GLOBAL JOURNAL OF AGRICULTURAL AND ALLIED SCIENCES (2024) VOL 5, S 1, PAGES 1-2

EXTENDED ABSTRACT
NAPA 2024 Conference Special Issue

Evaluating Machine Learning Algorithms to Model Time Series of Vegetation Indices in Tallgrass Prairie

Pradeep Wagle [1,*], Gopichandh Danala [2], Catherine Donner [3], Xiangming Xiao[4], Corey Moffet [1], Stacey A. Gunter [1], Wolfgang Jentner [2], and David S. Ebert [2]

[1]*USDA, Agricultural Research Service, Oklahoma and Central Plains Agricultural Research Center, El Reno, OK 73036, USA*
[2]*Data Institute for Societal Challenges, University of Oklahoma, Norman, OK 73019, USA*
[3]*Data Science and Analytics Institute, University of Oklahoma, Norman, OK 73019, USA*
[4]*School of Biological Sciences, Center for Earth Observation and Modeling, University of Oklahoma, Norman, OK 73019, USA*

Abstract history:
Received: July 15, 2024
Revised: November 18, 2024
Accepted: December 15, 2024

Keywords: Climate change, Enhanced vegetation index, Land surface water index, Random forests, XGBoost

Tallgrass prairie is one of the ecologically and economically important grassland ecosystems in the Great Plains of the United States of America (USA). A complex interplay of annual climatic conditions (e.g., temperature, solar radiation, and rainfall), plant species composition, geographic factors, disturbances, and management practices cause yearly variations in the timing of phenological events in tallgrass prairie. Unraveling the connection between climate and satellite-based vegetation indices (VIs) is key for predicting phenological events and productivity of tallgrass prairies under changing climate. Machine learning algorithms have become powerful tools in phenology research to find patterns and relationships between climatic factors and VIs using historical data. We hypothesized that the complex, non-linear response of prairie vegetation to climate requires advanced learning algorithms to capture these intricacies accurately. The objective of this study was to develop robust machine learning model(s) that can predict climate-induced phenological variability in tallgrass prairie by analyzing patterns of VIs and their climatic controls. We compared the performance of six machine learning algorithms - linear regression, eXtreme Gradient Boosting (XGBoost), random forests, decision tree, support vector regression, and K-nearest neighbors (KNN) - in modeling patterns of the enhanced vegetation index (EVI) and land surface water index (LSWI) derived from the Moderate Resolution Imaging Spectroradiometer. EVI, a greenness index, can be used as a proxy of productivity/biomass, while LSWI can be used to track drought conditions and ecosystem health in native tallgrass prairie in Central Oklahoma, U.S. We divided the dataset into three parts: training, testing, and validation. We randomly divided the 2000-2021 data into an 80% training set and a 20% testing set using a time series split. To test the temporal transferability of the models, we further evaluated the performance of the models on a completely new unseen validation dataset (2022-2023) from the same native prairie pasture.

The results showed that climate was a major driver of vegetation phenology in tallgrass prairie. Temperature was particularly important, as it influenced the rate of plant development. Consequently, air and soil temperatures showed the highest correlations with EVI ($r \geq 0.77$) and LSWI ($r \geq 0.56$). Solar radiation also influenced tallgrass prairie phenology. We observed low correlations ($r \leq 0.23$) of EVI and LSWI with contemporaneous rainfall or soil moisture suggesting vegetation's delayed response to these factors (i.e., vegetation responded to changes in rain or soil moisture with a time lag). The effects of other climatic factors such as relative humidity and wind speed were less pronounced. The study suggests that climate change will likely have a significant impact on the vegetation phenology of tallgrass prairie.

Decision tree, KNN, XGBoost, and random forests showed better performance in modeling EVI on the training dataset [coefficient of determination (R^2) = 0.94-1.0, Root Mean Squared Error (RMSE) <0.032, and Mean Absolute Error (MAE) <-0.024] (Fig. 1). Linear regression and SVR models showed relatively weaker performance (R^2 = 0.76-0.77). On the testing dataset, XGBoost, random forests, and KNN showed better performance (R^2 = 0.80-0.83, RMSE = 0.055-0.06, and MAE = 0.042-0.046). Linear regression and SVR showed slightly weaker performance (R^2 = 0.77-0.79), and the decision tree performed the worst (R^2 = 0.65). On the validation dataset, XGBoost and random forests showed the best performance (R^2 = 0.85, RMSE = 0.052-0.053, and MAE = 0.041), while linear regression, SVR, and KNN showed slightly weaker performance (R^2 = 0.71-0.74, RMSE = 0.07-0.072, and MAE = 0.054-0.061). The decision tree again performed the weakest (R^2 = 0.65).

On the training dataset, XGBoost, decision tree, and random forests showed better performance (R^2 = 0.88-1.0, RMSE = 0-0.053, and MAE = 0-0.04) than other models to model LSWI (Table 1). Linear regression, SVR, and KNN models showed weaker performance (R^2 = 0.62-0.69). On the testing dataset, XGBoost and random forests showed the best performance (R^2 = 0.69, RMSE = 0.089-0.09, and MAE = 0.066-0.068). Linear regression, SVR, and KNN showed weaker performance (R^2 = 0.57-0.63), and the decision tree showed the weakest performance (R^2 = 0.44). XGBoost and random forests showed the best performance (R^2 = 0.72, RMSE = 0.086, and MAE = 0.067), followed by linear regression, SVR, and KNN (R^2 = 0.62-0.65, RMSE = 0.096-0.01, and MAE = 0.08-0.085) on the validation dataset. The decision tree was the worst performer (R^2 = 0.38).

* *Corresponding author.* E-mail address: pradeep.wagle@usda.gov

https://doi.org/10.35251/gjaas.2024.004

The results showed that XGBoost and random forests were the best performers across all three datasets (training, testing, and validation) for modeling both EVI and LSWI. The decision tree yielded the weakest results, while linear regression showed a moderate performance. The strong performance of XGBoost and random forests revealed the intricate, non-linear nature of how climatic factors influenced prairie vegetation. These models are well-suited for capturing these complexities as the random forests capture complex patterns by combining predictions from multiple trees and XGBoost optimizes non-linear relationships through gradient boosting.

This study provides insights into the key climatic factors and underlying processes that control the vegetation dynamics of tallgrass prairie ecosystems. The machine learning models developed in this study can serve as valuable tools for predicting the timing of vegetation green-up and senescence, as well as forage production potential, helping ranchers plan grazing and haying activities and developing new strategies to manage tallgrass prairie ecosystems in the face of climate change.

Table 1. Comparison among six machine learning models for simulating land surface water index (LSWI) on the training, testing, and validation datasets. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) were used to compare model performance.

Models	Training Dataset		
	MAE	RMSE	R^2
Linear Regression	0.073	0.096	0.62
XGBoost	0.04	0.054	0.88
Random Forests	0.032	0.046	0.91
Decision Tree	0.0	0.0	1.0
Support Vector	0.069	0.091	0.66
KNN	0.065	0.087	0.69
	Testing Dataset		
Linear Regression	0.08	0.105	0.57
XGBoost	0.066	0.09	0.69
Random Forests	0.068	0.089	0.69
Decision Tree	0.091	0.12	0.44
Support Vector	0.077	0.101	0.60
KNN	0.073	0.097	0.63
	Validation Dataset		
Linear Regression	0.085	0.1	0.62
XGBoost	0.067	0.085	0.72
Random Forests	0.068	0.086	0.72
Decision Tree	0.101	0.128	0.38
Support Vector	0.083	0.096	0.65
KNN	0.08	0.096	0.65

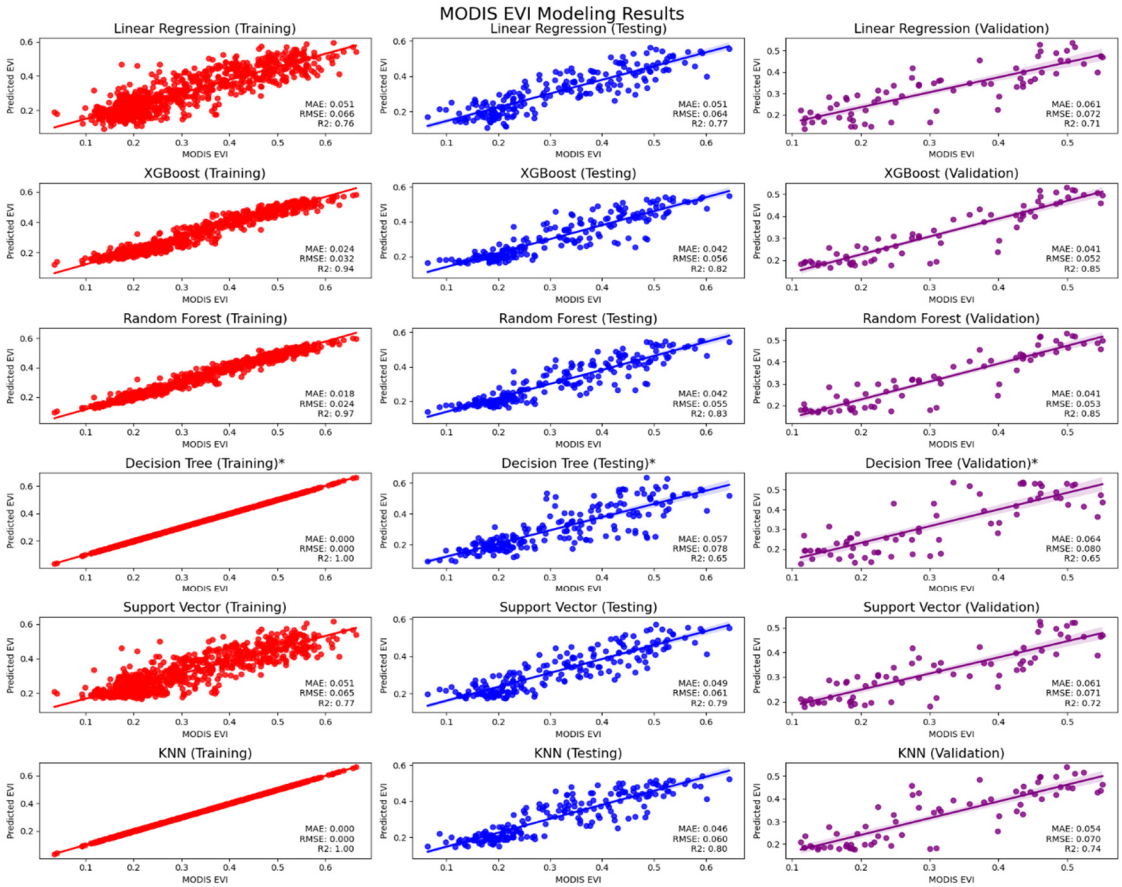

Figure 1. Comparison among six machine learning models for simulating enhanced vegetation index (EVI) on the training, testing, and validation datasets. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) were used to compare model performance.

Citation: Wagle, P., Danala, G., Donner, C., Xiao, X., Moffet, C., Gunter, S.A., Jentner, W., & Ebert D.S. (2024). Evaluating machine learning algorithms to model time series of vegetation indices in tallgrass prairie. *Global Journal of Agricultural and Allied Sciences, 5*(S1), 1-2 https://doi.org/10.35251/gjaas.2024.004.