A CRITICAL REVIEW OF MACHINE LEARNING AND TRADE INTELLIGENCE APPROACHES FOR FORECASTING TOBACCO YIELD AND EXPORT PERFORMANCE IN ZIMBABWE

Munashe Masomeke[1] and Rachael Chikoore[2]

[1]Data Science and Analytics, Harare Institute of Technology, Zimbabwe
[2]School of Information Science and Technology, Harare Institute of Technology, Zimbabwe

ABSTRACT

Formal predictive analysis remains limited in tobacco-dependent economies, where forecasting has largely relied on ARIMA-type time-series models. While widely used, these models impose linearity assumptions that restrict their ability to capture key structural drivers of production. This limitation is evident in recent studies where an ARIMA (1,1,0) model projected Zimbabwean tobacco yield at 1,511.78 kg/ha for 2023, underestimating the observed yield of 2,278 kg/ha by approximately 50.7% [1]*. Export forecasting is even less developed, with most existing studies remaining descriptive rather than predictive. The paper reviewed the literature related to tobacco yield forecasting, agricultural export modelling and the application of Machine Learning in crop prediction and trade intelligence systems. Data from across the fields confirm that the Machine Learning techniques Ridge Regression, Random Forest and Gradient Boosting offer superior results to statistical models. The analysis points out three main gaps. First, Machine Learning methods have not been widely applied to tobacco production in sub-Saharan Africa. Second, there is no formal export forecasting model for Zimbabwe that accounts for its multi-year shipment patterns. Third, there is no integrated framework that jointly modelsyield and exports within a unified decision-support system. These gaps highlight the need for more comprehensive, data-driven approaches to forecasting in the tobacco sector.*

KEYWORDS

Tobacco Yield Forecasting, Export Modelling, Machine Learning, Zimbabwe, ARIMA, Trade Intelligence

1. INTRODUCTION

Tobacco ranks as one of the world's most important non-food agricultural commodities in value. Agriculture contributes significantly to foreign exchange earnings, providing rural employment and is an important determinant of the trade balance of many developing countries. Zimbabwe is the biggest producer of tobacco in Africa and one of the six biggest exporters in the world. In 2024, it exported 243 million kilograms of tobacco worth more than US$1.3 billion[2]. The sector is estimated to contribute 10% of national gross domestic product, while also accounting for over half of the value of total agricultural exports[3].

Accurate forecasts are very important in this scenario. Dependable yield forecasts give insight on planning by the Tobacco Industry and Marketing Board (TIMB), government revenue estimates and the purchase strategies of merchants[1]. Still, Zimbabwe and the rest of Sub-Saharan Africa

largely remain wedded to conservative analytical frameworks. The most common institutional approach has been field-based crop assessments. Academic modelling has relied mainly on univariate time-series methods which predict yield as a function of past yield value alone, ignoring a range of structural, behavioural and institutional determinants.

Export forecasting suffers from even more severe constraints. Zimbabwe's tobacco exports volumes are not covered by any formal model. Zimbabwe's shipments of tobacco are not what they seem to be. They are in fact on a known multi-year shipment cycle where a portion of each season's harvest enters the market over multiple years. At present, analyses strictly report realized volumes descriptively without generating any forward-looking projections based on production, stock availability or global market conditions[4].

Agricultural forecasting has seen significant changes within the last 10 years with the advent of Machine Learning methods. Ensemble algorithms like Random Forest and Gradient Boosting outperform classical statistical algorithms when different feature sets are used, according to empirical evidence from systematic reviews[5]. The literature on agricultural economics and trade has at the same time developed richer models integrating production, stocks, prices, exchange rates and competitor supply[6]. Nonetheless, these forecasting improvements have not been systematically applied to tobacco forecasting in developing country context, notably Zimbabwe.

This paper critically reviews the literature in these inter-related domains with a purpose of three. First, it studies the evolution of tobacco yield forecasting from statistical methods to the emergence of Machine Learning methods. Also, it reviews the frameworks for forecasting agricultural exports, multi-year trade pattern analysis and more. Another contribution is mapping the most pressing research gaps and advocating for the development of an integrated forecasting framework, combining Machine Learning and trade intelligence for improved decision-making in Zimbabwe tobacco.

The paper is organized as follows. Section 2 provides sectoral and conceptual background. Section 3 reviews tobacco yield forecasting approaches. Section 4 addresses export forecasting and trade intelligence. Section 5 examines integrated forecasting frameworks. Section 6 presents critical analysis and research gaps. Section 7 concludes the review.

2. SECTORAL AND CONCEPTUAL CONTEXT

2.1. Global Tobacco Production and Trade Landscape

Tobacco is produced largely in few countries in the world. Global raw tobacco leaf production amounted to roughly 5.8 million tonnes in 2022. China accounted for nearly one-third of total global production.The next largest producers were Brazil, Zimbabwe, India, Indonesia and the United States[7]. Zimbabwe holds a unique distinction as Africa's largest producer of tobacco due to favorable natural conditions for the production of flue-cured and the recovery of the sector following the Fast Track Land Reform Programme which transformed production from large-scale commercial farming to smallholder-based systems through contracting farming.

The importance of good forecasting is clear at many levels in the institution. Coordination of marketing logistics and the establishment of procurement expectations based on production estimates is done by TIMB. Projections of government revenue depend directly on export volume forecasts. When forecasts deviate substantially from actual production outcomes, the entire supply chain is affected through resource misallocation, disrupted planning cycles and inefficient trade coordination [4].

2.2. Zimbabwe's Tobacco Industry: Structural Transformation

The structure of the tobacco industry in Zimbabwe has changed significantly since the year 2000. In the pre-land reform era, production was concentrated in the hands of 1,500 large scale commercial farmers who produced a maximum national output of approximately 260 million kilograms in 1998. After the redistribution of the farmland, production declined sharply to about 48 million kg by 2008. The new owners could not access formal land tenure, technical skills and were excluded from mainstream bank financing [3].

Starting in 2010, recovery escalated in the growth of contract farming arrangements in which tobacco buying companies provide smallholder farmers with inputs, extension services and guaranteed market access. In 2023, production at the national level exceeded 296 million kilograms, while the 2025 season had production above 354 million kilograms. The production environment currently observed in Zimbabwe's tobacco sector is structurally heterogeneous, comprising the pre-reform commercial era, the collapse period and the subsequent smallholder recovery phase, with such structural transitions carrying significant implications for forecasting method selection[1].

2.3. Importance of Agricultural Forecasting

Agricultural forecasting uses historical data, information from current conditions and forecasts and analytical models to estimate production, market and trade levels[5]. Yield forecasting refers to the estimates of expected output of production during a specified growing season. These forecasts facilitate crucial operational decisions regarding input allocation, marketing planning and storage management. Export forecasting estimates the quantity and schedule of goods flow into the international market over a horizon.

Export forecasting for tobacco is especially difficult. Tobacco is a product that can be kept for a long time without losing quality unlike perishable crops which are immediately exported after harvest. Accordingly, the allocations of a single season harvest get spread out over a number of subsequent years creating an export cycle that occurs over multiple reporting periods. Hence, contemporaneous production data is an insufficient predictor of annual export volume [8].

Figure 1 shows the classification of the forecasting techniques discussed in this paper. It broadly classifies techniques into two groups, namely classic techniques that include field evaluations, ARIMA models and econometric methods; and contemporary AI-driven techniques, supervised Machine Learning, deep learning, hybrid models and trade intelligence platforms. Sections 3, 4 and 5 review each component of this taxonomy.

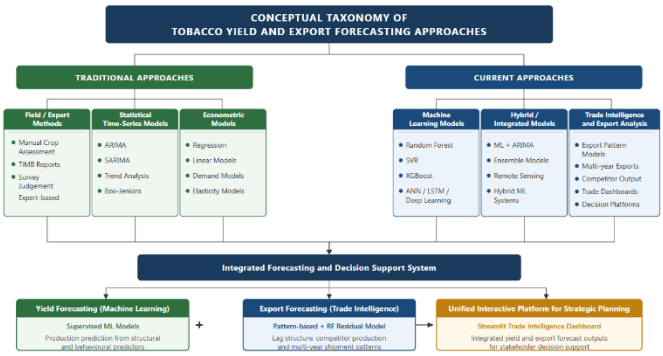

Figure 1. Conceptual taxonomy of tobacco yield and export forecasting approaches

3. TOBACCO YIELD FORECASTING APPROACHES

3.1. Traditional Forecasting Methods and Their Limitations

In Zimbabwe, expert assessments by agricultural extension officers are used to evaluate tobacco yield. Crop-cut surveys and ratio/ extrapolation methods evaluate yield based on seed purchase data and registered land area. Even though they are practically accessible, they are structurally deficient, since they are unable to generate reliable and consistent forecasts.

In Zimbabwe, Svotwa, Hakata and Murandu[9] directly compared traditional and remote-sensing-based approaches. They established that values derived from traditional methods are approximately 112 percent accurate. This denotes a systematic overestimation of 12 percent. Further, MODIS-derived NDVI estimates are 98.8 percent accurate. The disparity in the results illustrates the main weakness of survey-based methods: their reliance on subjective judgement and on incomplete official records means that given the difference, it is not possible to carry out a correction without the benefit of in-depth analysis.

Mapuwei et al. [1] fitted ARIMA (1,1,0) model to the national tobacco yield data for the period 1980 to 2018 and forecast that by 2023 the production will decline to 1511.78 kg/ha. The actual production was more than the projected amount, which was 2,278 kg/ha in 2023. The under forecasting was systematic with the forecast gap being sizeable. This shows an essential structural shortcoming: the specified ARIMA cannot explain nor do a proper model of the nonlinear expansion of contract farming, the rise in registered growers, or the interaction effects between policy incentives and input access. As noted by Mapuwei, such limitations call for the inclusion of structural variables and alternative methods in future studies.

Analysis of comparable situations in Africa confirms this pattern. According to Suleman and Sarpong[10], the ARIMA (2,1,0) model was applied to Ghanaian milled rice production, whose accuracy reduces due to the presence of structural breaks. Shoko and Belete[11] found similar trend on sorghum production in South Africa. The common thread in all this literature is the appropriateness of the ARIMA model for short run forecasts under stability of regimes but a severe deterioration when the data generating process has undergone structural change of the sort that Zimbabwe's tobacco sector has undergone since 2000.

Ordinary Least Squares regression enlarges the analytical portfolio by allowing for explanatory variables. Nonetheless, it assumes a constant linear connection between predictors and output, a condition often breached in dynamic agricultural systems. Due to the significant multicollinearity in area planted, seed sales and grower registration in Zimbabwe tobacco, OLS coefficient estimates can be unreliable. Hence, alternative regularized regressions are considered.

3.2. Machine Learning Approaches to Crop Yield Prediction

Data availability and computational capacity have improved which made shift to data-driven yield forecasting possible. Machine Learning techniques enable the estimationcomplex, non-linear relationships among many predictor variables at the same time. Liakos et al. [12] conducted an analysis of the numerous applications of Machine Learning in crop management, livestock, water and soil and concluded that using Machine Learning to integrate farm and sensor data is transforming agricultural systems into real-time analytical intelligence for decision support.

According to the systematic review by Van Klompenburg, Kassahun and Catal[5], which examined a total of 567 studies on crop yield prediction, the most applied family of algorithms consisted of Artificial Neural Networks, Deep Neural Networks and Recurrent Neural Networks.

This was followed by tree-based ensemble methods, such as Random Forest, Gradient Boosting, etc. In this review which evaluated many applications to compare various statistical techniques, ensembles (particularly Random Forests) achieved median R^2 values consistently greater than 0.85 in well-specified applications. In comparison, other classical statistical techniques had much lower median R^2 values. This review was able to reach a highly relevant methodological conclusion: that random train-test splits for time-series datasets severely inflate model accuracy and that temporally ordered validation strategies are required for credible evaluation.

The Random Forest algorithm, proposed by Breiman[13], teaches an ensemble of decision trees on bootstrap samples using random subsets of features. This architecture minimizes overfitting and controls multicollinearity by design, leading to trustworthy variable importance estimates, making it ideal for agricultural datasets with correlated predictors and non-linear relationships. Ridge Regression solves the problem of multicollinearity using L2 regularization. This method shrinks the estimates of correlated coefficients towards zero. In the case of multiple joint inclusion of predictors, standard OLS would be unstable. The tobacco production dataset is interesting from this perspective as the input variables, area planted, seed sales, grower registration, all appear to move together.

Scalable variants of Gradient Boosting have been one of the most successful algorithms on agricultural forecasts. Shahhosseini et al.[14] showed that a combination of Gradient Boosting with crop simulation model output can lower corn yield RMSE by 7% to 20% with respect to weather alone, demonstrating the value of mechanistic-data-driven model coupling. Paudel et al.[15] show that Machine Learning models that trained on crop simulation, weather, remote sensing and soil features depending on multiple European crop systems outperformed the traditional approach; moreover, Random Forest and Gradient Boosting achieved R^2 values in the range of 0.85 to 0.96 consistently.

According to a systematic review published in 2024 by Jabed and Azmi Murad, deep learning architecture such as convolutional neural networks and Long Short-Term Memory networks has a strong potential to capture spatial and temporal dependence in agriculture. Performance relies greatly on the quality and availability of data. The datasets on the agricultural sector of sub-Saharan Africa like the tobacco production in Zimbabwe often suffer from incompleteness and poor temporal richness that limits the application of models developed in data-rich settings[16].

3.3. Remote Sensing Integration in Tobacco Yield Estimation

Data from remote sensing, especially vegetation indices derived from satellite imagery, provide an objective and spatially continuous alternative to administrative estimation. According to Svotwa et al. [9], the MODIS-derived estimates have an accuracy of 98.8 percent for tobacco yield in Zimbabwe. In contrast, estimates generated through traditional field methods suffer from a 112 percent systematic overestimate. This demonstrates how remote sensing could radically enhance national production planning.

Yet, Zimbabwe's smallholder context limits the scalability of remote sensing. The spatial resolution of MODIS is very coarse, relative to the fragmented field patterns common in the resettlement areas, resulting in tobacco-specific spectral signals not being separated from neighbor crops. Using structural administrative predictors and vegetation indices could circumvent the problems of data quality and spatial resolution, but such a combined approach has

not been systematically evaluated for Zimbabwe. Methods of Machine Learning combined cluster sets such as Random Forest and Ridge Regression[13], [17]. Such feature sets are well-placed to take advantage of.

3.4. Comparative Evaluation and Benchmarking

The performance evaluation of agricultural forecasting models is based on a consistent set of regression metrics including the mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and coefficient of determination R^2. In the studies examined, ensemble methods consistently hit R^2 exceeding 0.85 in well-specified applications of crop yield prediction; ARIMA and OLS, conversely, exhibit much lower and more variable performance[5]. There are some methodological caveats that restrict comparisons across different studies. Train-test splits must be temporally ordered to optimize unbiased estimates. Dataset size dampens performance variation. Hyperparameter tuning practices were not uniformly reported in the literature surveyed [12].

Table 1 presents a structured overview of the forecasting methods covered throughout the review. It compares model, average empirical performance from literature, strengths, weaknesses, datasets and key references. Table 2 provides quantitative performance metrics for the ten most directly relevant prior studies, offering a quantitative benchmarking basis for comparison.

Table 1. Synthesis of Forecasting Approaches: Model Type, Strengths, Limitations and Key References

Approach	Model Type	Typical R^2	Strengths	Limitations	Dataset Context	Key Reference(s)
Expert / Survey	Descriptive	Variable	Accessible; low cost	Systematic bias; misses structural dynamics	Administrative, field data	Zimbabwe tobacco [9]
ARIMA / Box-Jenkins	Univariate time-series	Low when structural breaks present	Interpretable; easy to implement	Linearity assumption; fails under regime changes; no covariates	National yield series	Zimbabwe tobacco[1]
Ridge Regression	Regularised linear ML	High (~0.94 in comparable studies)	Manages multicollinearity; stable coefficients	Approximate linearity; limited non-linear capture	Crop inputs, grower data	Prior literature [17]
Random Forest	Tree ensemble ML	High (>0.85 median across studies)	Non-linear; robust; feature importance	Black-box; degrades with small samples	Multi-variable crop datasets	Multi-crop yield [5], [13]
Gradient Boosting /XGBoost	Boosting ensemble ML	High (>0.90)	Top accuracy; captures complex interactions	Computationally intensive; hyperparameter-sensitive	Satellite + administrative data	Crop yield [14]
Deep Learning / LSTM	Neural network ML	High (>0.90) when data-rich	Captures temporal-spatial dependencies	Requires very large datasets; limited interpretability	Large-scale time-series	Multi-crop review [16]

Remote Sensing + ML	Hybrid spatial-ML	High when combined with ML	Objective; spatially continuous; reduces survey bias	Resolution limits in fragmented smallholder landscapes	MODIS / Landsat data	Zimbabwe tobacco [9]
VAR Models	Multivariate time-series	Moderate to high	Captures feedback effects; internally consistent	Requires long stationary series; data-intensive	Trade, production, price panels	Agricultural trade [6]
Gravity Model	Cross-sectional econometric	N/A for time forecasting	Explains bilateral trade flows	Cross-sectional; unsuitable for single-country forecasting	Bilateral trade statistics	Agricultural trade [18]
Integrated DSS Platforms	Decision support systems	N/A (framework-level)	Combines models, data and interfaces for scenario planning	Requires significant institutional capacity	Multi-source: farm, weather, trade	Smart farming [19], [20]

Table 2. Performance Metrics from Prior Forecasting Studies

Study	Model	R^2	RMSE	MAE	MAPE	Context and Observations
Mapuwei et al. [1] (2022)	ARIMA(1,1,0)	N/R	High	N/R	~50.7%	Zimbabwe tobacco yield (1980-2018); forecast 1,511.78 kg/ha vs observed 2,278 kg/ha in 2023
Van Klompenburg et al. [5] (2020)	Random Forest (review median)	>0.85	Variable	Variable	Variable	Systematic review of 567 crop yield studies; ensemble methods superior to classical baselines
Shahhosseini et al. [14] (2021)	XGBoost + Crop Simulation	>0.90	RMSE -7 to -20%	Low	Low	Corn yield, US Corn Belt; coupling ML with simulation reduced RMSE by 7-20%
Paudel et al. [15] (2021)	RF and GBM	0.85-0.96	Low	Low	Variable	Multi-crop yield forecasting, Europe; outperformed classical statistical baselines
Jabed and Azmi Murad [16] (2024)	Deep Learning / LSTM	>0.90	Variable	Variable	Variable	Multi-crop systematic review; notes data availability constraints in sub-Saharan Africa
Svotwa et al. [9] (2013)	Remote Sensing (NDVI)	~0.99 accuracy	Low	Low	1.2% error	Zimbabwe tobacco; MODIS 98.8% accuracy vs survey

						112% overestimation
Suleman and Sarpong [10] (2012)	ARIMA(2,1,0)	Moderate under stability	Moderate	Moderate	Moderate	Ghana milled rice; adequate under stable regime; accuracy declined with structural changes
Shoko and Belete [11] (2017)	ARIMA-family	Moderate under stability	Moderate	Moderate	Moderate	Sorghum production, South Africa; confirms ARIMA limitations under regime change

The evidence assembled in Tables 1 and 2 confirms that Machine Learning methods, particularly regularized linear models such as Ridge Regression and tree-based ensembles such as Random Forest and Gradient Boosting, consistently outperform classical ARIMA and OLS frameworks when the underlying data-generating process involves structural changes and non-linear variable interactions. The comparative disparity is most pronounced in contexts like Zimbabwe where multiple production regimes coexist within a single historical series.

4. EXPORT FORECASTING AND TRADE INTELLIGENCE

4.1. Conceptual Foundations of Agricultural Export Modelling

Forecasting of agricultural exports is entirely different from yield forecasting in terms of both scope and complexity. Export volumes are not dependent upon domestic production alone. Export volumes emerge from the interaction of the levels of production, stocks carryover, conditions of global demand, price competitiveness and supply of competitors. The theoretical foundation for export supply modelling is the commodity balance identity: $Exports_{\mathrm{t}} = Production_{\mathrm{t}} + Opening\ Stocks_{\mathrm{t}} - Domestic\ Use_{\mathrm{t}} - Closing\ Stocks_{t}$. This framework, documented by FAO [21], makes explicit how export availability depends not only on current production but also on the disposition of prior-season inventories.

According to Anderson and Nelgen[6], agricultural trade flows are driven by stock-to-use ratios and price expectations, rather than contemporaneous supply. When stock levels are high relative to expected demand, it's likely there will be an increase in exports to prevent the buildup of stocks. Possible export response patterns introduce lag structures that contemporaneous-production models cannot capture. The basic dynamic nature makes forecasting of exports intrinsically more difficult than that of yields and justifies explicit lag-structure modelling.

4.2. Econometric and Statistical Export Forecasting Models

Multivariate econometric frameworks have largely governed export supply modelling. Goldstein and Khan's research[22] implies that import demand is a foreign income and price competitiveness determinant that drives export volumes. Although default parameter stability is assumed in the single-equation specifications, they are unlikely to hold in the increasing volatile agricultural export systems characterized by structural break or policy shocks.

The OECD-FAO [23] are integrated agricultural outlook systems that produce forecasts in model systems that capture feedback effects between production, prices and volumes of exports. The forecasts are more internally consistent than single-equation approaches. This evidence confirms

the analytical advantage that systems joint modeling of production trade has over separate modeling.

The economic mass and distance between two countries determine the level of trade between them. While these gravity modelsare widely used in aggregate trade analysis in agriculture, their cross-sectional orientation does not lend itself readily to predicting time series export volumes for a single country/commodity [18]. Consequently, they have little to offer in the way of direct applicability to the Zimbabwe tobacco export forecasting problem. However, they are useful in understanding export destination patterns.

4.3. Multi-Year Shipment Patterns in Tobacco Trade

Tobacco can be stored for long periods without any significant loss of quality unlike most perishable agricultural commoditieslike food and vegetables. As a result, exports from the harvest of a given season continue to enter global markets across one or two or three subsequent years, as appropriate to contracts, grading schedules or stock management decisions. According to FAO [21], at global level and across the world, stock carryovers are the norm on account of processing needs and delivery contracts. Anderson and Nelgen[6] provide the theoretical foundation that the agricultural trade flows adjust imperfectly and adjust gradually rather than instantaneously. Further, it happens due to price expectations and policy uncertainty.

The design of the export model is affected by this multi-period framework. The quantity of exports each year is not solely reliant on current production levels; decisions made in previous seasons regarding storage, contracts and market conditions collectively determine how much tobacco will ultimately reach export markets. Despite the global literature on the tobacco trade recognizing this characteristic, this has not been simulated in any Zimbabwe-specific published study. This is one of the major methodological gaps identified in the review.

4.4. Trade Intelligence and Decision Support Systems

Trade intelligence refers to the systematic integration of trade statistics, market analysis, predictive models and visualization tools within user-centric platforms that aim to assist regulators, exporters and industry stakeholders in strategic decision-making [24]. For instance, the International Trade Centre’s Trade Map provides an illustration of the institutional model embedded in interactive environments that facilitate proactive rather than retrospective market positioning.

Agricultural decision-support systems build upon this concept by integrating forecasting models, data analysis and interactive user interfaces. Wolfert et al.[19] highlighted the rise of smart farming platforms that integrate data from farms and environmental and market information to facilitate value chain decisions. The DSSAT cropping system model [20] presents an influential earlier instance of how simulation models can be embedded in decision support architectures integrating data management, model management and human– computer interaction. Nevertheless, the existing systems provide more focus on crop management as opposed to trade forecast and the export model is not integrated into decision support.

The three fundamental subsystems proposed by Sprague and Carlson [25] are data management, model management and user interaction. This agricultural trade intelligence platform brings together yield forecasting, export modelling and scenario design capabilities all in an integrated environment which is an architecture first. Agricultural decision support systems are evolving from static archival reporting to more dynamic scenario assessment planning through the incorporation of Machine Learning.

5. INTEGRATED FORECASTING FRAMEWORKS

5.1. Rationale for Integration

The commodity balance identity discussed in Section 4.1 makes an argument for integration directly: a yield estimate, once produced, should inform the export projection rather than be looked at in isolation. According to FAO 2017, most major agricultural outlook systems follow this sequential structure in which production estimates feed into balance sheet models that determine stock changes, domestic use and export availability. OECD-FAO systems enable commodity-market scale operations that link country-level projections to global trade and price equilibria [23].

Anderson and Nelgen[6] point out that supply conditions, stock movements and world market conditions jointly determine the outcome of agricultural trade. Essentially, if the production estimates lack integrated export estimates, then they are not very useful to plan with. The World Bank [26] included also competitiveness dimension whereby export performance is determined by the cost & quality position of the country. Any comprehensive forecasting model for Zimbabwe that is affected by output variations from its major competitor countries must have cross-country elements.

5.2. Conceptual Architecture for an Integrated Framework

An integrated forecasting framework for Zimbabwe's tobacco sector would have three linked elements based on the literature covered in sections 3 and 4. A conceptual representation is shown in Figure 1. The first is a yield forecasting module which takes in structural and behavioural production determinants at the national level and produces a production forecast that looks ahead in time.According to the literature, supervised Machine Learning methods, specifically, regularized regression and tree-based ensembles, are technically superior to ARIMA and OLS for this purpose [17].

The second component is a forecast of exports component which takes the yield estimate from the first component and distributes this over either current or future calendar years. This distribution is based on the multi-year lag structure documented by FAO in 2017. It is consistent with the trade flow adjustment evidence of Anderson and Nelgen. The distribution should seek to incorporate competitor supply signals across major producing countries as proposed by Goldstein and Khan [22] and the trade competitiveness analysis of Mutodi et al. [4].

A decision support system interface, which links both forecasting modules to end users, is the third component. The interface should enable institutional users, including TIMB planners, government revenue forecasters and export merchants, to enter production assumptions and receive forward-looking yield and export projections. It will be designed based on the DSS architecture of Sprague and Carlson [25] and the smart farming platform design of Wolfert et al. [19]. According to Jones et al [20], model outputs can be embedded in interactive decision support systems aimed at end-users.

The architecture proposed here directly addresses the three research gaps identified in this review. These gaps are a lack of supervised Machine Learning applied to Zimbabwe tobacco yield data, a lag structure export model which is formalized and a platform linking both in a decision support environment.

5.3. Decision Support Architectures in Agricultural Systems

According to Sprague and Carlson [25], there is a three-subsystem model for DSS. Jones et al. demonstrated its application in crop simulation through the DSSAT system, which links mechanistic crop models with soil, weather and management databases. A more recent example Wolfert et al. [19] suggests that smart farming platforms extend this model by combining various data streams and analytical layers into a unified, real-time decision environment.

According to literature, there is a consistent gap concerning the development of agricultural decision support systems (DSS) on farm-level crop management and commodity trade forecasting. The World Bank [26] stress that production forecasting systems and trade planning tools should be more closely linked, especially at national level in developing countries with weak institutional data capacity. This gap can be directly addressed by developing an integrated framework linking yield prediction with export modelling within a user-friendly interface.

5.4. Comparator Country Practices

Countries such as Brazil, India and China, major producers of tobacco, use more structured and data-driven methods to plan exports than are documented for Zimbabwe. These countries produce detailed statistics on production and trade, which are published on a regular basis and feed into sectoral planning cycles [2]. According to the OECD-FAO [23] agricultural outlook framework, national production estimates may be linked to global trade predictions through integrated analytical frameworks. Having more comprehensive forecasting capacity provides countries with planning advantages. This demonstrates the analytical gap's practical significance in Zimbabwe and strengthens the argument for the establishment of an integrated forecasting framework.

6. CRITICAL ANALYSIS AND RESEARCH GAPS

6.1. Methodological Critique of Yield Forecasting Literature

Analysis of tobacco yield forecasting literature shows it contains three major methodological flaws. First, ARIMA-based approaches dominate Zimbabwe-specific studies, despite relying on assumptions that are inconsistent with the structural characteristics of national production data. This is based on the structural features of national production data. Between 2000 and 2008, land reform disruption happened and the recovery led by smallholders afterwards resulted in regime changes that make historical stationarity assumptions untenable. Mapuwei et al.'s[1] ARIMA (1,1,0) model anticipated production of 1,511.78 kg/ha in the year 2023. However, the actual production observed in 2023 was 2,278 kg/ha. This discrepancy amounts to a deviation of approximately 50.7 percent. Such a significant failure indicates structural inadequacy. Namely, post-2020 production was not able to be anticipated by the model.

Additionally, there is an ongoing disconnect between the Machine Learning techniques employed in similar international contexts and the analytical toolbox that has been deployed in Zimbabwe-specific work. Van Klompenburg et al. [5] states that Random Forest and Gradient Boosting significantly outperform statistical baselines for 567 crop yields. Paudel et al. [15] confirmed this for large European crop systems and Shahhosseini et al. [14] demonstrated the benefits of coupling Machine Learning with simulation outputs. The mentioned approaches were not applied to national-scale data on tobacco production in Zimbabwe.

There is also a trade-off between predictive accuracy and interpretability in the literature. Tree-based ensemble methods have better metrics than linear models on regression, but the black-box

character makes them unsuitable for causal inference, required for the design of production interventions by policymakers[12]. While feature importance scores can provide some insights, there is no published systematic treatment of interpretability and predictive performance for Zimbabwe tobacco in the literature.

6.2. Limitations of Export Forecasting Literature

Zimbabwe's literature on tobacco export forecasting seems to lack formal model-based approaches. According to the available studies, export volumes realized are reported descriptively with no forward-looking projections developed [4]. There has been no formal lag-structure modelling of Zimbabwe's multi-year tobacco shipment cycle, despite the phenomenon being acknowledged in FAO [21].The omission of Zimbabwe-specific export analyses on production variables of other competitors is further a material omission given that competition from global supplies determines export market share [6], [22].

6.3. Gaps in Integrated Forecasting Literature

Academic studies on tobacco consistently call for yield forecasting, plus export modelling, to be housed in a single methodology. Yet this construction has not been found. Integrated agricultural outlook systems relate production forecasts to stocks, trade and prices in an internally consistent framework [21]. The literature on agricultural decisions support system mentions up-to-date platforms that combine data pipelines, model layers and interactive interfaces for scenario-driven analytics. On the literature review, there is no published study that jointly deals with Zimbabwe's national tobacco yield and export to within a single interactive decision-support environment.

Without such a framework, establishing yield-related forecasts is difficult and even if reliable yield forecasts are available, converting them into actionable export forecasts requires significant manual analytical work. The foundational DSS architecture established by Sprague and Carlson [25] and the smart farming platform evidence of Wolfert et al.[19] both point toward integrated, automated pipelines as the appropriate solution; however, this architecture has not been implemented for the Zimbabwe tobacco context.

6.4. Research Gap Summary

This review identifies three substantive and inter-connected gaps in research. The first problem is that yield forecasting for Zimbabwe is methodologically inadequate because the exclusive use of ARIMA has led to demonstrably inaccurate forecasts and the absence of any application of supervised Machine Learning alternatives with demonstrated superiority in comparable contexts to national-scale data. The second gap is a complete absence of formal export forecasting models that explicitly incorporates behaviour of multi-year shipments, stock carryover dynamics and global competitor supply. The third gap is the lack of a platform-based framework that simultaneously incorporates production and export models and enables the presentation of both outputs through a trade intelligence interface.

7. CONCLUSION

A review of the literature on tobacco yield forecasting, agricultural export modelling, Machine Learning applications and trade intelligence systems, focused on Zimbabwe's tobacco sector, identified the research gaps that motivate an integrated forecasting framework.

The review leads to three central conclusions. The methodological limitations of tobacco yield forecasting literature in Zimbabwe lie in use of classical time-series approaches whose structural assumptions are contradicted by the data. In systematic reviews, Ridge Regression, Random Forest and Gradient Boosting have been found to outperform other methods in similar agricultural settings. As evidenced by quantitative material in the literature, ensemble methods yield R^2 values above 0.85 consistent, while ARIMA systematically underestimated Zimbabwe's actual 2023 tobacco yield by 50.7 percent.

Additionally, Zimbabwe's tobacco sector export forecasting remains at a pre-analytical stage. None of the published models contain any formal incorporation of the multi-year shipment cycle, stock carryover dynamics and competitive supply signals from major producer countries. At TIMB, the institutional data infrastructure required for building such a model exists. The theoretical foundations for such a model can be found in the agricultural trade literature [6], [22]. There is a methodological gap: the application of structured lag models and trade intelligence variables to Zimbabwe-specific shipping data.

Another issue is the lack of a comprehensive integrated platform-based framework for linking yield prediction to export modelling within a single decision support system. The literature on agricultural decision support systemsand integrated agricultural outlooksdenotes that this integration is essential for evidence-based planning of the sector.

This review's contribution to future research is threefold. The assessment provides the first systematic comparison of yield and export forecasting methods applied to the Zimbabwe tobacco context based on quantitative evidence from previous literature assembled in Table 1 and Table 2. It identifies the Zimbabwe tobacco export system's multi-period lag structure as a necessary structural input to formal export modelling based on theoretical grounds established by FAO [21]. This study sets out initial thoughts on how to establish a yield-forecasting integrated platform that connects DSS and trade intelligence frameworks. It envisions a forecasting platform that will make DSS framework proposals operational.

Future studies should test supervised Machine Learning algorithms for the prediction of Zimbabwean tobacco yield utilizing multi-predictor feature sets that can capture the structural, behavioural and temporal production dynamics identified in this review. Consistent with the trade flow adjustment evidence of Anderson and Nelgen[6], export modelling can integrate lag allocation patterns with competitor production inputs. An interactive platform allowing future scenario analysis access to non-expert institutional users should incorporate both components. The methodological template constructed for Zimbabwe can be generalized to other export-oriented agricultural commodities in data-sparse developing economies where classical forecasting techniques have similarly failed to keep pace with structural transformation.

ACKNOWLEDGEMENT

The authors acknowledge the support of the Harare Institute of Technology, School of Information Science and Technology and express gratitude to the Tobacco Industry and Marketing Board of Zimbabwe for providing access to production and export data used in the broader research programme from which this review originates.

REFERENCES

[1] T. W. Mapuwei, J. Ndava, M. Kachaka, and B. Kusotera, "An Application of Time Series ARIMA Forecasting Model for Predicting Tobacco Production in Zimbabwe," Am. J. Model. Optim., vol. 9, no. 1, pp. 15–22, Dec. 2022, doi: 10.12691/ajmo-9-1-3.

[2] Trading Economics, "Trading Economics." Accessed: Oct. 09, 2025. [Online]. Available: https://tradingeconomics.com/zimbabwe/exports/tobacco-manufactures-tobacco-substitutes
[3] I. Scoones, N. Marongwe, B. Mavedzenge, F. Murimbarimba, J. Mahenehene, and C. Sukume, "Zimbabwe's land reform: challenging the myths," J. Peasant Stud., vol. 38, no. 5, pp. 967–993, Dec. 2011, doi: 10.1080/03066150.2011.622042.
[4] K. Mutodi, E. T. Maziriri, and T. Chuchu, "The response of Zimbabwe tobacco exports to real exchange rates volatility: 1980–2019," J. Agribus. Rural Dev., vol. 56, no. 2, pp. 201–219, Jun. 2020, doi: 10.17306/J.JARD.2020.01241.
[5] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Comput. Electron. Agric., vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.
[6] K. Anderson and S. Nelgen, "Trade Barrier Volatility and Agricultural Price Stabilization," World Dev., vol. 40, no. 1, pp. 36–48, Jan. 2012, doi: 10.1016/j.worlddev.2011.05.018.
[7] FAOSTAT, "FAOSTAT." Accessed: Oct. 09, 2025. [Online]. Available: https://www.fao.org/faostat/en/#home
[8] FOA, "Knowledge Repository ::Home." Accessed: Feb. 17, 2026. [Online]. Available: https://openknowledge.fao.org/home
[9] E. Svotwa, A. J. Masuka, B. Maasdorp, A. Murwira, and M. Shamudzarira, "Remote Sensing Applications in Tobacco Yield Estimation and the Recommended Research in Zimbabwe," ISRN Agron., vol. 2013, pp. 1–7, Dec. 2013, doi: 10.1155/2013/941873.
[10] N. Suleman and S. Sarpong, "Forecasting Milled Rice Production in Ghana Using Box-Jenkins Approach," Int. J. Agric. Manag. Dev., vol. 2, no. 2, pp. 79–84, Jun. 2012.
[11] R. R. Shoko and A. Belete, "Efficient planning of sorghum production in South Africa – Application of the Box-Jenkin's method," J. Agribus. Rural Dev., vol. 46, no. 4, pp. 835–841, Dec. 2017, doi: 10.17306/J.JARD.2017.00352.
[12] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," Sensors, vol. 18, no. 8, Aug. 2018, doi: 10.3390/s18082674.
[13] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[14] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," Sci. Rep., vol. 11, no. 1, p. 1606, Jan. 2021, doi: 10.1038/s41598-020-80820-1.
[15] D. Paudel et al., "Machine learning for large-scale crop yield forecasting," Agric. Syst., vol. 187, p. 103016, Feb. 2021, doi: 10.1016/j.agsy.2020.103016.
[16] Md. A. Jabed and M. A. Azmi Murad, "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," Heliyon, vol. 10, no. 24, p. e40836, Dec. 2024, doi: 10.1016/j.heliyon.2024.e40836.
[17] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," 1970.
[18] K. Head and T. Mayer, "Gravity Equations: Workhorse,Toolkit, and Cookbook," in Handbook of International Economics, vol. 4, Elsevier, 2014, pp. 131–195. doi: 10.1016/B978-0-444-54314-1.00003-3.
[19] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big Data in Smart Farming – A review," Agric. Syst., vol. 153, pp. 69–80, May 2017, doi: 10.1016/j.agsy.2017.01.023.
[20] J. W. Jones et al., "The DSSAT cropping system model," Eur. J. Agron., vol. 18, no. 3, pp. 235–265, Jan. 2003, doi: 10.1016/S1161-0301(02)00107-7.
[21] FOA, "Issues in the Global Tobacco Economy: Selected Case Studies." Accessed: Feb. 17, 2026. [Online]. Available: https://www.fao.org/4/y4997e/y4997e00.htm
[22] M. Goldstein and M. Khan, "Income and price effects in foreign trade," in Handbook of International Economics, vol. 2, in Handbook of International Economics, vol. 2. , Elsevier, 1985, pp. 1041–1105. Accessed: Feb. 19, 2026. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:intchp:2-20
[23] OECD and Food and Agriculture Organization of the United Nations, OECD-FAO Agricultural Outlook 2021-2030. in OECD-FAO Agricultural Outlook. OECD Publishing, 2021. doi: 10.1787/19428846-en.
[24] "ITC - Transforming trade. Changing lives." Accessed: May 25, 2026. [Online]. Available: https://www.intracen.org/

[25] "Sprague, R.H. and Carlson, E.D. (1982) Building Effective Decision Support Systems. Prentice-Hall International Inc., London, 329pp. - References - Scientific Research Publishing." Accessed: May 23, 2026. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=1198383

[26] W. World Bank, "The World Bank Group Annual Report 2025," World Bank. Accessed: Feb. 17, 2026. [Online]. Available: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099503310092520301