Leveraging Geospatial Data and Machine Learning to Improve Regional Crop Price Forecasts

Dr. Hari V. Dube[1], Jayprakash H. Pardeshi[2]

[1]Assistant Professor, Department of Computer Science, Padmashri Vikhe Patil College of Arts, Science and Commerce, Pravaranagar, Maharashtra Email ID: dubehvd72@gmail.com
[2]Assistant Professor, Department of Computer Science, Padmashri Vikhe Patil College of Arts, Science and Commerce, Pravaranagar, Maharashtra Email ID: jayprakashpardeshi@gmail.com

Abstract
For farmers, merchants, legislators, and other players in the agri-supply chain, accurate crop price forecasting is essential to lowering market uncertainty and enhancing income stability. However, due to significant spatial variability, seasonal production patterns, climatic shocks, and poor market integration across several geographic zones, regional crop price prediction is still difficult. In order to enhance regional crop price estimates, this study suggests an integrated strategy that makes use of machine learning (ML) and geospatial data. The study makes use of a variety of sources, such as historical market pricing data from agricultural marketing systems, satellite-derived vegetation indicators like the Normalized Difference Vegetation Index (NDVI), and meteorological parameters like temperature and rainfall. Data cleaning, dataset spatial alignment, feature extraction (NDVI trends, rainfall anomalies, and distance-based market connectedness), and predictive modeling utilizing machine learning techniques like Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks are all part of the methodology. Error-based metrics such as MAE, RMSE, and MAPE are used to evaluate model performance, while feature importance and comparative model testing with and without geographical variables are used to analyze the role of geospatial predictors. The anticipated result is a forecasting system that incorporates spatial crop health and weather data to show enhanced robustness in regional price prediction. The study comes to the conclusion that using geographical information in conjunction with machine learning can improve crop price forecasting systems' interpretability and utility for local decision-making, allowing for more proactive market planning and risk reduction.

Keywords: Geospatial analytics, Crop price forecasting, NDVI, Regional markets, remote sensing; Weather features, Machine learning, XGBoost.

1. Introduction

A crucial part of contemporary agricultural decision-making is crop price forecasting. Farmers' decisions on crop selection, planting schedules, storage, and selling tactics are influenced by price signals. Price projections help governments and agricultural organizations make decisions about import-export, buffer stock allocation, procurement planning, and price stabilization. Better forecasts also lessen supply chain interruptions and enhance logistics and inventory control for traders and agriculture companies.

Crop price forecasting at regional levels is still challenging, despite its significance. Regional markets exhibit considerable diversity in contrast to national-level averages because of local production patterns, climate, infrastructural limitations, demand concentration, and market connection. Agricultural markets are not fully integrated in many nations. Due to transportation obstacles, limited cold storage, information asymmetry, and localized demand shocks, prices may not fluctuate consistently throughout districts. Traditional forecasting techniques often model crop prices as time series, focusing mostly on historical price trends. However, agricultural prices are not solely driven by past price movements. They are also influenced by supply-side signals such as crop growth conditions, rainfall variability, pest incidents, and seasonal yield fluctuations. These drivers are inherently spatial and temporal. A region experiencing drought or delayed monsoon may see reduced output and localized price increase, even when other regions remain stable.

Crop health and environmental conditions are continuously and regionally localized thanks to geospatial data sources, especially gridded meteorological databases and remote sensing photography. A popular indicator of crop growth and yield potential, the Normalized Difference Vegetation Index (NDVI) is generated from satellite sensors and represents vegetation vigor. Geospatial indicators can assist in capturing non-linear interactions between crop health, price movement, and climate when combined with machine learning models.

Thus, this study investigates how machine learning in conjunction with geographic data might enhance regional agricultural price projections. The goal is to develop a forecasting framework that includes regional indicators of agricultural conditions in addition to price histories. This method can produce forecasts that are more accurate, particularly in situations where the climate is unpredictable and the market is unstable.

Review of Literature

Traditional Crop Price Forecasting Approaches

Econometric and statistical techniques have been the mainstays of traditional agricultural price forecasts. To model past price trends, methods including autoregressive integrated moving average (ARIMA), exponential smoothing, and moving averages have been often employed. These models work well when price series are stable and show regular seasonal patterns because they make the assumption that future price behavior is a function of previous price history. However, sudden shocks from weather unpredictability and shifts in crop levels frequently cause standard models to fail in agriculture. Although market integration and price transmission across regions have been studied using sophisticated econometric techniques like vector auto regression (VAR) and cointegration analysis, these methods necessitate substantial assumptions about linearity and stationarity. Furthermore, these models frequently rely on the availability of several trustworthy economic indicators, which might not be regularly documented at regional scales.

Machine Learning in Agricultural Price Prediction

Because machine learning can model intricate non-linear interactions, it has gained popularity for price predictions in recent years. When applied to agricultural pricing datasets, algorithms like Random Forest, Support Vector Regression (SVR), Gradient Boosting Machines, and XGBoost frequently demonstrate better performance than traditional techniques. A variety of inputs, including price lags, market arrivals, inflation indicators, and seasonal dummy variables, can be incorporated into machine learning models.

Time series forecasting has also drawn interest in deep learning techniques, particularly recurrent neural networks (RNN) and Long Short-Term Memory (LSTM) networks. Because they can learn long-term dependencies and capture seasonal impacts without explicit manual feature building, LSTM models are especially pertinent. However, in order to prevent over fitting, deep models usually need huge datasets and careful tuning.

Remote Sensing, NDVI, and Weather Influence

Crop classification, yield estimation, drought monitoring, and crop health assessment have all made extensive use of remote sensing in agricultural monitoring. One of the most popular measures of vegetative health, NDVI has a substantial correlation with production and biomass. Numerous studies have shown that combining NDVI with meteorological data enhances drought risk assessment and yield prediction.

Crop health indicators have an impact on prices and predicted production levels from a market standpoint. Reduced supply is probable when the NDVI indicates low vegetation growth, which could lead to price pressure, particularly in local markets. Crop growth, harvesting schedules, and market arrivals are all impacted by weather factors such heat stress and irregular rainfall. Weather and satellite data can therefore serve as early indicators of possible market imbalance.

Research Gaps

Relatively few research concentrate on coupled geospatial–ML techniques explicitly targeted at regional agricultural price forecasting, despite the literature supporting the value of ML for forecasting and remote sensing for agriculture. Significant gaps consist of:

1. Spatial variability is only partially incorporated into price models.
2. Inadequate assessment of weather and NDVI data as indicators of changes in regional prices.
3. Inadequate frameworks for interpreting how geography affects price projections.
4. The absence of end-to-end processes that reliably synchronize market pricing, weather, and satellite information at regional resolution.

By putting forth a structured forecasting methodology that specifically integrates geographical data and assesses their impact via comparison model analysis, this study fills up these gaps.

Objectives of the Study

The objectives of the present study are:

1. To use multi-source datasets (market prices, satellite indices, and weather factors) to create a framework for forecasting regional agricultural prices.
2. To extract and include geographic characteristics into forecasting models, such as rainfall patterns, NDVI trends, and spatial market connections.
3. To evaluate the effectiveness of several machine learning models (Random Forest, XGBoost, and LSTM) for forecasting regional prices.
4. To compare model outcomes with and without geographical features in order to assess the impact of geospatial predictors.
5. To offer useful advice on how to use forecasting results for local agricultural decision-making.

Research Methodology

Data Collection Sources

Three main categories of data are suggested for use in this study:

1. Market Price Information: Government price reporting systems or regulated agricultural markets may provide historical crop price information. These datasets usually include market arrivals, minimum/maximum prices, and daily or weekly modal prices. Data is combined at the district or market level.
2. Satellite-Based Vegetation Data: Sentinel-derived products or satellite sources like MODIS can provide NDVI and associated indices. Crop growth data is regularly provided by the NDVI, which is calculated from red and near-infrared reflectance measurements.
3. Weather Data: Rainfall, minimum and maximum temperatures, humidity, and heat stress indicators can be found

in gridded weather datasets or station-based observations. To match the granularity of market data, these meteorological variables can be combined at the district and regional levels.

Data Preprocessing
Preprocessing entails:
- Using seasonal averages or interpolation to handle missing pricing entries.
- Eliminating anomalies resulting from reporting mistakes or atypical market closures.
- To minimize noise and match satellite revisit times, everyday data is transformed into weekly averages.
- To guaranty uniform mapping between market regions and satellite/weather grids, regional labels should be standardized.
- Time window synchronization ensures that weather indications and NDVI are in line with pertinent pricing periods.

Feature Extraction and Engineering
The basis of geospatial-ML forecasting is feature engineering.
NDVI-based Features:
- The region's mean NDVI at different stages of crop growth
- The growth rate of the NDVI (weekly change)
- Seasonal peak and timing of the NDVI
- NDVI anomaly with relation to the long-term average

Weather-based Features:
- Total rainfall over the past two to four weeks
- Deviation from average rainfall levels
- Indicators of high heat and average temperature
- When applicable, growing degree days (GDD)

Spatial and Market Connectivity Features:
- The distance between major trading markets
- Prices in the spatial neighborhood (weighted averages of neighboring markets)
- Trends in market arrivals (if available)
- Phase indicators for the crop calendar (sowing, vegetative, harvesting)

Time-Series Features:
- Prices that lagged (t-1, t-2, t-4 weeks)
- Rolling standard deviation and rolling mean
- Seasonality can be captured using month/season indicators.

Machine Learning Models Used
The study suggests assessing several models because of variations in learning capacities:
1. Random Forest Regression (RF): Provides feature importance, manages non-linear interactions, and is robust to noise.
2. XGBoost Regression: This gradient boosting model performs well, manages intricate patterns, and facilitates regularization.
3. LSTM Networks: Suitable for sequential learning, particularly when price evolution is influenced by long-term dependencies.
4. Baseline models: simple lag-based regression or ARIMA for comparison.

Evaluation Metrics
Standard forecasting metrics are used to assess the models' performance:
The average absolute variation between actual and expected prices is measured by the Mean Absolute Error (MAE).
- Root Mean Square Error (RMSE): Indicates forecast stability and penalizes excessive errors.
- Mean Absolute Percentage Error (MAPE): This measure, which is helpful for interpretation across various crops and pricing scales, expresses accuracy in percentage terms. Furthermore, a comparative analysis is carried out:
- Geospatial feature-free model performance (price-only baseline) Performance of the model with geographical features (NDVI + weather + spatial) the actual contribution of geographical predictors is made clear by this comparison.

Proposed Framework / Model
The proposed model follows a structured workflow to ensure reproducibility and regional adaptability:
Step 1: Data Collection
- Acquire past crop pricing information from local markets.
- To determine the regional boundary, download and aggregate NDVI imagery.
- Gather meteorological variables that correspond to the same area and period of time.

Step 2: Data Preprocessing
- Standardize time intervals and eliminate inconsistent entries.
- Align datasets according to timestamp and region.

- Use time-series imputation techniques to fill in the gaps.

Step 3: Feature Engineering

- Extract NDVI anomalies and summary by stage.
- Create lag-based rainfall indicators and weather summaries.
- Create spatial features, such as pricing indications in the neighborhood.

Step 4: Model Development

- Use time-aware validation to split data (test on later periods, train on earlier years).
- Utilize designed characteristics to train Random Forest and XGBoost models.
- Use uniform scaling and sequential input windows to train the LSTM model.

Step 5: Forecast Generation

- Depending on the use-case, project crop prices over the following one to four weeks.
- Use model uncertainty estimation or ensemble variability to create confidence bands.

Step 6: Model Evaluation and Interpretation

- Determine each region's MAE, RMSE, and MAPE.
- Evaluate how well different models and feature sets perform. To determine whether geographic drivers enhance forecasts, use feature significance plots (RF/XGBoost).

Step 7: Output and Decision Support

- Provide forecast reports per region.
- Risk alarms with substantial expected price volatility should be flagged.
- Make it possible for dashboards and advisory services to be integrated.

This framework facilitates both practical implementation and research experimentation.

Results and Discussion

It is anticipated that adding geospatial elements will enhance regional crop price estimates in a number of significant ways. Geospatial–ML forecasting can identify early indicators of supply changes brought on by weather and crop health fluctuations, in contrast to strictly price-based forecasting, which mostly relies on the continuity of past pricing trends.

Improvement from Geospatial Predictors

Models are better able to predict anticipated production outcomes when they include NDVI and weather indications. For instance, inadequate biomass development may be indicated by a fall in NDVI during critical crop growth phases, which could eventually result in fewer market arrivals and higher pricing. Similar to this, rainfall shortfalls during flowering or sowing seasons frequently result in decreased yields, which have an effect on regional pricing patterns. Because market prices frequently react after supply changes are apparent in arrivals, price-only models may miss this relationship in its early stages. On the other hand, models can predict price movements before markets fully respond thanks to geospatial indicators, which provide earlier evidence.

Regional Forecasting Performance

The responsiveness of regional markets to supply and demand shocks varies considerably. While some markets are quite isolated, others have strong ties to major trading hubs. By capturing cross-market influence, spatial connectivity indicators like neighborhood pricing averages or distance to large markets can enhance forecasts.

Because of their capacity to learn feature interactions, models like XGBoost are anticipated to perform well. For example, price changes may not necessarily result from NDVI anomalies unless they are accompanied by unfavorable variations in rainfall. Compared to more straightforward linear regression, XGBoost is better able to capture such mixed effects.

When price movements exhibit long seasonal cycles and substantial temporal dependence, LSTM models may be beneficial. However, the usefulness of LSTM in regional forecasting relies on steady reporting patterns and the availability of lengthy historical datasets.

Interpretability and Practical Value

The ability of Random Forest and XGBoost to produce interpretable feature significance metrics is one of its main advantages. Decision-makers gain from understanding whether crop stress signals, rainfall anomalies, or market structure variables have an impact on the model forecast in realistic forecasting systems. This forecasting tool's interpretability promotes confidence.

It is crucial to understand that geographic features do not ensure consistently higher accuracy for all crops and geographical areas. Because to intercropping, mixed cropping, or cloud contamination in satellite data, some crops may have low NDVI sensitivity. Similarly, geographical indicators may have less of an impact in areas where price is heavily influenced by foreign demand or policy initiatives.

Therefore, rather than making exaggerated claims about flawless prediction, improvements should be understood as increased resilience and earl Therefore, improvements should be interpreted as enhanced robustness and early warning potential, rather than unrealistic claims of perfect prediction. The main value lies in reducing forecasting errors in volatile periods and improving the reliability of decision-making at the regional level.y warning potential. Reducing forecasting errors during volatile times and enhancing the dependability of regional decision-making are the primary benefits.

Conclusion

This study offered an integrated research strategy that uses machine learning and geospatial data to improve regional agricultural price forecasts. Farmers, markets, and policymakers depend on crop price forecasting, but regional forecasting is still difficult because of regional variations in agricultural production, local climate unpredictability, and imperfect markets.

The suggested architecture integrates satellite-derived NDVI, historical market price data, and meteorological factors like temperature and rainfall. The forecasting system may capture non-linear correlations between crop growth circumstances and regional price changes by integrating geographical variables into machine learning models like Random Forest, XGBoost, and LSTM. Price-only models and geospatially enriched models can be systematically compared in terms of performance thanks to evaluation using MAE, RMSE, and MAPE.

The study indicates that by offering early indicators of supply-side changes and enhancing model robustness during climatic shocks, geospatial features can improve regional forecasting. Geospatial-ML forecasting has great potential for developing more flexible agricultural market intelligence systems, despite possible drawbacks including satellite noise, resolution limitations, and market policy implications.

Suggestions / Future Scope

This work can be expanded in the following ways by future study and development:

1. Real-Time Forecasting Dashboards: Provide farmers and traders with intuitive dashboards that display price projections, NDVI trends, and weather risk alerts.
2. Extension to Additional Crops and Areas: Expand the framework to encompass a variety of crops, including pulses, soybeans, onions, and maize, as well as a range of agro climatic zones.
3. Higher Resolution Satellite Data: To improve NDVI dependability, use better cloud-masking techniques and higher resolution images (like Sentinel-2).
4. Integration with Market Arrivals and Trade Flows: To enhance price transmission
5. models, use market arrival data, transportation limitations, and regional trade volumes.
6. Policy and Advisory Integration: Include projections in government decisions for farmer advice, price stabilization, and procurement planning.
7. Quantification of Uncertainty: Use probabilistic forecasting models that offer confidence intervals to assist in making risk-conscious decisions.

These enhancements can strengthen the reliability and adoption of crop price forecasting systems, enabling more resilient agricultural markets.

References

1. Aghabozorgi, S., Seyed Shirkhorshidi, A., & Wah, T. Y. (2015). Time-series clustering – A decade review. Information Systems, 53, 16–38.
2. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175.
3. Sakai, M., & Yasuoka, Y. (2008). Monitoring crop phenology using NDVI time series. International Journal of Remote Sensing, 29(2), 1–12.
4. Patel, P. N., Oza, M. P., & Srivastava, R. (2020). Remote sensing-based vegetation indices and crop condition monitoring: Review and applications. Journal of
5. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
8. Doraiswamy, P. C., Moulin, S., Cook, P. W., & Stern, A. J. (2003). Crop yield assessment from remote sensing.Photogrammetric Engineering & Remote Sensing, 69(6), 665–674.
9. FAO. (2017). the future of food and agriculture: Trends and challenges. Food and Agriculture Organization of the United Nations.
10. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
11. Jain, M., Singh, B., & Rao, P. (2019). Application of remote sensing and machine learning for crop yield prediction: A review. International Journal of Remote Sensing, 40(22), 1–24.
12. Justice, C. O., Townshend, J. R. G., Vermote, E., et al. (2002). An overview of MODIS land data processing and product status. Remote Sensing of Environment, 83(1–2), 3–15.
13. Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. Remote Sensing of Environment, 164, 324–333.
14. NDMC (National Drought Management Centre). (2016). Drought indicators and early warning systems. Government publication.
15. Agrometeorology, 22(2), 145–156.
16. Rao, N. H., & Rogers, P. P. (2006). Assessment of agricultural sustainability. Current Science, 91(4), 439–448.