AATCC Review

02 September 2023: Received
22 October 2023: Revised
08 January 2024: Accepted
28 January 2024: Available Online

www.aatcc.peerjournals.net

Research Article Open Access

Remote Sensing Data-Based Sugarcane Acreage Estimation and Yield Forecasting

V. B. Virani[1]*and Neeraj Kumar[2]

[1]Agricultural Meteorological Cell, Navsari Agricultural University, Navsari, Gujarat, India
[2]Department of Agronomy, College of Agriculture, Navsari Agricultural University, Bharuch, Gujarat, India

ABSTRACT

This study employs three distinct methods: Stepwise Multiple Linear Regression (SMLR), Ridge Regression, and XG-Boost was used to forecast the sugarcane yield of the Navsari district of Gujarat, India. The integration of meteorological data and remote sensing-derived Vegetation Indices (VIs) is a key component in these approaches. A Sentinel-2 satellite image from May 2023 was utilized for accurate acreage estimation, revealing measuring 8.53% error compared to government data. Ridge regression emerges as the most accurate model for yield forecasting, demonstrating consistency across validation years. The combination of remote sensing data, meteorological data, and machine learning algorithms proves effective in predicting sugarcane yield, offering a cost-effective, time-efficient, and error-free alternative. This approach not only enhances the accuracy of crop yield forecasts but also addresses the challenges associated with traditional methods, such as human error, expense, and time consumption. Overall, this study underscores the effectiveness of remote sensing in conjunction with meteorological data and machine learning for precise and efficient sugarcane yield forecasting, it may provide valuable insights for stakeholders such as policymakers, crop insurance companies, and agro-processing entities. A constraint of this study lies in the presence of cloudy images, especially during the months from June to September. The presence of cloudy conditions introduces contamination, thereby presenting a specific challenge to accurately forecast the yield, particularly for Kharif crops. Another limitation of this study is the low temporal resolution of Landsat satellite imagery, making it challenging to obtain real-time data on crop conditions within very short intervals.

Keywords: *Acreage estimation, Machine learning, Meteorology, Landsat, Sugarcane, Remote sensing, GIS and Yield forecasting.*

INTRODUCTION

Precise prediction of crop yields well in advance of the harvest is of utmost importance, particularly in a country like India, marked by erratic weather patterns. The need for early yield assessment at both regional and national levels is significant for a variety of stakeholders, including agricultural planners, policymakers, crop insurance companies, agro-processing companies (i.e., sugar factories) and the research community [6]. Crop yield estimation relying on traditional methods of data collection through experimentation can be subjective, expensive, time-consuming, and susceptible to substantial errors stemming from incomplete on-ground observations [4]. Remotely sensed data offers significant opportunities for agricultural decision-makers via improved accuracy in crop yield forecasting and rapid crop loss assessments [2]. In recent times, with the successful deployment of a range of remote sensing satellites, such as Landsat, Sentinel, SPOT-VGT and MODIS, remote sensing has emerged as a valuable resource for offering precise, free cost, efficient, and rapid assessment in crop acreage and yield estimation. Moreover, remote sensing-based methods have already demonstrated their effectiveness and success in mapping sugarcane cultivation areas [7] and in predicting sugarcane yield using various vegetation indices (VIs) [1]. This paper investigates sugarcane acreage estimation using sentinel 2 data and develops Landsat 7/8 and meteorological data-based sugarcane yield forecasting model for the Navsari district of Gujarat.

*Corresponding Author: ***V. B. Virani***

DOI: https://doi.org/10.58321/AATCCReview.2024.12.01.133

METHODOLOGY

Study Area

This study focuses on Navsari district in Gujarat, India (Fig. 1), located between 20°32′ and 21°05′ North Latitude and 72°42′ and 73°30′ East Longitude. Situated in the western part of India, the district falls under the South Gujarat heavy rainfall agroclimatic zone. Its total geographical area spans 2204 km^2.

Acreage Estimation

The study employed a Sentinel-2 satellite image from May 2023 due to its high spatial resolution, which allows for precise identification of sugarcane fields. Additionally, the choice of a May month satellite image was deliberate, as minimal crop presence during this period simplifies the classification process. To accurately identify sugarcane fields, the Google Earth Pro software was used to collect 300 sugarcane field ground-truthing points (GTPs) across the study area (Fig. 2). These points were then saved as a shapefile for machine training. The majority (80%) of the collected ground truth points (GTPs) were used to train the model, which allowed for optimal learning and generalization. The remaining 20% of GTPs were reserved for accuracy assessment.

To create a false colour composite (FCC), one to seven bands of sentinel 2 images are used to make a band set in QGIS. After the

creation of a training sample of different classes (water, bare land, sugarcane, orchard, build area, and other crops), maximum likelihood algorithms were employed to classify the whole image into different classes. After classifying, extract the sugarcane class and reclassify this class using a random forest algorithm into binary class (sugarcane vs non-sugarcane pixels). NDVI time series of sugarcane was used to identify the sugarcane and non-sugarcane pixels. After identifying pure sugarcane pixels, the Google satellite base map in QGIS served as a reference layer to digitize each field boundary where pure sugarcane pixels were present. This yielded a shapefile for calculating the total sugarcane area.

Remote Sense Data Preparation

The monthly mean of vegetation indices (VIs) extracted from Landsat 7 and Landsat 8 satellite imagery (refer to Table 1) was computed using Google Earth Engine (GEE) across the time range from 2003 to 2021. Table 1 displays the vegetation indices (VIs) derived from Landsat 7 and 8 satellite images, which were subsequently employed in yield forecasting models.

Meteorological Data Preparation

The weekly average from the 6th SMW (standard meteorological week) to the 42nd SMW was calculated from the daily weather data and then by using this weekly data weighted and unweighted weather indices were computed. For the calculation of weight weather indices, the correlation coefficient between yearly sugarcane crop yield and weather variables for corresponding weeks was used as weight. Table 2 illustrates the unweighted and weighted weather indices utilized in the model development.

- Un-weighted weather indices = Sum (each weekly weather variable)
- Weighted weather indices = Sum (each weekly weather variable x correlation coefficient between yield and particular week weather variable)

Yield Forecasting

Step Wise Multiple Linear Regression

The SMLR (Stepwise Multiple Linear Regression) method was employed as the straightforward approach for the development of the yield forecast model. The yield forecasting model was developed using SPSS software, integration of remote sensing-derived monthly mean of vegetation indices (VIs) and both weighted and unweighted weather indices as independent variables, while sugarcane yield data (kg/ha) served as the dependent variable in the regression model.

Machine learning Algorithms (MLAs)

In this approach, we employed two machine learning algorithms, XG Boost and Ridge Regression. XG Boost is favoured for its exceptional execution speed and model performance [3]. Ridge regression serves as a method to mitigate overfitting in data by introducing a slight degree of bias to the regression estimates [5] and it is particularly useful when dealing with multicollinearity issues in the input features. Both models were constructed in a Python environment using the sci-kit-learn library. The training process involved allocating 80% of the datasets and reserving the remaining 20% for testing. Additionally, a three-year holdout dataset spanning 2019 to 2021 was set aside for model validation.

Fig. 3 illustrates the methodology flow chart for acreage estimation and yield forecasting.

Model Evaluation

The criteria employed to determine the model performance included statistical metrics such as R^2, root mean square error (RMSE), mean absolute percentage error (MAPE), percentage accuracy (PA) and mean absolute error (MAE). These statistical metrics were calculated in the Agricultural and Meteorological Software website (https://agrimetsoft.com/calculators/).

RESULTS AND DISCUSSION

Validation of Sugarcane Acreage Estimation

The preliminary assessment for the sugarcane area of Navsari district in 2024 yielded a figure of 14,121 hectares, with an overall accuracy of 86% for the sugarcane class. In comparison, the average sugarcane area over the past five years (2017-2021), as reported by the Directorate of Agriculture (Government of Gujarat) was 15,438 hectares. The percentage error between the two estimates is 8.53%, indicating a favourable alignment between the remote sensing and government estimates for the sugarcane area. Fig. 4 illustrates digitized boundaries delineating sugarcane field areas of Navsari district.

Yield forecasting

Stepwise Multiple Linear Regression

In the validation phase of Stepwise Multiple Linear Regression (SMLR), the three-year holdout data exhibited an RMSE value of 5.08 t/ha, an MAE value of 4.50 t/ha, and a MAPE value of 6.4% (refer to Table 3). For the last year of holdout data in 2021, the predicted yield was 70,909 kg/ha, while the observed yield was 74,652 kg/ha. Analysing the regression models (see Table 4), the highest R^2 value reached 0.986 with an adjusted R^2 value of 0.975. In contrast, the lowest R^2 was 0.70, attributed to considering only Vegetation Indices (VIs) in this model. The elevated R^2 value was attributed to the integration of VIs with meteorological data.

Ridge regression and XG-Boost

The sugarcane yield forecasting model, developed using ridge regression and XG-Boost, demonstrated strong performance as outlined in Table 3. During validation, the three-year holdout data showcased an RMSE value of 3.86 t/ha, an MAE value of 2.41 t/ha, and a MAPE value of 3.3%. Specifically for the last year of holdout data in 2021, ridge regression predicted a yield of 81,328 kg/ha, while the observed yield stood at 74,653 kg/ha. Comparing the performance across SMLR, ridge regression, and XG Boost, ridge regression consistently exhibited high accuracy across holdout years. In summary, among the three models, ridge regression emerges as a robust choice for sugarcane yield forecasting. During the validation stage of XG-Boost, three-year holdout data show an RMSE value of 6.50 t/ha, an MAE value of 6.16 t/ha, and a MAPE value of 8.9%. XG-Boost regression predicted a yield of 70,090 kg/ha for the year of 2021.

CONCLUSION

In this study, three distinct methods, namely SMLR (Stepwise Multiple Linear Regression), ridge regression, and XG-Boost, were employed to forecast sugarcane yield in Navsari district. The integration of meteorological data and remote sensing-derived Vegetation Indices (VIs) was a key aspect in all three approaches. The results showed that ridge regression predicts accurate yield as compared to others. The remote sensing technique employed for acreage estimation demonstrated a precise estimate, exhibiting an error rate of 8.53% when

compared to the data released by the government of Gujarat. Integrating remote sensing data with meteorological information and machine learning algorithms resulted in a robust prediction of sugarcane yield. In summary, the remote sensing approach for acreage and yield forecasting proves cost-effective, and time-saving and eliminates the potential for human errors.

Future scope: In the current investigation, Landsat data with a temporal resolution of 16 days and a spatial resolution of 30 meters were employed. However, for future research, we plan to transition to Sentinel-2 data, which offers a more frequent temporal resolution of 5 days and improved spatial resolutions of 10 meters and 20 meters. This shift is intended to enhance precision and enable the acquisition of more real-time data. Utilizing machine learning algorithms, the integration of Sentinel-2 data with meteorological information enhances the robustness of yield prediction. Alternatively, an alternative approach involves leveraging various MODIS product data in conjunction with meteorological data for the same purpose.

Conflict of interest: The authors have declared that no conflict of interest exists.

Acknowledgement: The authors are grateful to USGS, ESA, and Google Earth Engine for providing free data. Additionally, thanks to the Department of Agrometeorology, NAU, Navsari, Gujarat for furnishing essential weather data. Their contributions have been instrumental in the successful conduct of this research.

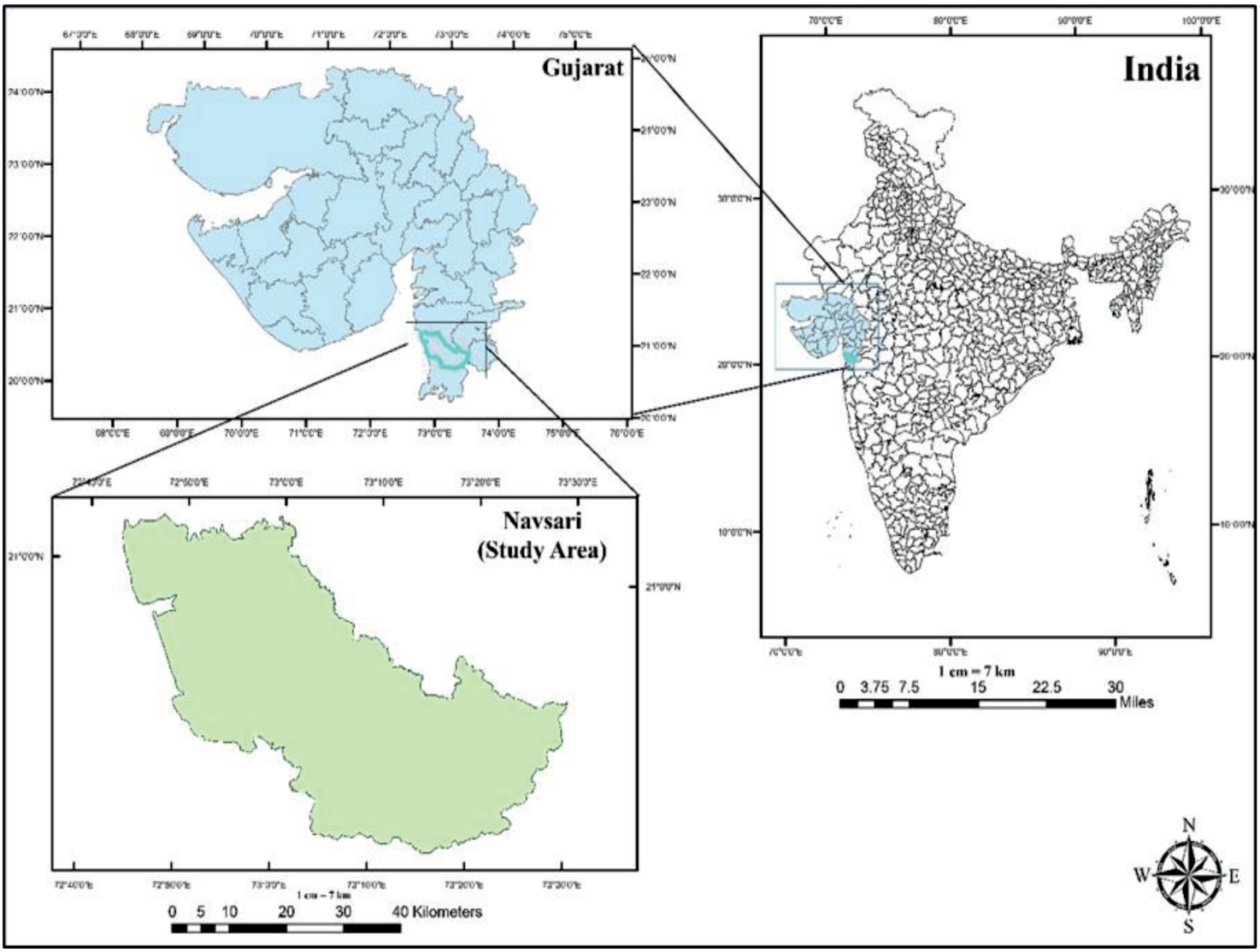

Figure 1: Study Area

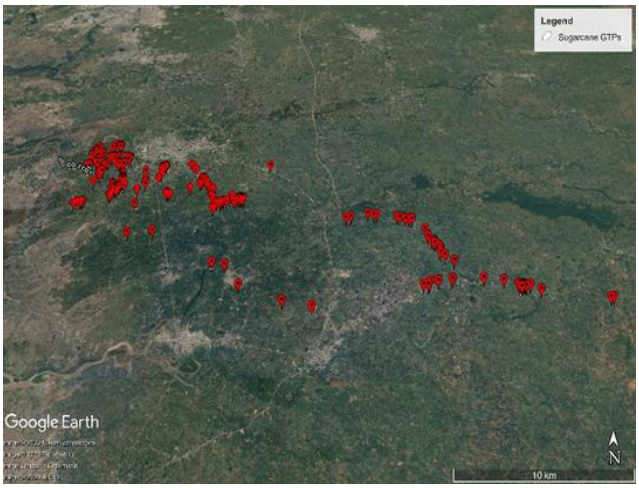

Figure 2: Ground truthing points (GTPs) of sugarcane field across study area

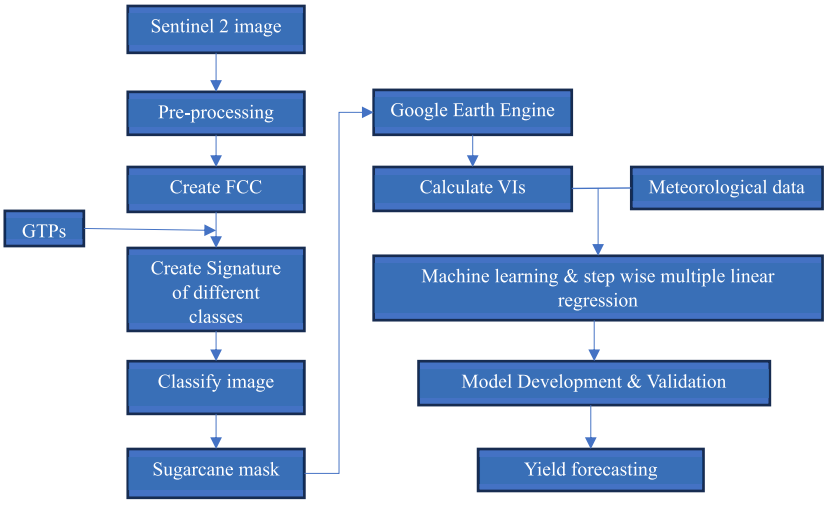

Figure 3: Methodology flow chart of sugarcane acreage estimation and yield forecasting

Table 1: Vegetation indices (VIs) used in sugarcane yield fore casting

Vegetation Indices (VIs)	Formula	Description
Green normalized difference vegetation index (GNDVI)	GNDVI = (NIR+Green)/(NIR−Green) **For Landsat 8 =** (Band5+Band3)/(Band5−Band3)	• Higher GNDVI values typically indicate healthier and more dense vegetation, while lower values may suggest stressed or sparse vegetation. • Range: -1 to +1
Simple Ratio (SR)	SR = NIR/RED **For Landsat 8 =** Band5/Band4	• Higher values of the Simple Ratio generally indicate healthier and more abundant vegetation.
Enhanced Vegetation Index (EVI)	$EVI = 2.5 \times \frac{NIR-RED}{NIR+6 \times RED-7.5 \times Blue+1}$ **For Landsat 8 =** $2.5 \times \frac{Band5-Band4}{Band5+6 \times Band4-7.5 \times Band2+1}$	• It was formulated to enhance the shortcomings of NDVI by refining the vegetation signal, especially in regions characterized by a high leaf area index (LAI).
Normalized Difference Moisture Index (NDMI)	NDMI = (NIR+SWIR)/(NIR−SWIR) **For Landsat 8 =** (Band5+Band6)/(Band5−Band6)	• NDMI is particularly effective in assessing changes in vegetation water content, making it valuable for applications like drought monitoring and irrigation management.

Table 2: Unweighted and weighted weather indices for models development

Parameter	Unweighted Weather Indices						Weighted Weather Indices					
	Tmax	Tmin	RF	RH-I	RH-II	BSSH	Tmax	Tmin	RF	RH-I	RH-II	BSSH
Tmax	Z10						Z11					
Tmin	Z120	Z20					Z121	Z21				
RF (rainfall)	Z130	Z230	Z30				Z131	Z231	Z31			
$RH\text{-}I_{morning}$	Z140	Z240	Z340	Z40			Z141	Z241	Z341	Z41		
$RH\text{-}II_{afternoon}$	Z150	Z250	Z350	Z450	Z50		Z151	Z251	Z351	Z451	Z51	
BSSH	Z160	Z260	Z360	Z460	Z560	Z60	Z161	Z261	Z361	Z461	Z561	Z61

Table 3: Models evaluation

Algorithms	Accuracy (%)			RMSE (t/ha)	MAE (t/ha)	MAPE (%)
	2019	2020	2021			
XG-Booster	92.35	87.16	93.88	6.50	6.16	8.9
Ridge Regression	**98.14**	**96.12**	**91.04**	**3.86**	**2.41**	**3.3**
SMLR	96.67	89.10	94.99	5.08	4.50	6.4

Table 4: Regression models develop using SMLR using integrated VIs and meteorological data.

Models	R^2	Adjusted R^2
Y = 53497.601 + ($GNDVI_{octo.}$ * 42284.073) + ($SR_{octo.}$ * -4668.67) + ($NDMI_{may}$ * 52832.722) + (Z61 * 635.776) + ($SR_{nov.}$ * -2069.585) + ($NDMI_{oct.}$ * 7944.212) + (Z361 * 1.908) + ($NDMI_{sept}$ * -4534.329)	0.986	0.975
Y = 50429.419 + ($GNDVI_{octo.}$ * 48833.775) + ($SR_{octo.}$ * -5853.34) + ($NDMI_{may}$ * 38044.148) + ($SR_{nov.}$ * -1783.850) + (EVI_{oct} * 11445.829) + (Z61 * 712.07)	0.969	0.954
Y = -18363.969 + ($GNDVI_{octo.}$ * 59912.629) + (Z150 * 0.542) + (Z121 * 15.485) + (Z241 * -3.450)	0.907	0.881
Y = 54384.943 + ($GNDVI_{octo.}$ *55889.658) + ($SR_{octo.}$ * -4358.044) + ($NDMI_{may}$* 26180.0)	0.890	0.868
Y = 52212.454 + ($GNDVI_{octo.}$ *56543.547) + ($SR_{octo.}$ * -5274.356) + ($EVI_{octo.}$ * 18463.546)	0.884	0.861
Y = 13865.278 + ($GNDVI_{octo.}$ * 52867.184) + (Z141 * 2.934)	0.815	0.792
Y = 45073.16 + ($GNDVI_{octo.}$ *58005.173)	0.700	0.683

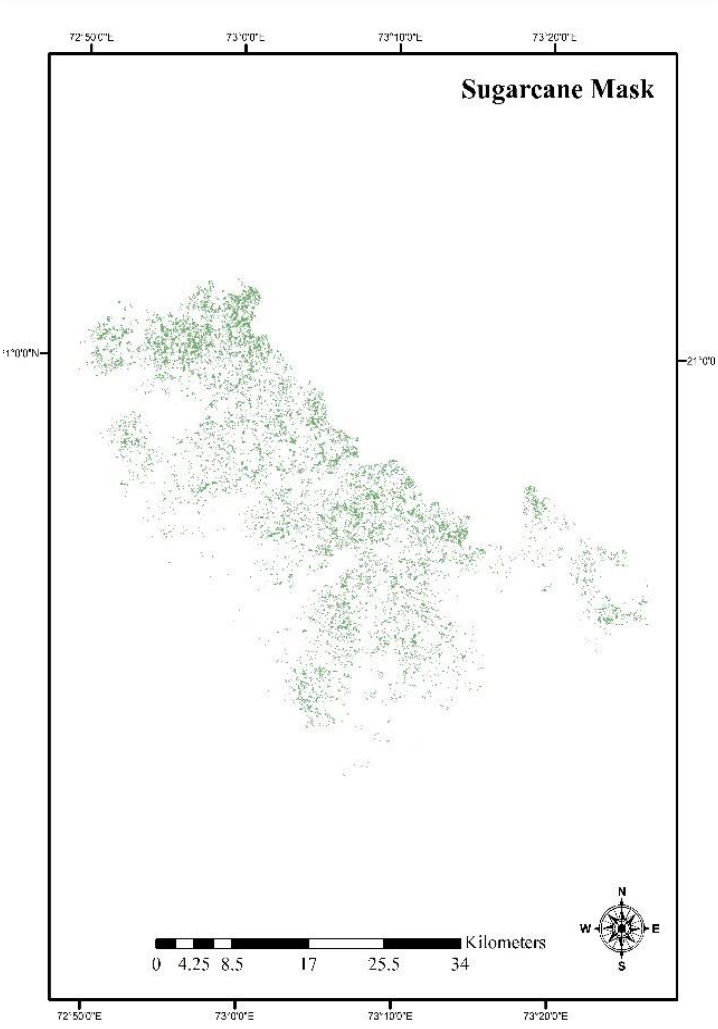

Figure 4: Sugarcane area of Navsari district in 2024

REFERENCES

1. Dubey S K, Gavli A S, Yadav S K, Sehgal S and Ray S S (2018) Remote sensing-based yield forecasting for sugarcane (*Saccharum officinarum* L.) crop in India. *J. Indian Soc. Remote Sens.*, 46: 1823-1833.

2. Lobell D B (2013) The use of satellite data for crop yield gap analysis. *Field Crops Res.*, 143: 56–64.

3. Ravi R and Baranidharan B (2020) Crop yield Prediction using XG Boost algorithm. *Int. J. Recent Technol. Eng*, 8(5): 3516-3520.

4. Sapkota T B, Jat M L, Jat R K, Kapoor P and Stirling C (2016) Yield estimation of food and non-food crops in smallholder production systems. Methods for measuring greenhouse gas balances and evaluating mitigation options in smallholder agriculture, 163-174.

5. Setiya P, Satpathi A, Nain A S and Das B (2022) Comparison of weather-based wheat yield forecasting models for different districts of Uttarakhand using statistical and machine learning techniques. *J. Agrometeorol.*, 24(3): 255-261.

6. Van Wart J, Kersebaum K C, Peng S, Milner M and Cassman K G (2013) Estimating crop yield potential at regional to national scales. *Field Crops Res.*, 143: 34–43.

7. Yedage A S, Gavali R S and Patil R R (2013) Remote sensing and GIS base crop acreage estimation of the sugarcane for Solapur district, Maharashtra. *Golden Res. Thoughts*, 2(11): 1-12.