SOYBEAN CROP YIELD PREDICTION BY INTEGRATION OF REMOTE SENSING AND WEATHER OBSERVATIONS

J.D.Mohite[1,*], S.A.Sawant[2], A.Pandit[3], R.Agrawal[3], S. Pappula[4]

[1] TCS Research and Innovation, Tata Consultancy Services, Mumbai, India - jayant.mohite@tcs.com
[2] TCS Research and Innovation, Tata Consultancy Services, Pune, India - suryakant.sawant@tcs.com
[3] TCS Research and Innovation, Tata Consultancy Services, Indore, India - (ankur.pandit, rishabh.agrawal3)@tcs.com
[4] TCS Research and Innovation, Tata Consultancy Services, Hyderabad, India - srinivasu.p@tcs.com

KEY WORDS: Yield Forecasting, Soybean Crop, Remote Sensing, Weather Data, Random Forest Regression

ABSTRACT:

The main objective of this study is the in-season forecasting of soybean crop yield using the integration of satellite remote sensing and weather observations. The study was carried out in the Paraná state of Brazil. The soybean crop in the study region is sown during Oct.–Nov. month and harvested between Feb.–Mar. of the next year. Municipality-level soybean yield data for 15 municipalities was obtained from the AGROLINK portal of Brazil, from the 2005–06 season to the 2020–21 season. The crop yield data constituted yearly municipality-wise yield in kg/ha. Remote sensing-based indicators such as the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) and Land Surface Temperature (LST), and Rainfall data from CHIRPS was considered in the study. Regression modelling was carried out between municipality-level yield as the dependent variable and features generated from remote sensing and weather observations as independent variables. Performance evaluation of tuned random forest regression (RFR) and tuned support vector regression (SVR) were performed against multiple linear regression (MLR). A comparison of results in terms of algorithms shows that RFR performed better than SVR and MLR. Further, a root-mean-square-error (RMSE) of 414 kg/ha and an R^2 value of 0.748 were achieved by the best RFR model. Validation of developed RFR model was performed on the data from the new soybean season, i.e., 2020–21. We have achieved an R^2 value of 0.693 with a RMSE of 585 kg/ha. Although the model performance on the data of 2020-21 season is slightly reduced, R^2 and RMSE are in good agreement with test results. This study showed that, integration of remote sensing and weather observations would be useful for in-season yield forecasting of soybean at municipality level.

1. INTRODUCTION

Soybean (Glycine max (L.) Merrill) is one of the most important sources of protein and oil. United States, Brazil, Argentina, and China account for around 90% of the global soybean production (Embrapa, 2018, USDA, 2019). Brazil is the second largest producer of soybean in the world, after the United States, and the crop is grown in almost all of the country’s states. The production of soybean in Brazil has shown a steady increase over the years, driven by factors such as technological advancements, favourable weather conditions, and government policies that promote agricultural development. According to data from the Brazilian National Company for Food Supply Companhia Nacional de Abastecimento (CONAB), Brazil’s soybean cultivation area has increased from around 15 million hectares in year 2000 to more than 38 million hectares in the 2020-2021 harvest season (CONAB, 2021). The majority of soybean cultivation in Brazil is concentrated in the southern and central regions of the country, which have favourable soil and weather conditions for the crop. Parana is the second largest producer of soybeans in Brazil, accounting for approximately 19% of the country’s total soybean production in 2020-21 harvest season. The largest producer is Mato Grosso, which accounts for approximately 28% of the country’s total soybean production (CONAB, 2021). The Parana state has favourable soil and weather conditions for soybean production, and has invested heavily in research and development of the crop, including the development of high-yielding soybean varieties and advanced production techniques. The area under soybean cultivation in Parana has also increased significantly over the years, from around 2 million hectares in year 2000 to more than 5 million hectares in year 2021 (CONAB, 2021).

* Corresponding author

Accurate and timely statistics on soybean crop area and yield in Brazil and Parana are crucial for crop management, food security, and economic planning. Many approaches have been developed in the past for in-season yield assessment ranging from field based survey, physical process based models (Jones et al., 2003), use of non-invasive technologies, etc. The Brazilian government collects and publishes regular data on soybean production, including crop area, yield, and production volume, which are widely used by farmers, policymakers, and analysts. However, manually collecting the data in yield is costly and time consuming process. Alternatively, crop cutting experiments are being conducted by various state and private agencies, which have limited scalability and involves huge cost and time (Mohite et al., 2019).

Over the years, there has been significant progress in the use and development of physical process-based crop growth simulation models. Few examples of crop growth simulation models being used for soybean yield prediction include Agricultural Production System Simulator (APSIM) (Keating et al., 2003), Decision Support System for Agro-Technology Transfer (DSSAT) (Jones et al., 2003), SoySim (Wilhelm et al., 2004), STICS (Brisson et al., 1998). Crop growth simulation models basically uses mathematical equations to simulate the growth and development of crops and used to assess the response of various management practices on yield. Although, physical process based models are quite accurate, and a large amount of ground data is needed to run the simulations. Model performance is largely affected by the quality and quantity of ground data available.

On the other hand, non-invasive technologies mainly involve the use of remote sensing observations in machine learning based models for in-season yield forecast (Peralta et al., 2016, Schwalbert et al., 2018, Xia et al., 2020, Li et al., 2020). With the advent of global coverage and openly available remote sensing satellite data in Optical and Synthetic Aperture Radar electromagnetic spectrum at regular intervals. Thereby, resulted into ease in accessing the time series of satellite data and couple machine learning algorithms to provide value added insights into the agriculture sector. Additionally the compute platform such as Google Earth Engine (GEE) (Gorelick et al., 2017) facilitated the spatio-temporal scalability. Remote sensing data can be used to estimate various crop-related parameters, such as vegetation indices, surface temperature, and soil moisture, which are closely related to crop growth and yield. These parameters can be used to develop models for predicting crop yield. In recent years, several studies have been conducted on soybean crop yield prediction using remote sensing data in Brazil. These studies have focused on different aspects of the problem, such as the choice of remote sensing data sources, the development of appropriate models, and the validation of the results. Moreover, integration of weather data along with remote sensing observations has proven effective in improving the model's performance (Peng et al., 2018, Cai et al., 2019). Considering the advantages and limitations of the prior art, we have attempted to use the integration of satellite and weather data for in-season yield forecasting.

The main objective of this study is the in-season forecasting of soybean crop yield using the integration of satellite remote sensing and weather observations. The study wasconducted in selected municipalities of the Paraná state of Brazil.

2. MATERIALS AND METHODS

This section covers information about the study area, insights into the satellite and ground reference data, and a detailed approach for yield prediction using various machine learning algorithms.

2.1 Study Area

We have conducted this study in selected municipalities in Parana State, Brazil 1. Parana is one of the southern states of Brazil, bordered on the north by Sao Paulo state, on the east by the Atlantic Ocean, on the south by Santa Catarina state and the province of Misiones, Argentina, and on the west by Mato Grosso do Sul and Paraguay, with the Parana River as its western boundary line (Chisholm, 1911). The annual mean air temperature ranges between 15 and 24 °C, with the highest temperatures found in the northwest and the lowest around Palmas (Aparecido et al., 2016). Precipitation is less than 1,200 mm (47 in) a year in the north of the state, rising to above 1,800 mm (71 in) in the southwest and southeast of the state (Aparecido et al., 2016). Agriculture is one of the main economic drivers of the state, and about 15% of Parana's GDP comes from agriculture. Parana is a leading producer of crops such as soybeans, corn, coffee, and wheat.

2.2 Satellite Datasets Used

We have mainly used satellite data products from the Moderate Resolution Imaging Spectroradiometer (MODIS) and rainfall data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS).

2.2.1 MODIS Based Vegetation Indices MOD09GA and MYD09GA are two products of the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors onboard the Terra and Aqua satellites, respectively (Vermote and Wolfe, 2015). These products provide daily surface reflectance data at a 1-kilometer spatial resolution. The data are corrected for atmospheric effects and provide a consistent surface reflectance time series over the lifetime of the MODIS sensors. Version 6 of the products include improvements such as better cloud and cloud shadow screening, and updated aerosol retrieval algorithms. The surface reflectance bands, such as Red, NIR were used to estimate the Normalized Difference Vegetation Index (NDVI), However, surface reflectance in blue, red and NIR bands was used to estimate Enhanced Vegetation Index (EVI). The data was accessed from the GEE during the Soybean crop season i.e. for the months November, December and January between 2005-2021. Monthly mean composites at 1 km spatial resolution were generated and used in the analysis.

2.2.2 MODIS Based Land Surface Temperature We have utilized land surface temperature data from MODIS (MOD11A1 and MYD11A1 version 6 products) (Wan, 2015). The MODIS Land Surface Temperature (LST) product is a remotely sensed dataset that provides information about the temperature of the Earth's surface (Wan et al., 2002). These products are generated using a split-window algorithm that utilizes the 3.7 μm and 11 μm thermal bands to retrieve LST. The data is available at a daily temporal scale and 1 km spatial resolution during daytime and nighttime modes. We have utilized the data available in the daytime mode. LST data was accessed from GEE for the months of November, December, and January between 2005 and 2021 and used in the analysis.

2.2.3 Rainfall Data from CHIRPS The Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) daily rainfall product is a remotely sensed dataset that provides information about daily rainfall amounts across the globe. The product is derived from a combination of satellite data and ground station observations (Funk et al., 2015) and has a 30+ year quasi-global coverage. The product provides daily global coverage of rainfall amounts at a spatial resolution of 0.05 degrees, or approximately 5 km. We created monthly accumulated composites from the daily data and resampled it to 1 km spatial resolution to match it with MODIS-based vegetation indices and LST. We accessed the data for the months of November, December, and January between 2005 and 2021 from the GEE.

2.3 Ground Reference Data on Yield

AGROLINK is a web portal focused on the agricultural sector in Brazil. The portal was created in year 1996 and provides a wide range of information and services related to agriculture, livestock, agribusiness, and rural development (AGROLINK, 2019). AGROLINK offers a variety of content, including news, articles, interviews, market analyses, and technical information. The portal covers a wide range of topics, such as crop production, livestock management, pest and disease control, irrigation, precision agriculture, biotechnology, sustainability, and rural

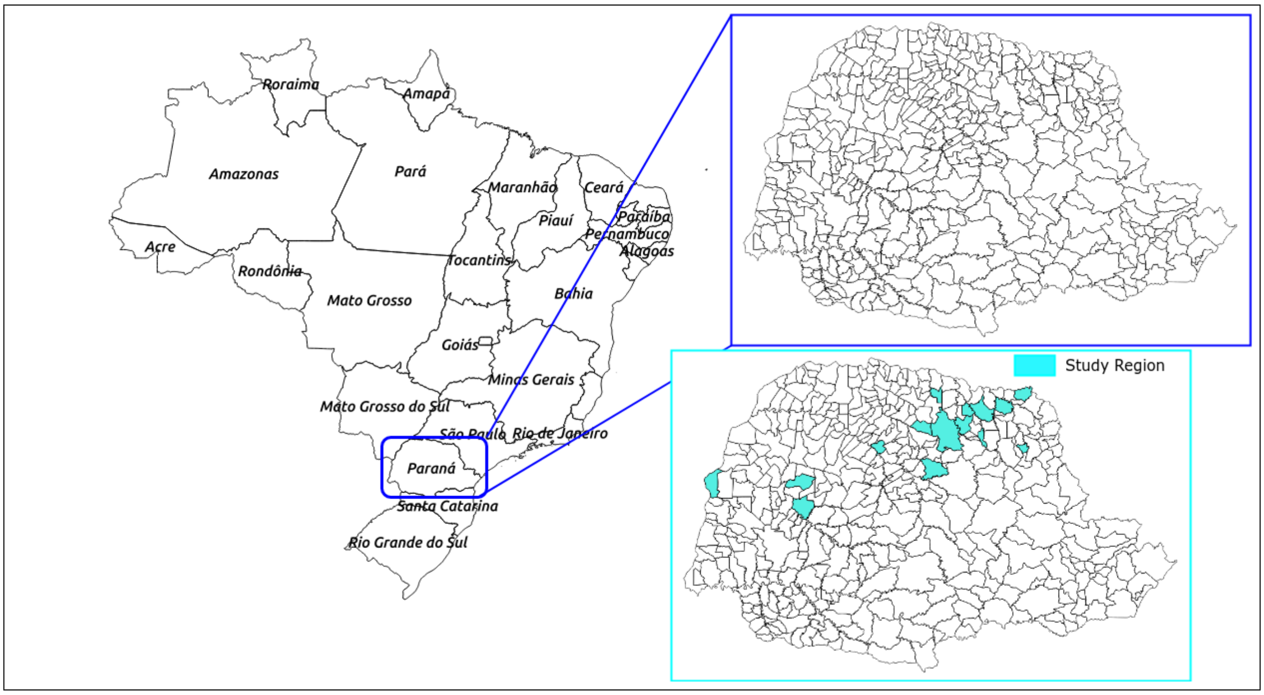

Figure 1. Study area

policies. AGROLINK provides crop yield data at the municipality level for various agricultural crops for various historical years. We accessed the yield data for the soybean crop for 15 municipalities in the state of Parana, Brazil, for the soybean crop seasons from 2005-06 to 2020-21. The data was represented as the yield value per hectare per municipality for specific year.

2.4 Overall Framework

We obtained municipality-level soybean yield data for 15 municipalities from AGROLINK (AGROLINK, 2019) for the soybean crop seasons from 2005-06 to 2020-21. The data was represented as yield value per hectare per municipality for specific year and used as ground truth for modeling purposes. For the municipality-level analysis, we considered observations from MODIS satellite and remote sensing-based indicators, such as Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Land Surface Temperature (LST), as well as rainfall data from CHIRPS. For NDVI and EVI, we considered the municipality-level maximum, minimum, and average values for the period from 1 Nov. to 31 Jan.. For LST, we estimated the municipality-level maximum, minimum, and average values for the months of November, December, and January, and accumulated rainfall for Nov., Dec., and Jan. was considered features in the study. We utilized a total of 18 features, including 3 for NDVI, 3 for EVI, 9 for LST, and 3 for rainfall. The dataset contained 248 samples from various municipalities and seasons. Regression modeling was performed between municipality-level yield and features derived from remote sensing and weather observations. Municipality-level yield data was considered as the dependent variable, while features derived from MODIS and CHIRPS data were used as independent variables in the regression analysis. The modeling approach involved dividing the data into training (80%) and validation (20%) sets. The performance of tuned Random Forest Regression (RFR) and tuned Support Vector Regression (SVR) was evaluated against the Multiple Linear Regression (MLR). Finally, the best model was used to generate yield maps for the study region.

2.5 Machine Learning Algorithms used in the study

We mainly used three different Machine Learning algorithms, namely Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Support Vector Regression (SVR).

2.5.1 Multiple Linear Regression MLR also known as Multiple regression is a statistical method that uses multiple explanatory variables to predict the outcome (Stepanov et al., 2020). MLR models the linear relationship between explanatory variables and target variables. We used MLR as the baseline regression model in our study, where NDVI, EVI, LST, and rainfall were used as explanatory variables with yield as the target variable.

2.5.2 Random Forest Regression Random Forest is a well-known machine learning algorithm used for ensemble learning, which combines multiple decision trees to make more accurate predictions. This algorithm generates a large number of decision trees from a subset of random training data, and the output from individual trees is used to make the final decision (Breiman, 1996). The method is widely used for various remote sensing applications, including crop yield prediction, due to its capability to resist over-fitting and require less computational time (Inglada et al., 2015, Belgiu and Csillik, 2018, Were et al., 2015). However, like many other algorithms, RFs are sensitive to the choice of hyperparameters and the training data used. In this study, we followed a K-fold cross-validated grid search approach. In this approach, the dataset was first randomly split into a training set (80% of the total dataset) and a test set (20% of the total dataset). The test set was kept for final evaluation, while the training set was further split into three subfolds. The model was iteratively trained on K-1 (2) of the folds and validated on the remaining one to optimize the performance by tuning RF hyperparameters, such as the number of trees and minimum number of variables available at each split.

2.5.3 Support Vector Regression SVR is a type of supervised learning that can handle non-linear relationships between variables and high-dimensional datasets, making it a powerful tool for yield prediction. SVR has been extensively used for various agricultural applications (Khosla et al., 2020, Anandhi and Chezian, 2013, Brdar et al., 2011). Similar to other ML algorithms, the performance of SVR could be improved by having the optimum choice of hyperparameters. In this study, we have tuned the SVR for two hyperparameters, namely C and Sigma. The train-test splitting and modeling strategy were kept similar to RFR.

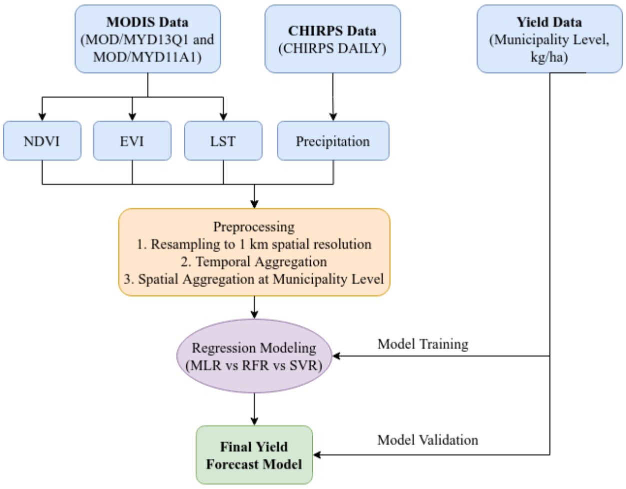

Figure 2. Overall analysis approach

3. RESULTS AND DISCUSSION

This section describes the performance of various ML models, such as MLR, RFR, and SVR. Additionally, we have also discussed the predicted versus actual yields at the municipality level for a new season.

3.1 Implementation of Various ML Models

We mainly used three algorithms in our study: MLR, RFR, and SVR. There are no hyperparameters for MLR, but algorithms such as RFR and SVR are sensitive to the choice of hyperparameters and training data. The right selection of these hyperparameters is crucial for optimal results. For hyperparameter selection, we followed a 3-fold cross-validated grid search strategy for RFR and SVR. In this approach, the dataset was first randomly split into a training set (80% of the total dataset) and a test set (20% of the total dataset). The test set was kept for final evaluation, while the training set was further split into 3 sub-folds. The model was iteratively trained on K-1 (2) of the folds and validated on the remaining one to optimize the performance by tuning hyperparameters. RFs were tuned for two parameters: the number of trees (ntree) and the minimum number of variables available at each split (mtry). We tried ntree values between 50-400 with an interval of 50 (8 ntree values) and mtry between 2-8 (7 mtry values) with an interval of 1. A total of 56 models were trained, considering each combination of ntree x mtry, and the best model was considered for testing on the remaining 20% data. SVR was tuned for two hyperparameters: C and Sigma. We tried C values of 0.1, 1, 10, 100 and Sigma values of 0.001, 0.01, 0.1, and 1. Therefore, 16 models were trained, considering each combination of C x Sigma, and the best model was considered for testing on the remaining 20% data. In the case of MLR, we used all 80% of the data for model training and the remaining data for testing of the model. Finally, we chose the best model from MLR, RFR, and SVR, predicted the yield for the season 2020-21, which was not considered in the modeling process, and validated our results.

Table 1. Performance of MLR

SN	TrainSet	R^2	RMSE
1	TrainSet 1	0.578	822
2	TrainSet 2	0.545	854
3	TrainSet 3	0.549	853

3.2 Performance of Various ML Models

Table 1 shows the performance of MLR for three different simulations (training sets). Since the training data was randomly selected, we ran the models three times to assess the performance of MLR with changing training data. The average R^2 value was found to be 0.557 and RMSE of 843 kg/ha with a standard deviation of 0.018, indicating that the model is not sensitive to training data and performed equally well when the training data was changed. However, the overall ^{2} value across all the models is relatively lower.

Tables 2 and 3 show the performance of the RFR model in terms of R^2, RMSE for various combinations of ntree and mtry. The results show that the best RFR model achieved a Root Mean Square Error of 414 kg/ha and an R^2 value of 0.748, which was obtained for a model trained using ntree=200 and mtry=5. It can be observed that as the number of trees increased, there was a steady increase in R^2 from 50 to 350 for mtry values of 2-4. However, for mtry values between 5-8, there was an increase in R^2 only for ntree values from 50-200, and R^2 either reduced or remained constant after an ntree of 200. Comparing MLR and RFR, it can be seen that there was a significant improvement in R^2 for all the combinations of ntree x mtry.

Table 4 and Table 4 show the performance of the SVR model for various combinations of C and Sigma. The best R^2 value of 0.722 was achieved for a C value of 10 and a Sigma value of 0.01. For the same combination of C and Sigma, the RMSE reached up to 502 kg/ha. Furthermore, there was a significant improvement in performance compared to MLR; however, the results were not as good as RFR.

Finally, we chose the best RFR model (trained with ntree=200 and mtry=5) and applied it to the data from the 2020-21 season for yield prediction. We estimated the R^2 and RMSE between the predicted yield and actual yields for the season provided by AGROLINK. We achieved an R^2 value of 0.693 with an RMSE of 585 kg/ha. Although the model's performance on the data from the 2020-21 season is slightly reduced, the R^2 and RMSE are in good agreement with the test results. The figure 3 shows the percent change between the ground reference yield and predicted yield at the municipality level. We can observe that half of the municipalities are within a 10% deviation. However, remaining half of the municipalities are within a 15% deviation, and for the one municipality, the deviation goes beyond 15%.

4. SUMMARY AND CONCLUSIONS

We have developed in-season yield forecasting models for soybean crop in the Parana state of Brazil. The models were developed using historical municipality level yield data from the

https://doi.org/10.5194/isprs-archives-XLVIII-M-1-2023-197-2023 | © Author(s) 2023. CC BY 4.0 License.

Table 2. Performance of RFR in terms of R^2

ntree ↓ \ mtry →	2	3	4	5	6	7	8
50	0.63	0.644	0.667	0.68	0.678	0.634	0.661
100	0.63	0.661	0.678	0.684	0.689	0.699	0.699
150	0.645	0.673	0.671	0.691	0.711	0.727	0.7
200	0.68	0.701	0.712	**0.748**	0.733	0.73	0.732
250	0.69	0.691	0.718	0.744	0.733	0.738	0.74
300	0.716	0.72	0.733	0.739	0.727	0.744	0.73
350	0.722	0.733	0.74	0.741	0.734	0.742	0.74
400	0.689	0.69	0.692	0.721	0.704	0.712	0.698

Table 3. Performance of RFR in terms of RMSE

ntree ↓ \ mtry →	2	3	4	5	6	7	8
50	731	722	700	678	623	710	723
100	701	689	659	601	600	621	700
150	700	677	622	507	588	521	631
200	685	612	572	**414**	522	471	552
250	633	606	512	434	462	464	498
300	544	576	473	455	445	430	477
350	498	502	444	471	438	444	456
400	512	555	502	498	476	465	578

Table 4. Performance of SVR in terms of R^2

C ↓ \ Sigma →	0.001	0.01	0.1	1
0.1	0.645	0.666	0.665	0.678
1	0.689	0.7	0.7	0.7
10	0.712	0.722	0.71	0.71
100	0.682 3	0.655	0.642	0.642

Table 5. Performance of SVR in terms of RMSE

C ↓ \ Sigma →	0.001	0.01	0.1	1
0.1	722	687	555	598
1	643	600	515	543
10	599	502	509	522
100	576	544	536	519

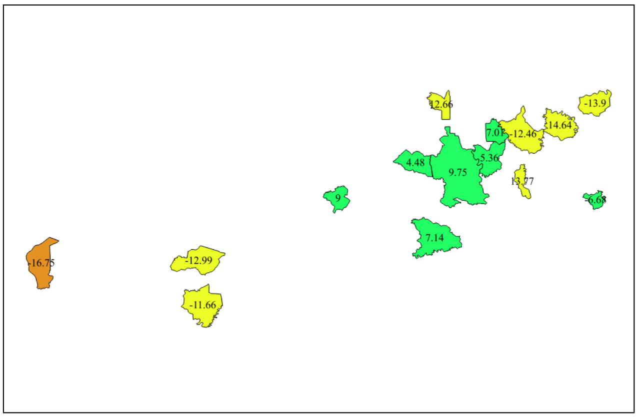

Figure 3. Percent deviation between actual and predicted yield at municipality level

AGROLINK portal of Brazil, as well as remote sensing and weather-based indices. We used regression modeling with municipality level yield data as the dependent variable and remote sensing and weather indices as independent variables. The performance of tuned Random Forest Regression (RFR) and tuned Support Vector Regression (SVR) was evaluated against the Multiple Linear Regression (MLR). The results showed that RFR outperformed SVR and MLR, achieving a Root Mean Square Error of 414 kg/ha and an R^2 value of 0.748 for the best RFR model trained with ntree=200 and mtry=5. We validated the developed RFR model with data from the 2020-21 soybean season, achieving an R^2 value of 0.693 with an RMSE of 585 kg/ha. Although the model performance on the data from the 2020-21 season was slightly reduced, the R^2 and RMSE were in good agreement with the test results. This study showed that the integration of remote sensing and weather observations would be useful for in-season yield forecasting of soybean at the municipality level.

5. FUTURE WORK

In the existing work, we have considered municipality-level data from 15 municipalities of Parana state in Brazil and used well-known machine learning algorithms such as MLR, RFR, SVR. In the future, we would like to assess the performance of deep learning algorithms by adding data from other municipalities in Parana and other states of Brazil. Additionally, we would like to validate the developed approach for upcoming seasons.

REFERENCES

AGROLINK, 2019. AGROLINK - O Portal do conteúdo Agropecuário. https://www.agrolink.com.br/.

Anandhi, V., Chezian, R. M., 2013. Support vector regression in forecasting. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(10), 4148-4151.

Aparecido, L. E. d. O., Rolim, G. d. S., Richetti, J., Souza, P. S. d., Johann, J. A., 2016. Koppen, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Parana, Brazil. *Ciencia e Agrotecnologia*, 40(4), 405-417.

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote sensing of environment*, 204, 509–523.

Brdar, S., Culibrk, D., Marinkovic, B., Crnobarac, J., Crnojevic, V., 2011. Support vector machines with features contribution analysis for agricultural yield prediction. *Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment (EcoSense 2011), Belgrade, Serbia*, 43–47.

Breiman, L., 1996. Bagging predictors. *Machine learning*, 24, 123–140.

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M.-H., Ruget, F., Nicoullaud, B., Gate, P., Devienne-Barret, F., Antonioletti, R., Durr, C. et al., 1998. STICS: A generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. *Agronomie*, 18(5-6), 311–346.

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L. et al., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and forest meteorology*, 274, 144–159.

Chisholm, H., 1911. Paraná. *Encyclopædia Britannica*, 20, Cambridge University Press, 766.

CONAB, 2021. Acompanhamento da safra brasileira de grãos: Safra 2020/2021 - décimo levantamento. Retrieved from http://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos.

Embrapa, 2018. Soja em números. https://www.embrapa.br/soja/cultivos/soja1/ dados-economicos.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. *Scientific data*, 2, 150066.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202, 18–27.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P. et al., 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9), 12356–12379.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L., Wilkens, P. W., Singh, U., Gijsman, A. J., Ritchie, J. T., 2003. The DSSAT cropping system model. *European journal of agronomy*, 18(3-4), 235–265.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N., Meinke, H., Hochman, Z. et al., 2003. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, 18(3-4), 267–288.

Khosla, E., Dharavath, R., Priya, R., 2020. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability*, 22, 5687–5708.

Li, Y., Wang, X., Li, J., Wu, B., Zhang, C., 2020. Soybean Yield Prediction Using Remote Sensing Data and Machine Learning Algorithms in Northeast China. *Sensors*, 20(12), 3522.

Mohite, J., Sawant, S., Sakkan, M., Shivalli, P., Kodimela, K., Pappula, S., 2019. Spatialization of rice crop yield using sentinel-1 sar and oryza crop growth simulation model. *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, IEEE, 1–6.

Peng, B., Guan, K., Pan, M., Li, Y., 2018. Benefits of seasonal climate prediction and satellite data for forecasting US maize yield. *Geophysical Research Letters*, 45(18), 9662–9671.

Peralta, N. R., Assefa, Y., Du, J., Barden, C. J., Ciampitti, I. A., 2016. Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. *Remote Sensing*, 8(10), 848.

Schwalbert, R. A., Amado, T. J., Nieto, L., Varela, S., Corassa, G. M., Horbe, T. A., Rice, C. W., Peralta, N. R., Ciampitti, I. A., 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. *Biosystems engineering*, 171, 179–192.

Stepanov, A., Dubrovin, K., Sorokin, A., Aseeva, T., 2020. Predicting soybean yield at the regional scale using remote sensing and climatic data. *Remote Sensing*, 12(12), 1936.

USDA, 2019. USDA foreign agricultural service. https://www.fas.usda.gov/regions/ brazil.

Vermote, E., Wolfe, R., 2015. MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1km and 500m SIN Grid V006 [data set].

Wan, Z., 2015. MOD11A1 MODIS/Terra land surface temperature and the emissivity daily l3 global 1km sin grid. NASA LP DAAC.

Wan, Z., Zhang, Y., Zhang, Q., Li, Z., 2002. Quality assessment and validation of the MODIS global land surface temperature. *International Journal of Remote Sensing*, 23(14), 2661–2674.

Were, K., Bui, D. T., Dick, Ø. B., Singh, B. R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, 52, 394–403.

Wilhelm, W. W., Johnson, J. M., Hatfield, J. L., Voorhees, W. B., Linden, D. R., Berry, L. D., 2004. SoySim: simulation model of soybean growth and management. *Agronomy Journal*, 96(1), 139–149.

Xia, H., Wang, J., He, M., Chen, J., 2020. Soybean yield prediction using remote sensing and machine learning techniques: A case study in the Heilongjiang Province, China. *Journal of Applied Remote Sensing*, 14(2), 026506.