. Continuity in the captured band values is necessary for the statistical ML/DL analysis. Consider for example, satellite image analysis to identify the crop grown areas, requires continuous dynamic behaviour exhibited by the crops over its life cycle. But many optical earth observation satellites give 'gaps in image data' due to atmospheric attenuation of the signal especially during monsoon season. Cloud cover/fog attenuates the signal drastically and no values or incorrect values are recorded by the optical satellite sensors, creating a gap or missing data. So, to get gap free image data, a technique is needed [3] to fill or compensate for these missing values due to atmospheric conditions.

1.1. Objectives

Every crop has its own specific spectral signature, any images over a monitoring period carrying the attributes specific to a crop. This spectral information plays a crucial role and contributes to specimen spectral signature which differentiates the different crops and its varieties as well [4]. Spectral information is reflected in the band values collected by the sensors carried aboard by the satellite.

In this paper, a study has been done to investigate the best estimate of the missing pixel values. In ML/DL, predictions can be modelled with optimum accuracy if the model is trained with gap free data. With this as an objective, timely based satellite images of agricultural land grown with sugarcane are collected. This data is used as training data for different ML/DL techniques, to model gap free data. Google Earth Engine (GEE) is used as Platform as a Service (PaaS) to analyse the observations. Research work is conducted to generate a gap-free time series data, for satellite image analysis. This work includes (i) choosing a particular Indian crop, study region and its phenology study. (ii) selecting satellites covering the study region (iii) applying statistical methods to fill gaps in time series data and validation using in situ data (ground truthing).

1.2. Selection of Crop

Sugarcane is a principal commercial crop that is grown in the subtropical and tropical regions of the country. India has the largest area under the sugarcane cultivation throughout the globe and is second largest producer next only to Brazil [5]. In India, sugarcane production and sugar industry play an important role in socioeconomic development in rural areas by utilising resources and creating job opportunities and higher income as well. About 8.0% of the agricultural population and about 45 million sugarcane farmers, their dependents, and an oversized population of agricultural labour are associated with sugarcane cultivation, harvesting, and accessory activities in India [6]. Sugarcane accounts for the largest value of production in the country and holds top position among other commercial crops. It is a popular choice for cultivation among farmers wherever geographical and climatic conditions favour its growth [7]. Sugarcane crop grown as Eksali (Annual crop) is selected for study.

1.3. Selection of Satellite

Three satellite datasets were explored from open-source data catalogues of GEE spanning over the study region namely MODIS, Landsat, Sentinel. Sentinel datasets were selected due to its good spatial resolution (10m), acceptable spectral resolution (13 bands) having sufficient revisit time (5 days) as compared with MODIS and Landsat over study region. Sentinel datasets are mainly used for Land use / land cover applications provided by Copernicus, United States

Geological Survey (USGS). Sentinel data catalogue includes surface reflectance products and top of atmosphere reflectance products.

1.4. Application of Statistical Methods

Statistical methods employed using the inbuilt functions of GEE namely joins, masks and filters are-

- Moving Average: It completely changes the value of the time series. Average value at each point in the time series image was taken and replaced by taking either mean or median over the selected time window.
- Gap Filling: It changes only the missing / masked time series data. The first cloud free pixel was picked and replaced in mosaic. Proper selection on time window requires phenology knowledge and experience.
- Interpolation: It creates a series of empty images / time series, then temporal neighbourhood values are used to fill the gaps in the data.
- Savitzky Golay (SG) filter: It uses a polynomial to do curve fitting by the available data. Smoothening of the time series pattern was achieved which helps in trend analysis easily. Selecting the proper order for polynomials was challenging and required little expertise.

Predicting missing data in image is done with following steps:

i) Each image will be added with time – band.
ii) The image collection will be joined with itself to get before and after images as well.
iii) Interpolated images are created by replacing masked pixels with new processed values.
iv) Check and validate the results

The investigation demonstrated that Join, Mask, Filter in GEE can effectively fill gaps of time series remote sense data.

2. Materials and Methods

2.1. Study Region

The study area selected is Desur village, located at 15.74250 N (Latitude), 74.50530 E (Longitude), in Belagavi district, Karnataka. The study region is a part of the 'Sugar Bowl' of Karnataka state in India, which is known for accounting to 35% of state sugarcane production [8]. Sugarcane crop is grown as an annual crop in the selected study region.

After an interaction with cultivators, the crop calendar for annual (Eksali) sugarcane crop is as shown in Figure 1. Depending on the date of planting; sugarcane has 4 four growth phases namely Germination, Tillering, Grand growth, Maturity is as listed in Table 1. There is a practice called "Ratooning" where the crop is regrown for 2 – 3 two to three times every year after harvest, by preserving the roots of the sugarcane. The sugarcane varietal grown in the study area are mainly Co86032, CoC671, Co-09004. The study area was envisaged for the application of geospatial technology in crop identification and assessment of model in identifying sugarcane crop grown area in the selected study areas, using Satellite imageries and Machine Learning Techniques.

Figure 1.
Sugarcane crop grown at study region and its growth phases (Eksali variety – annual crop).

Table 1.
Growth phases of the sugarcane crop (Annual - Eksali).

Crop growth phase	Schedule at Study Region
Germination	Feb-Mar
Tillering	Apr-June
Grand growth	July-Sep
Maturity	Oct-Dec

2.2. Remote sensing data

GEE platform (an open source) was used to download remote sense images. Different composites of satellite imageries are available for public access to download and analyse. In this study, Sentinel 2, Surface Reflectance (SR) composites with 13 bands, 10 m spatial resolution and 5 days temporal resolution images were used. Multispectral data collected by the instrument on board and also metadata is available for analysis. GEE is also used as a computing resource as the Earth Engine API is available in Python and JavaScript, making it easy to harness the power of Google's cloud for our own geospatial analysis [9]. The web-based code editor is used for fast, interactive algorithm development with instant access to petabytes of data. This rendered a great flexibility on local machines by connecting to a powerful computing machine and its storage as an advantage. The data catalogue contains products with geometric, radiometric and atmospheric corrected satellite images. With 72 images every year for a period of 4 years will be 288 images to process and analyse. This information is vital for implementing efficient management strategies and anticipating potential yield losses.

2.3. In-Situ Sugarcane Crop Data Collection

In this study, study site Desur was visited where records of agriculture related information of the villages constituting the Taluka was collected. For verifying/ground truthing satellite data, and to understand the cropping patterns and issues, a detailed social survey and field survey were conducted. The field-based ground-truths were collected periodically (quarter) for the years 2023, 2024. Ground truths, for validating satellite data, were collected through Global Positioning System (GPS) [Garmin OREGON 650] device and Quantum Geographic Information System (QGIS), an open-source software is used to create shape files that can be uploaded as assets in GEE by surveying the study region. Spectroradiometer (Spectra Vista Corporation's portable instrument – SVC H512i data) as shown in Figure 2 is used to get gap-free sensor values, collected as a periodic visit to the study region. Vegetation index, Normalized Difference Vegetation Index (NDVI) which is also called crop growth profile indicator [4] was calculated using the values recorded from the devices mentioned above using expression (1). NDVI calculated for the years 2023 and 2024 are compared with statistical methods to find the best fit for predicting the missing values.

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{R})}{(\mathrm{NIR} + \mathrm{R})} \qquad (1)$$

where R and NIR represent pixel reflectance values in the red and near-infrared bands, respectively.

time= 08-18-2023 7:00:31 AM, 08-18-2023 7:13:33 AM
longitude= 07429.2657E , 07429.2740E
latitude= 1546.2247N , 1546.1499N
gpstime= 103342.000 , 104640.000
comm=
memory slot= 59, 69
factors= 1.000, 1.000, 1.000 [Overlap: Preserve, Matching Type: None]
inclinometer x offset= -2, -8
inclinometer y offset= -1, -1
data=

336.6	19221.94	449.75	2.34
338.2	21433.01	504.49	2.35
339.9	23188.79	535.78	2.31
341.6	24549.53	564.56	2.30
343.2	25017.53	558.80	2.23
344.9	26875.50	585.12	2.18
346.5	28350.00	616.88	2.18
348.2	29930.23	639.67	2.14
349.8	32223.13	678.63	2.11

Figure 2.
Sugarcane crop data collection using portable devices in field and reports generated.

2.4. Statistical Approach for Time Series Data Generation

Time series analysis describes, explains, and predicts changes in a phenomenon through time. People have utilised techniques that add a distinctive spatial dimension to this type of analysis [10]. Major applications of spatiotemporal analysis include forecasting yield, analysing the development of crops in croplands, and forecasting/back casting economic risks in case of shortage or surplus accordingly. Crop identification and grown area estimation can be predicted with a gap free pattern from satellite images.

Multispectral Instrument (MSI) carried by optical satellites being used for tracking continuous events on land use activities. MSI contain different types of sensors which can acquire land cover features like vegetation, soil, water bodies, and coastal areas, including information on their health and changes over time; this data is used for monitoring land use, vegetation dynamics, and environmental changes across various scales [11]. These features are recorded in the form of seamless univariate or multivariate time-series data. Very often, however, the data contains missing data which disrupts the continuity of the data making it difficult to analyse the data. The missing part of the data needs to be ascribed to make the remaining available data applicable.

Choosing the proper ascribing method is crucial for fruitful analysis and extracting underlined features from the data [3]. NDVI is used as an important performance indicator. Different statistical methods namely Moving average, Gap filling, Interpolation, Savitzky Golay filter are used.

The moving average is calculated for each individual missing NDVI value using expression

$$Y(i)_{j=0}^{M-1} = \left\{\frac{1}{M}\right\} \sum X[i+j] \qquad (2)$$

where Y: NDVI value, X: time stamp, j: window size

The sharp rise and falls in NDVI values were smoothened by Moving average using (2) are depicted in Figure 3, but long gaps were unattended. Creating a regularly spaced time series and then using temporal neighbours NDVI to interpolate the missing values was done using (3).

$$Y = Y_1 + [Y_2 - Y_1] * \frac{t - t_1}{t_2 - t_1} \qquad (3)$$

where Y = interpolated image,

Y1 = before image, Y2 = after image

t = interpolation timestamp

t1 = before image timestamp, t2 = after image timestamp

Also, application of SG filter was performed which indeed produced the best results after hyper tuning the parameters. NDVI composite after Savitzky Golay filtering is shown in Figure 3. These values were found close to ground truth values after validation. Savitzky Golay polynomial is optimised for the time window size and order of the polynomial.

2.5. Statistical Methods for Gap Free NDVI

Four years of sentinel data namely 2020, 2021, 2022 and 2023 are downloaded for the selected study region and NDVI are computed which is shown in Table 1. As evident the gaps in time series data will hamper the performance of machine learning / deep learning techniques [12].

In Shibuya, et al. [13] statistical methods namely interpolation, mean and median were tested for sugarcane crop using Venμs satellite time series with NDVI composites and achieved an overall accuracy of 0.81, used for Brazil agriculture landscapes. Also, better discrimination of annual crops having high temporal dynamics by improvising or optimising the statistical methods is achievable. The study in this work, revealed improved performance of the statistical methods.

These gap free NDVI will be essentially important hyperparameters for Machine learning techniques or Deep learning techniques to analyse the crop status in mid-season, forecast the yield which aids in framing policies, planning strategies, decision making.

3. Results

The gap-free NDVI plots using the Moving average, Gap filling, Interpolation, SG filter methods, for the years 2020, 2021, 2022, 2023 are shown in Figure 3. It is observed that aberrations and assignment of missing values in SG filter is near to the real values, with minimum error.

As evident from Table 2, monsoon data is not available from remote sensed images and accordingly it is not possible to calculate the NDVI values during those timelines, which in fact is a crucial phase in the crop calendar of sugarcane crop as steep growth of crop and farming activities are intensely engaging during monsoon. Number of sampling points directly maps to the accuracy of the data prediction model. Unavailability of data during these ‘three to four months’ needs a solution to be addressed that is more reliable, accurate, simple for implementation. SG filtering gave good results (Please refer to Figure 3) over NDVI composites.

Table 2.
NDVI values from satellite time series data for years 2020, 2021, 2022 and 2023.

Year 2020		Year 2021		Year 2022		Year 2023	
Date	**NDVI**	**Date**	**NDVI**	**Date**	**NDVI**	**Date**	**NDVI**
15 Jan 20	0.176	19-Jan-21	0.224	04-Jan-22	0.389	04-Jan-23	0.211
20-Jan-20	0.155	24-Jan-21	0.199	09-Jan-22	0.411	14-Jan-23	0.218
25-Jan-20	0.178	03-Feb-21	0.139	19-Jan-22	0.35	19-Jan-23	0.225
30-Jan-20	0.161	08-Feb-21	0.152	29-Jan-22	0.324	24-Jan-23	0.194
04-Feb-20	0.151	13-Feb-21	0.138	03-Feb-22	0.313	13-Feb-23	0.236
19-Feb-20	0.141	28-Feb-21	0.141	08-Feb-22	0.304	18-Feb-23	0.258
29-Feb-20	0.119	05-Mar-21	0.156	13-Feb-22	0.27	28-Feb-23	0.254
15-Mar-20	0.21	10-Mar-21	0.171	18-Feb-22	0.246	05-Mar-23	0.231
20-Mar-20	0.276	15-Mar-21	0.189	23-Feb-22	0.195	10-Mar-23	0.276
30-Mar-20	0.244	20-Mar-21	0.156	28-Feb-22	0.179	30-Mar-23	0.333
09-Apr-20	0.241	04-Apr-21	0.217	05-Mar-22	0.156	04-Apr-23	0.321
19-Apr-20	0.247	09-Apr-21	0.173	10-Mar-22	0.131	09-Apr-23	0.33
24-Apr-20	0.297	19-Apr-21	0.272	15-Mar-22	0.153	19-Apr-23	0.319
09-May-20	0.324	29-Apr-21	0.321	30-Mar-22	0.205	24-May-23	0.46
14-May-20	0.371	04-May-21	0.401	26-Oct-22	0.531	29-May-23	0.47
24-May-20	0.367	26-Oct-21	0.482	31-Oct-22	0.511	03-Jun-23	0.522
05-Nov-20	0.525	25-Nov-21	0.555	30-Nov-22	0.444	26-Oct-23	0.581
10-Nov-20	0.52	20-Dec-21	0.462	20-Dec-22	0.396	15-Dec-23	0.332
30-Nov-20	0.497	25-Dec-21	0.39	25-Dec-22	0.259	25-Dec-23	0.264
15-Dec-20	0.427	30-Dec-21	0.401	30-Dec-22	0.244	30-Dec-23	0.24

Table 3.
RMSE and F1 scores of statistical methods used for study.

Metric/ Method	Moving Average	Gap filling	Interpolation	SG filter
RMSE	0.0851	0.0812	0.0731	0.0646
F1 score	0.645	0.664	0.725	0.789

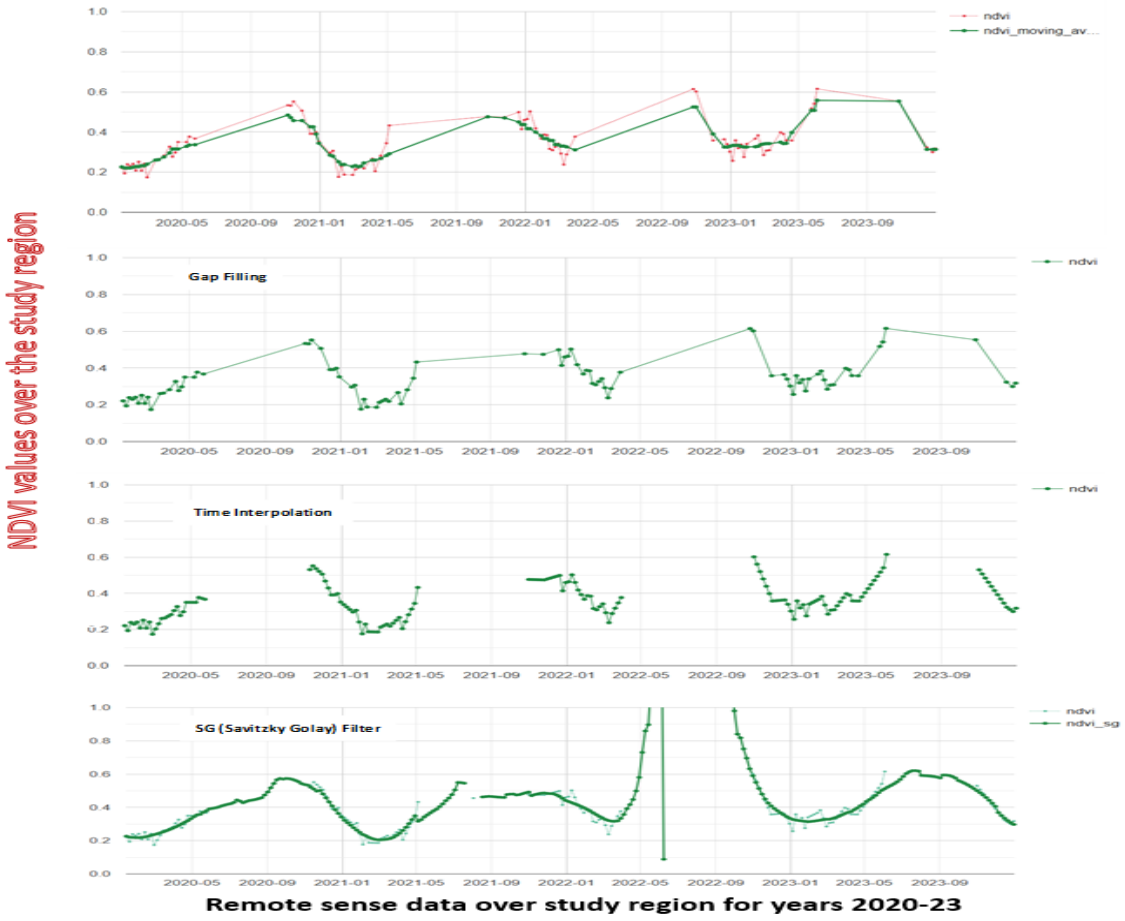

Figure 3.
NDVI missing value assignment by statistical methods.

4. Discussion

Accuracy assessment was performed using Root Mean Square Error (RMSE), F1 score as metrics for evaluation. Results obtained are as tabulated in Table 3. For sugarcane crops, the results were validated using spectroradiometer readings collected in situ, using spectral range 350 nm - 1050 nm with 512 channels. In [14] have used to study the best estimates for agriculture crop characteristics by optimising wavebands and widths. The reflectance spectroradiometer readings are used to get physicochemical variables of soil [15] paving way towards precision agriculture; but requires sophisticated interpretation (due to huge data). This study investigated the usefulness of spectroradiometer reflectance values in validating the statistical methods estimates of missing pixel values. The aberrations smoothened by SG filter will be a data preprocessing task to feed ML / DL models with a gap free data. Crop dynamics has a specific pattern and SG filter method supported more compared to MA, GAP filling, Interpolation methods to generate the gap free time series data.

References

[1] K. Pawlak and M. Kołodziejczak, "The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production," *Sustainability*, vol. 12, no. 13, p. 5488, 2020. https://doi.org/10.3390/su12135488

[2] P. Shanmugapriya, S. Rathika, T. Ramesh, and P. Janaki, "Applications of remote sensing in agriculture-A Review," *International Journal of Current Microbiology and Applied Sciences*, vol. 8, no. 01, pp. 2270-2283, 2019. https://doi.org/10.20546/ijcmas.2019.801.238

[3] K. Lee *et al.*, "Fast and operational gap filling in satellite-derived aerosol optical depths using statistical techniques," *Journal of Applied Remote Sensing*, vol. 16, no. 4, pp. 044507-044507, 2022. https://doi.org/10.1117/1.JRS.16.044507

[4] S. S. Panda, D. P. Ames, and S. Panigrahi, "Application of vegetation indices for agricultural crop yield prediction using neural network techniques," *Remote Sensing*, vol. 2, no. 3, pp. 673-696, 2010. https://doi.org/10.3390/rs2030673

[5] A. Bégué *et al.*, "Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI," *International Journal of Remote Sensing*, vol. 31, no. 20, pp. 5391-5407, 2010. https://doi.org/10.1080/01431160903349057

[6] A. Narmilan, F. Gonzalez, A. S. A. Salgadoe, U. W. L. M. Kumarasiri, H. A. S. Weerasinghe, and B. R. Kulasekara, "Predicting canopy chlorophyll content in sugarcane crops using machine learning algorithms and spectral vegetation indices derived from UAV multispectral imagery," *Remote Sensing*, vol. 14, no. 5, p. 1140, 2022. https://doi.org/10.3390/rs14051140

[7] K. Malik *et al.*, "Sugarcane production and its utilization as a biofuel in India: Status, perspectives, and current policy," *Sugarcane Biofuels: Status, Potential, and Prospects of the Sweet Crop to Fuel the World*, pp. 123-138, 2019. https://doi.org/10.1007/978-3-030-18597-8_6

[8] A. Lonare, B. Maheshwari, and P. Chinnasamy, "Village level identification of sugarcane in Sangali, Maharashtra using open source data," *Journal of Agrometeorology*, vol. 24, no. 3, pp. 249-254, 2022. https://doi.org/10.54386/jam.v24i3.1688

[9] M. Aghababaei, A. Ebrahimi, A. A. Naghipour, E. Asadi, and J. Verrelst, "Vegetation types mapping using multi-temporal landsat images in the google earth engine platform," *Remote Sensing*, vol. 13, no. 22, p. 4683, 2021. https://doi.org/10.3390/rs13224683

[10] S. T. Arab, R. Noguchi, S. Matsushita, and T. Ahamed, "Prediction of grape yields from time-series vegetation indices using satellite remote sensing and a machine-learning approach," *Remote Sensing Applications: Society and Environment*, vol. 22, p. 100485, 2021. https://doi.org/10.1016/j.rsase.2021.100485

[11] M. Faisal and R. Makar, "Development of a simplified technique for gap filling of Normalize Difference Vegetation Index (NDVI) time series data," *Journal of Applied and Natural Science*, vol. 14, no. 4, p. 1500, 2022. https://doi.org/10.31018/jans.v14i4.4095

[12] T. Butkhot and P. Reungsang, "Assessment of machine learning on sugarcane classification using landsat-8 and sentinel-2 satellite imagery," *Asia-Pacific Journal of Science and Technology*, vol. 26, no. 4, pp. 1-11, 2021. https://doi.org/10.14456/apst.2021.38

[13] D. H. Shibuya, G. M. Pereira, G. K. Figueiredo, A. C. d. S. Luciano, R. A. Lamparelli, and G. le Maire, "Evaluation of time series gap-filling of venμs satellite for land use classification," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021: IEEE, pp. 4244-4247.

[14] P. S. Thenkabail, R. B. Smith, and E. De Pauw, "Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 6, pp. 607-622, 2002.

[15] A. D. Vibhute, K. V. Kale, S. C. Mehrotra, R. K. Dhumal, and A. D. Nagne, "Determination of soil physicochemical attributes in farming sites through visible, near-infrared diffuse reflectance spectroscopy and PLSR modeling," *Ecological Processes*, vol. 7, pp. 1-12, 2018. https://doi.org/10.1186/s13717-018-0138-4