FUNCTIONAL TIME SERIES MODELS FOR K-STEP PREDICTION OF SUGAR PRODUCTION: A CASE STUDY OF ESWATINI, KENYA, AND SOUTH AFRICA

Abstract

Sugar, produced from cane or beet, is a vital energy source and a globally traded commodity. Most business happens in a futures exchange market where speculators, hedgers, and institutional consumers and producers, make decisions based on their understanding of the future supply and demand situation. Sugar is sometimes bought and sold before the cane or beet is planted. Hence, improving sugar production forecast accuracy is vital to maximize gains and enable effective planning. The study considered the annual sugar production total as a function of the monthly output rather than a scalar quantity and analyzed historical data using three functional time series techniques. In particular, the methods used for k-steps and dynamic updating prediction of sugar production include Local Polynomial Regression (LPR), Ridge Regression (RR), and Penalized Least Squares (PLS). The performance of the three models was compared to that of the automatic ARIMA method using the Mean Squared Error (MSE) and the Mean Absolute Percentage Error (MAPE) measures to establish the suitability of each technique in sugar production forecasting. The LPR, RR, and PLS outperformed the ARIMA method in all three cases. However, the prediction accuracy was lower for highly volatile datasets from countries that overly depend on rainfall for cane development.

Keywords: Local Polynomial Regression, Ridge Regression, Penalized Least Squares, Functional Time Series

UNDER PEER REVIEW

Introduction

Most of the global sugar trade occurs in a futures market setting where parties buy or sell the commodity through financial contracts that obligate them to execute the agreement at a future date. In recent years, the sugar market has been one of the most volatile alongside forex and equities. Participants in the sugar futures market make key trading decisions based on their knowledge of future production and demand situations. Hedgers, for example, are in the business to reduce the risk of financial loss and maximize the value of their assets from price movements. While speculators intend to make profits from price changes (U.S. Commodity Futures Trading Commission, 2024). Significantly large variations in sugar production forecasts could have financial ramifications for various market parties. Hence, there is a need to consistently improve methods of predicting output.

Irrespective of the prediction model chosen, the number of historical values used is a key parameter in determining the accuracy of the final forecast. Insufficient past data makes the model less flexible, but a large volume of historical observations can also be troublesome in nonparametric statistics due to the asymptotic patterns of exponentially decaying estimates. According to Shang and Hyndman (2011), functional ideas, such as splitting the time series into several segments, effectively resolve the problem that arises from incorporating a large volume of past data in a forecasting model. Many time series also evolve in a constantly changing environment where data patterns vary over time. Hence, it is reasonable to design methods that consider possible alterations that affect the study phenomenon.

Different methodologies have been applied to forecast sugar and sugarcane output and yields during the last decade. One prominent method is the random forest algorithm, which has been implemented to predict yields for various crops, including mangoes and switch grass (Fukuda et al., 2013; Tulbure et al., 2012). Despite its extensive application, the technique has produced mixed results regarding prediction accuracy. Garc´ia-Gutierrez et al. (2015) noted that while random forest models have demonstrated significant predictive power in numerous cases, there have been several instances where conventional time series forecasting algorithms outperformed them.

Mehmood et al. (2019) applied the ARIMA time series technique to forecast sugarcane production in Pakistan. Using the Box-Jenkin methodology, the study fitted an ARIMA (2, 1, 1) model on 71 yearly cane output data to develop a k-step ahead forecast for the next 12 years' annual sugarcane crop outturn. This approach highlighted the ARIMA model's robustness in providing reliable forecasts over an extended period.

Paswan et al. (2022) combined the ARIMA technique with ANN to create a model for predicting sugarcane output in the Bihar region of Pakistan. The study utilized 76 observations for training and the remaining five data points for testing the models' accuracy. It found that the ANN method outperformed ARIMA in forecast accuracy using common evaluation metrics, including the root mean squared error (RMSE) and the mean absolute percentage error (MAPE), particularly producing a smaller MAPE and RMSE.

Beyond traditional time series models, geospatial and remote sensing techniques have also been explored for yield forecasting. Bezuidenhout and Singels (2007) used satellite mapping to predict sugarcane yields in South Africa. The study combined remote sensing techniques with conventional statistical models, requiring frequent satellite flyovers to capture the desired images and identify crop changes, an exercise noted to be potentially expensive in the long run. Luciano et al. (2021) also used remote sensing techniques to predict cane yield in Brazil. They incorporated several agronomic and

UNDER PEER REVIEW

meteorological variables using the random forest model to enhance predictions from Landsat images, taking four years to calibrate the model for actual yield estimation.

Other innovative approaches have contributed to the diversity of methodologies in sugarcane yield forecasting. Gupta and Agarwal (2021) modeled sugar output using multivariate linear regression on a time series of annual production totals collected between 1931 and 2018. The analysis considered various factors, including milling capacity, cane tons per hectare yields, cane crushed volumes, and area under cane. The researchers split the historical data into two segments, using 80 percent as the training sample and the remaining 20 percent for model validation. Similarly, Megha et al. (2019) set out to identify the most suitable approach for predicting sugar production in India using a 36-year dataset collected between 1990 and 2015. They fitted several linear and non-linear methods, concluding that the cubic model outperformed all others in prediction accuracy.

Several studies within the East African setup have also aimed to develop frameworks for forecasting sugar or sugarcane crop outturn. Kwamboka et al. (2019) modeled sugarcane yields in Kenya using the seasonal ARIMA (SARIMA) methodology, selecting SARIMA (0,1,1) (0,0,0) 12. The study recommended adopting alternative techniques to provide comparative forecasts. Mwanga et al. (2017) also predicted Kenya's sugarcane crop yield using the seasonal ARIMA technique, choosing SARIMA (2,1,2) (2,0,3) 4 for quarterly yield forecasting.

In Uganda, Yuma et al. (2023) evaluated three machine learning techniques, including the random forest algorithm, decision tree methodology, and multiple linear regression, to establish the most suitable method for cane yield estimation. They found that the random forest methodology, which provided up to 94.6 percent accuracy, was the most suitable model for forecasting Ugandan cane yields. The study recommended future investigations employ more data and explore web-based machine-learning techniques and deep-learning methods to improve forecast accuracy.

Most studies have relied on ARIMA models, which perform well for the short-term point forecast but rapidly accumulate errors, leading to less reliable long-term predictions. Again, the majority of the investigations have concentrated on sugarcane yield forecasting, not directly predicting sugar output. Studies have also shown that functional time series models outperform conventional time series approaches in producing more accurate forecasts in other sectors (Shang and Hyndman, 2011).

Other researchers modeled sugar and sugarcane production using remote sensing and other geospatial information systems (GIS) crop mapping techniques. The GIS or remote sensing technique could be expensive, especially when it involves flying drone cameras over the fields every few days. Also, sugarcane takes 12-19 months to mature in Africa, making it difficult to distinguish the current

UNDER PEER REVIEW

year from next year's crop by only examining satellite images. Also, the model calibration to enable Landsat imagery in crop prediction takes a long time before a suitable solution is obtained.

Additionally, most recent investigations have been conducted outside the African continent. In some countries, such as Brazil, sugarcane production is almost fully mechanized, unlike in many African producer nations. While some forecasting models developed outside Africa could be applied to project sugar output in the continent, others may not be directly useful.

The choice of the models proposed in this study is based on the comparative analysis from previous studies. Shang and Hyndman (2011) showed that the PLS and RR produced more accurate functional time series forecasts than other methods such as the block moving, conventional time series forecast, and the ordinary least squares. Also, Kihara (2013) proposed local polynomial regression as a suitable solution in estimating trends when the underlying distribution cannot be explicitly specified. In our case, the underlying distribution of the sugar production trends in the three countries is unknown. Hence the need to train, test, and compare the performance of several models to identify the most appropriate strategy for adoption.

Methodology

The study analyzed data collected between 1990 and 2023 from the international sugar organization, Eswatini Sugar Association, and Agriculture and Food Authority-Kenya**.** Production conditions differ in various locations across the region. In East Africa, sugar milling happens all year in most countries, while in southern Africa, most sugarcane harvesting and processing happens between April and December. Also, in some producer states, the crop is fully irrigated, and cane development is more mechanized. Therefore, this study used data from Kenya, South Africa, and Eswatini, representing the varied nature of the sugar production landscape within Sub-Saharan Africa. South Africa and Eswatini are surplus producers, while Kenya is a net deficit country. The weather patterns are also different making the three suitable representatives of the industry representation.

Local Polynomial Regression (LPR)

Local Polynomial Regression (LPR) is essential for analyzing random variables with unknown distributions. For random pairs $(Y_0, Y_1), \ldots, (Y_{n-1}, Y_n)$ the dependent variable Y_i follows the model:

$$Y_i = m(Y_{i-1}) + v(Y_{i-1})^{\frac{1}{2}} \epsilon_i \quad (1)$$

where ϵ_i is an independent and identically distributed standard normal random variable; $v(\cdot)$ is the variance function; and $m(\cdot)$ is the function to be estimated.

UNDER PEER REVIEW

Let Y_i be the annual sugar production for the i^{th} year, represented as;

$$Y_i = \sum_{j=1}^{p} x_{ij} \quad (2)$$

where x_{ij} denotes the monthly sugar production for month j in year i with p being the number of production months annually. Assuming that the $Y_i's$ have a common but unknown probability distribution function f for all $i's$. The researcher uses weighted least squares to fit a polynomial of degree q and approximate the kernel estimator $\hat{m}(Y; q, h)$ for random variable pairs (Y_i, Y_{i-1}).

The bandwidth h satisfies the limiting conditions:

$$\lim_{n\to\infty} h = 0 \text{ and } \lim_{n\to\infty} nh = \infty$$

The coefficients $\hat{\beta} = \left(\widehat{\beta_0}, \ldots, \widehat{\beta_q}\right)^T$ are obtained by minimizing

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{q}\beta_j (Y_{i-1} - c)^j\right)^2 K_h(Y_{i-1} - c) \quad (3)$$

Subject to the β_j coefficients and $K_h(Y_i - c)$ is a Gaussian kernel function of the form;

$$K_h(Y_i - c) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(Y_i - c)^2}{2h^2}\right) \quad (4)$$

where *c* is the point at which the kernel is centered and *h* is the bandwidth.

The estimators of the model coefficients *βj*, for example, the estimator of *β1* regression coefficient at a particular point *c*, is then obtained as follows:

$$\hat{m}(c; q, h) = e_1^T (X_c^T W_c X_c)^{-1} X_c^T W_c Y = e_1^T \hat{\beta}_j(c) \quad (5)$$

where Y is the response set, W_c is an $n \times n$ diagonal matrix of weights, X_c is $n \times (q+1)$ design matrix with 1s in the first column; and $e_1 = [1 \quad 0 \ldots \quad 0]^T$ (Kihara, 2013).

After estimating the $\hat{\beta}_j$ coefficients, the predicted annual sugar production for year n is obtained as:

$$Y_{n\,forecast} = \hat{\beta}_0 + \widehat{\beta_1} Y_{n-1} + \hat{\beta}_2 Y_{n-1}^2 + \hat{\beta}_3 Y_{n-1}^3 \quad (6)$$

Ridge Regression

The Ridge Regression (RR) framework modifies the standard linear regression model by introducing a constant λ to the diagonal elements of the matrix $(X^T X)$. The linear regression model $\hat{\beta} = (X^T X)^{-1} X^T Y$ is improved to obtain the RR equation given as:

UNDER PEER REVIEW

$$\hat{\beta}_{ridge} = \left(X^T X + \lambda I_p\right)^{-1} X^T Y \quad \textbf{(7)}$$

The $\hat{\beta}_{ridge}$ parameters are then obtained by minimizing

$$S_y = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \textbf{(8)}$$

The penalty term λ, restricts the $\hat{\beta}_{ridge}$ parameters, resulting in a reduction of the sum of squared residuals. If the covariates are independent, then

$X^T X = nI_p$, leading to $\hat{\beta}_{ridge} = \frac{n}{n+\lambda}\ \hat{\beta}_{OLS}$,

where $\hat{\beta}_{ridge}$ is the Ridge regularization estimator, n is the number of observations, and $\hat{\beta}_{OLS}$ is the ordinary Least squares estimator. The degrees of freedom associated with the regularization parameter λ can be expressed as $df(\lambda) = tr\left(X\left(X^T X + \lambda I_p\right)^{-1} X^T\right) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$

Where $df(\lambda)$ is the degrees of freedom as a function of λ, X is the design matrix of input features , I_p is the identity matrix of size $p \times p$ and d_j are the singular values of the matrix X. Additionally, the variance-covariance matrix is given by:

$$var\left(\hat{\beta}_{ridge}\right) = \sigma^2\left(X^T X + \lambda I_p\right)^{-1} X^T X \left(X^T X + \lambda I_p\right)^{-1} \quad \textbf{(9)}$$

When $\lambda = 0$, there is no penalization, and all the p parameters are retained . For $p = 2$, RR corresponds to a circle, $\sum_{j=1}^{p} \beta_j^2 < c$. Future production values are predicted as;

$$\hat{Y} = X\hat{\beta} \quad \textbf{(10)}$$

where Y and X are the future response and design matrix, respectively.

Penalized Least Squares

Penalized Least Squares (PLS) is an improvement of the Ordinary Least Squares (OLS) approach designed to reduce over-fitting and enhance curve smoothness. Studies by Shang and Hyndman (2011) and Kan et al. (2019) showed that PLS provides more accurate forecasts than OLS when dealing with longitudinal data. The random variables are modeled as:

We assume the random variable Y_i is associated with Y_{i-1} and that there is an unknown nonlinear interrelationship between Y_i and Y_{i-1}. The random variables are modeled as:

$$Y_i = f(Y_{i-1}) + \epsilon_i \quad \textbf{(11)}$$

where ϵ_i is independent and identically normally distributed with zero mean and constant variance. Y_i is the annual sugar production for the ith year. In OLS, β and $f(Y_{i-1})$ are estimated by minimizing:

$$S(f) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(Y_{i-1})\right)^2 \quad \textbf{(12)}$$

UNDER PEER REVIEW

The PLS approach introduces a penalty function to obtain a function f that best interpolates $f(Y_{i-1})$ with an acceptable degree of smoothness. In PLS, the function to be minimized is:

$$S_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(Y_{i-1})\right)^2 + \lambda J_2(f) \quad \textbf{(13)}$$

where $J_2(f)$ is the penalty designed to achieve smoothness in f. The penalty is approximately equal to the square of the second derivative of the function f. $S_\lambda(f)$ is broken down into two parts: the first represents the goodness of fit , while the second, which contains the λ parameter, measures the smoothness. An increase λ emphasizes smoothness, while a decrease improves the goodness of fit. The function $f_\lambda(Y_i)$ is approximated as:

$$f_\lambda(Y_i) = \theta_0 + \sum_{j=1}^{n}\theta_j Y_{i-1} + \sum_{j=1}^{n}\delta_j B_j(Y) \quad \textbf{(14)}$$

where B_j is the basis function and $Y = [Y_1, \dots, Y_n]^T$. Fourier series basis functions effectively represent periodic functions and are suitable for signal processing (Leith, 2024). Also, (Saputro et al., 2019) noted that the Fourier series trigonometric functions cosine and sine are flexible and can easily adjust to the local patterns in data. Hence, the cosine and sine trigonometric polynomials are suitable for modeling data when the underlying distribution is unknown. Therefore, $B_j\ (Y_{i-1}) = \sin\left(\frac{j\pi Y_{i-1}}{p}\right)$ or $B_j\ (Y_{i-1}) = \cos\left(\frac{j\pi Y_{i-1}}{p}\right)$, where p is the period of the function, in this case, the number of production months in a year. Parameter λ is obtained by minimizing:

$$V(\lambda) = \frac{\left(\frac{1}{n}\right)\|I - H(\lambda))\,Y\|^2}{\left[\left(\frac{1}{n}\right)tr(I - H(\lambda))\right]^2} \quad \textbf{(15)}$$

which is the Generalized Cross Validation (GCV) function, and $H(\lambda)$ is the hat matrix.

$$H = X(X^TX)^{-1}X^T \quad \textbf{(16)}$$

Matrix H projects the observed values Y onto the space spanned by the columns of X, giving the fitted values $\hat{Y}$. The predicted annual sugar output is obtained by plugging in the calculated values of the coefficients into the equation($\mathbf{14}$) , which serves as the forecast of Y_i .

The performances of the LPR, RR, and PLS techniques proposed for predicting future years' sugar production are then compared to that of the ARIMA model, the most common prediction method used in the sugar industry. The automatic ARIMA model handles both seasonal and non-seasonal data and is generally expressed as ARIMA (p,d,q)(P,D,Q)s where p is the order of the non-seasonal autoregressive 24 (AR) part, d is the degree of non-seasonal differencing, and q is the order of the non-seasonal moving average (MA) part. Also, P is the order of the seasonal

UNDER PEER REVIEW

AR, D is the degree of seasonal differencing, Q is the order of the seasonal MA, and s is the seasonal period. Ideally, an ARIMA (p,d,q)(P,D,Q)s model would be of the form;

$$\Phi(B).\Phi_s(B^s).\Delta^D.\Delta_s^D.y_t = \Theta(B).\Theta_s(B^s).\epsilon_t \quad (57)$$

where:

$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ is the non-seasonal AR part,

$\Phi_s(B^s) = 1 - \Phi_1 B^s - \Phi_2 (B^s)^2 - \cdots - \Phi_P (B^s)^P$ is the seasonal AR part.

$\Delta^D = (1 - B)^D$ is the seasonal differencing operator, $\Delta_s^D = (1 - B^s)^D$ is the seasonal differencing operator, $\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ is the non-seasonal MA polynomial,

$\Theta_s(B^s) = 1 + \Theta_1 (B^s) + \Theta_2 (B^s)^2 + \cdots + \Theta_Q (B^s)^Q$ is the seasonal MA polynomial,

y_t is the value of the time series at time t, ϵ_t is the white noise at time t, B is the backshift operator for non-seasonal differencing and B^s is the backshift operator for seasonal differencing.

Results

Table 1 gives a summary of the descriptive statistics of the data used in the study. Among the three countries, South Africa is the largest sugar producer, followed by Eswatini.

Table 1: Annual Sugar Output 1990 to 2023

Country	Mean	Standard Deviation	Minimum	Maximum	Median
Eswatini	586,971	77,887	413,945	746,981	594,921
Kenya	499,024	98,254	328,844	796,554	489,603
South Africa	2,119,434	363,183	1,259,208	2,844,165	2,121,366

Also, from Table 1, Kenya's sugar production trend has the highest volatility with a 20% coefficient of variation (standard deviation as a percentage of the mean). The Eswatini dataset has the lowest volatility (13%) compared to South Africa (17%) and Kenya.

Model Performance for K-steps Sugar Production Forecasting

To evaluate the performance of the LPR, RR, and PLS models, annual production data (variable Y) was normalized using the standard deviation approach as:

UNDER PEER REVIEW

$$Y_{normalized} = \frac{Y - \mu_Y}{\sigma_Y} \tag{68}$$

The most appropriate parameters in each case were determined though cross validation. The results are shown in Table 2.

Table 2: Model Performance for K-Step Sugar Production Forecasting

Local Polynomial Regression						
Country	Bandwidth (h)	LPR order	MAPE Training	MAPE Test	RMSE Training	RMSE Test
Eswatini	0.25	0	3.6696	7.0201	26914	63681
		1	3.1761	9.1728	23379	70766
		2	2.4905	5.8759	18645	48936
		3	1.6486	5.3566	12264	47784
		4	0.9771	13.0971	10538	102028
		5	1.0178	105.1289	10667	787355
Kenya	0.25	0	6.0882	20.9415	34393	141532
		1	4.9932	29.3922	29346	156815
		2	4.3095	38.4024	26819	208070
		3	3.2915	47.5184	20499	252748
South Africa	0.25	0	10.3118	6.6104	256725	142348
		1	9.4728	14.9377	227576	322420
		2	7.2571	18.1532	189139	381917
		3	6.5453	11.6525	162504	245142
		4	5.0809	32.3652	136196	780787
Penalized Least Squares						
	λ	α	MAPE Training	MAPE Test	RMSE Training	RMSE Test
Eswatini	0.28	0	3.6194	5.9520	24522	45400
		1	6.5324	7.1995	45196	56206
		0.1	4.0392	5.8812	26762	45841
Kenya	0.08	0	5.5549	31.8729	31699	174329
		1	6.4341	28.6934	36150	160465
		0.1	5.5939	31.5715	31858	172891
South Africa	0.1	0	9.3030	9.1949	240667	221719
		1	11.3004	9.2498	273080	194749
		0.1	9.4033	8.9816	241565	215493
		0.2	9.5490	8.7172	243055	210121
		0.3	9.6961	8.5516	245163	205399
Ridge Regression						
	λ		MAPE Training	MAPE Test	RMSE Training	RMSE Test
Eswatini	0.08		6.2148	8.1095	44137	66070
Kenya	0.10		5.0714	18.0759	39755	119759
South Africa	0.04		14.9503	10.0011	350660	220955
Automatic ARIMA						
	MAPE (Training)		MAPE (Test)	RMSE (Training)		RMSE (Test)
Eswatini	5.1227		9.8293	40134		75878
Kenya	9.6899		29.3891	54893		158417
South Africa	14.4385		12.1043	346042		245506

UNDER PEER REVIEW

In all three cases, 25 percent of the range was selected as the most appropriate bandwidth through cross-validation. LPR models of degrees 0 to 5 were fitted to the data successively, and the following results were obtained. From Table 2, the quadratic local polynomial (order 2) outperformed both the Nadaraya-Watson kernel regression (order 0) and the linear local polynomial regression (order 1) for Eswatini. The cubic polynomial (order 3) was the most accurate, achieving a training MAPE of 1.6486% and a test MAPE of 5.3566%, with RMSE values of 12,264.4 and 47,783.6. Higher-order polynomials (order 4 and above) overfitted the training data, leading to poor generalization.

For South Africa, the Nadaraya-Watson kernel regression (order 0) was the most suitable, with a training MAPE of 10.3118% and a test MAPE of 6.6104%, showing consistent performance and lower error margins compared to higher-order models. Similarly, for Kenya, the Nadaraya-Watson kernel regression (order 0) provided the best results, with a training MAPE of 6.0882% and a test MAPE of 20.9415%, indicating reasonable forecasting accuracy.

The most suitable values of the penalty λ were determined as 0.08 for Eswatini, 0.10 for Kenya, and 0.04 for South Africa. From the result in Table 2, the RR model achieved a MAPE of 5.0714% on the training data and 18.0759% on the test data for Eswatini, with RMSE values of 39,755 and 119,759, respectively. For South Africa, the RR model resulted in a MAPE of 14.9503% on the training data and 10.0011% on the test data, with RMSE values of 350,660 and 220,955, respectively. Overall, the RR model fits the Eswatini sugar data better than the data for South Africa or Kenya. It provides highly accurate results for Eswatini and South Africa and a good forecast for Kenya's annual sugar output.

The PLS model with $\alpha = 0.1$ and $\lambda = 0.28$ achieved high accuracy for Eswatini, with a training MAPE of 4.0392% and a test MAPE of 5.8812%. RMSE values were 26,762 for training and 45,841 for testing, indicating an effective balance between fit and smoothness. For South Africa, the optimal PLS model used ridge regularization ($\alpha = 0$) with $\lambda = 0.1$, resulting in a training MAPE of 9.3030% and a test MAPE of 9.1949%. The RMSE values were 240,667 for training and 221,719 for testing. The lasso regularization ($\alpha = 1$) model under-fitted the training data and may need feature engineering for better performance. For the Kenyan dataset, the best PLS model had $\alpha = 1$ and $\lambda = 0.08$, yielding a training MAPE of 6.4341% but a much higher test MAPE of 28.6934%. While it provided reasonable forecasts, it was less effective compared to the models for Eswatini and South Africa, reflecting difficulties in capturing Kenyan sugar production trends.

ARIMA is the most widely used technique for forecasting sugarcane yields and production based on our investigations. It auto-ARIMA model provided accurate predictions for Eswatini's annual sugar

UNDER PEER REVIEW

production, with MAPE values of 5.1227% and 9.8293% and RMSE values of 40,134 and 75,878 for the training and test samples, respectively. However, the ARIMA method could not match the performance of the LPR, RR, and PLS techniques. Among the models, PLS ($\lambda = 0.28$, $\alpha = 0.1$) proved to be the most suitable for Eswatini, with LPR and RR also outperforming auto-ARIMA. LPR (order 3, bandwidth = 0.25) achieved MAPE values of 1.6486% and 5.3566%, while RR ($\lambda = 0.08$) resulted in MAPE values of 6.2148% and 8.1095%. For Kenya, auto-ARIMA fitted the training data well (9.6899% MAPE) but performed poorly on the test data (29.3891% MAPE). Although none of the models were highly accurate for Kenya, RR ($\lambda = 0.1$) was the most effective, with MAPE values of 5.0714% and 18.0759% for the training and test samples. In modeling the South African sugar data, PLS ($\lambda = 0.1$, $\alpha = 0$) was the best for forecasting yearly sugar output, outperforming other techniques, including auto-ARIMA, which under-fitted the data. RR and LPR models also showed lower RMSE compared to auto-ARIMA. Overall, LPR, RR, and PLS were superior to auto-ARIMA for k-step sugar production forecasting.

UNDER PEER REVIEW

Five-Year Annual Sugar Production Forecast

We forecasted annual sugar production quantities for South Africa, Eswatini, and Kenya for the five years from 2024 to 2028.

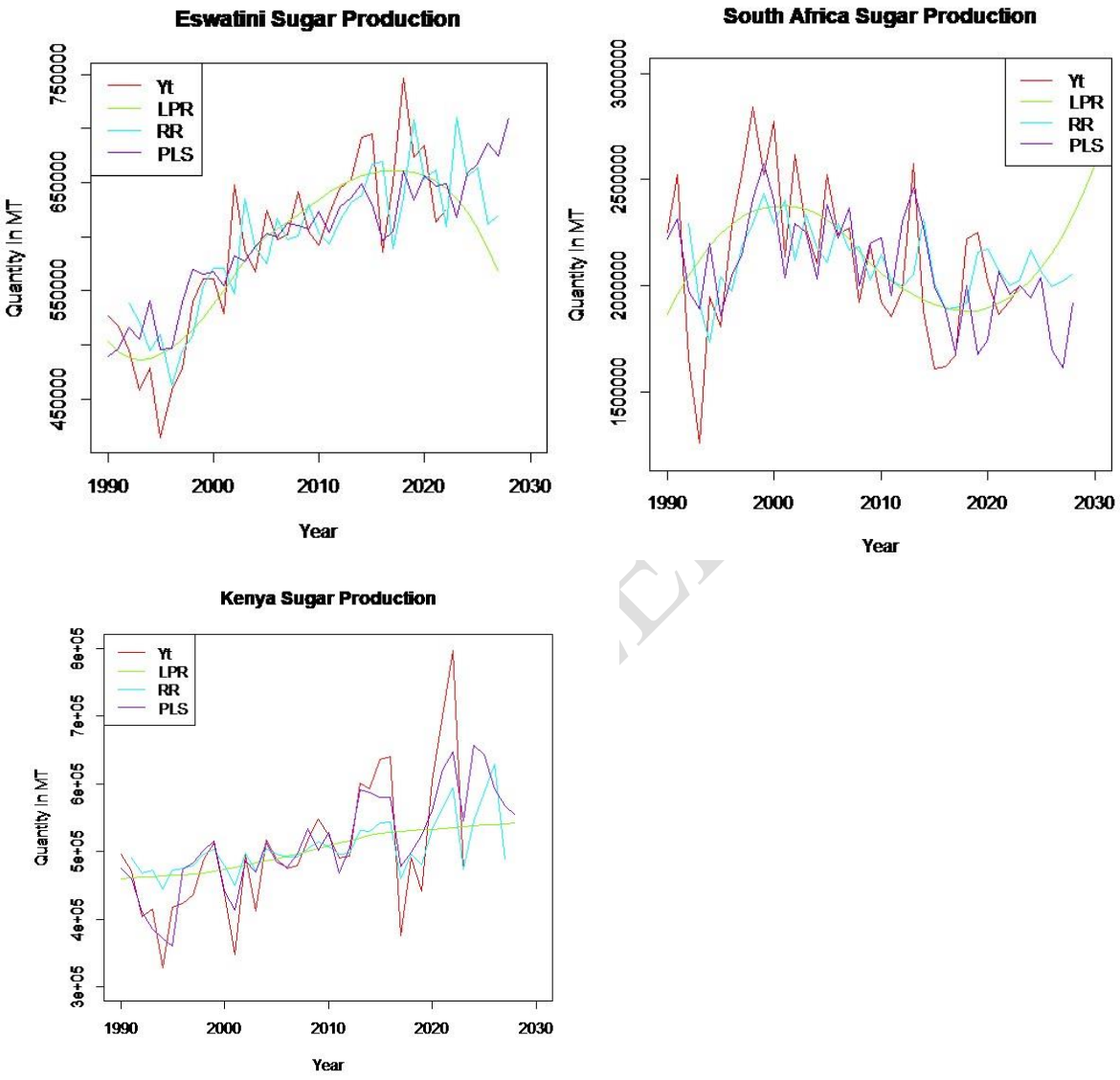

Figure 1: Actual (1990-2023) and Predicted (2024-2028) Sugar Output.

The predicted pattern obtained using LPR, RR, and PLS are shown in figure 1 with Y_t being the actual output. While the predicted production quantities are given Table 3 where the bold values are forecast obtained from the most suitable model.

Table 3: Five Year Sugar Production Forecast Using the Proposed Models (in metrics tons)

Model	2024	2025	2026	2027	2028
Eswatini					
LPR	635132	623012	607813	589380	567560
RR	710567	655835	664158	611614	620140
PLS	**658831**	**666793**	**686043**	**674421**	**710540**
Kenya					
LPR	537676	538831	539985	541139	542293
RR	**473323**	**544630**	**586864**	**629036**	**487363**
PLS	657348	642293	592436	567096	552973
South Africa					
LPR	2153064	2235038	2331690	2443849	2572345
RR	2169082	2066650	1995415	2023522	2056982
PLS	**1941827**	**2034203**	**1687262**	**1601052**	**1906542**

The PLS model ($\lambda = 0.28$, $\alpha = 0.1$) is the best forecasting approach for Eswatini's annual sugar output, providing the lowest error margins. Based on the selected PLS technique, Eswatini's sugar output is expected to rise gradually over the next three years (2024-2026). In 2027, output is predicted to be slightly lower than in 2026 but still above the long-term average production level.

On the other hand, Kenya's sugar production trend has experienced significant shocks in recent years. Among all three countries, Kenya's sugar output trend exhibits the highest volatility (20%), making it difficult to consistently and accurately predict the yearly sugar output. Even the selected ridge regression model ($\lambda = 0.1$), which produced the lowest error margins among all three techniques for forecasting the country's yearly output, still had up 18% error margin. Based on the chosen RR technique, annual production is expected to increase gradually between 2024 and 2027. However, it could fall by about 23 percent in 2028.

For South Africa, a PLS model with $\lambda = 0.1$ and assuming ridge regularization, is the best overall technique for predicting South Africa's yearly sugar production. The most suitable LPR and RR models were unreliable, as they significantly underfit the training dataset even after being optimized through cross-validation. Therefore, it may be necessary to do feature engineering on the South African data before implementing LPR or RR. Using the selected model, South Africa's sugar production is predicted to be between 1.6 million and 2 million tons annually over the next five years. The historical data shows a standard deviation of 361,183, about 17% of the long-term average, indicating higher volatility than Eswatini (13%), but lower than Kenya.

UNDER PEER REVIEW

Conclusions & Recommendations

The ARIMA model is the most common methodology for sugar, sugarcane quantity, and sugarcane yield forecasting. However, all the techniques suggested in this study outperformed the ARIMA in predicting the total annual sugar production in the three selected countries (South Africa, Kenya, & Eswatini). The LPR, RR, and PLS all produced significantly lower error margins and more consistent results when compared in terms of the MAPE and RMSE measures on the training and test samples.

In particular, the PLS ($\lambda = 0.28;\ \alpha = 0.1$) and PLS ($\lambda = 0.1;\ \alpha = 0$) were the best overall models for forecasting future years' sugar production volumes for Eswatini and South Africa, respectively. While RR ($\lambda = 0.1$) was the chosen technique for predicting production outcomes for the Kenyan case. Still, all the proposed methodologies (LPR, RR, & PLS) and the auto-ARIMA methodology produced higher error margins when implemented on the Kenyan data than the Eswatini and South African sugar datasets. Kenya's sugar production data had the highest volatility (20%), adversely affecting the prediction accuracy.

The Kenya sugar production trend showed more pronounced shocks (extreme values) than Eswatini and South Africa. In Kenya, sugarcane is grown under rain-fed agriculture, while the crop in Eswatini is almost fully irrigated. Whereas for South Africa, about 25% of sugar is produced from irrigated farms while the rest is rainfall-dependent. Also, the milling capacity is consistently increasing in Kenya but almost constant in Southern Africa. Therefore, we can conclude that overreliance on rainfall and milling capacity changes leads to increased volatility of sugar production, making the outcomes more difficult to predict.

We recommend the PLS, LPR, and RR models for adoption in predicting annual sugar output in Eswatini, South Africa, Kenya, and the wider East and Southern Africa for countries with similar experiences. Where cane production is overly dependent on rainfall, it could be important to investigate whether an additional categorical weather variable can improve the accuracy levels. Also, for highly volatile datasets, we recommend that the methods can be used together with dynamic updating models discussed in our other study to improve forecast accuracy.

UNDER PEER REVIEW

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

References

U.S. Commodity Futures Trading Commission (2024). Commodity Futures Trading Commission. Retrieved March 20, 2024

Shang, H. L. and Hyndman, R. (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation,* 81:1310–1324.

Garc´ıa-Gutierrez, J., Mart ınez-Alvarez, F., Troncoso, A., and Riquelme, J. C. (2015). A comparison of machine learning regression techniques for lidar-derived estimation of forest variables. *Neurocomputing*, 167:24–31.

Mehmood, Q., Sial, M., Riaz, M., and Shaheen, N. (2019). Forecasting the production of sugarcane crop of Pakistan for the year 2018–2030. Using Box-Jenkin's methodology. *J. Anim. Plant Sci,* 29:1396–1401.

Paswan, S., Paul, A., Paul, A., and Noel, A. S. (2022). Time series prediction for sugarcane production in Bihar using ARIMA & ANN model. *The Pharma Innovation Journal,* 11(4):1947– 1956.

Bezuidenhout, C. and Singels, A. (2007). Operational forecasting of South African sugarcane production: Part 1 – system description. *Agricultural Systems,* 92(1):23–38.

Luciano, A. C., Picoli, M. C. A., Duft, D. G., Rocha, J. V., Leal, M. R. L. V., and le Maire, G. (2021). Empirical model for forecasting sugarcane yield on a local scale in brazil using landsat imagery and random forest algorithm. *Computers and Electronics in Agriculture,* 184:106063.

Gupta, S. K. and Agarwal, A. P. (2021). Predicting total sugar production using multivariable linear regression. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 465–469.

Megha, J., Havaldar, Y., Pavithra, N., Jyoti, B., and Kumar, V. K. (2019). Identification of the best model for forecasting of sugar production among linear and non-linear model. *Int. J. Curr. Microbiol. Appl. Sci,* 8(3):2556–2560.

Kihara, P. N. (2013). *Estimation of finite population total in the face of missing values using model calibration and model assistance on semiparametric and nonparametric Models.* PhD thesis.

Kwamboka, L., Orwa, G. O., and Muga, Z. M. (2019). Forecasting monthly sugar cane yields using Box-Jenkin's predictive models in Kenya.

UNDER PEER REVIEW

Mwanga, D., Ong'ala, J., and Orwa, G. (2017). Modeling sugarcane yields in the kenya sugar industry: A sarima model forecasting approach. *International Journal of Statistics and Applications*, 7(6):280–288.

Yuma, E., Jossy, N., and Gusite, B. (2023). Development of a machine learning regression model for accurate sugarcane crop yield prediction, jinja – Uganda. *Journal of Applied Sciences, Informaon and Compung (JASIC)*, 25–33.

Shang, H. L. and Hyndman, R. (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation,* 81:1310–1324

Kan, H. J., Kharrazi, H., Chang, H.-Y., Bodycombe, D., Lemke, K., and Weiner, J. P. (2019). Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PloS one*, 14(3):e0213258.

Leith, U. (2024). Modeling a periodic signal using fourier series. *Journal of Applied Mathematics and Physics*, 12(3):841–8

Saputro, D., Sukmayanti, A., and Widyaningsih, P. (2019). The nonparametric regression model using Fourier series approximation and penalized least squares (pls) (case on data poverty in East Java). *Journal of Physics: Conference Series*, 1188:012019.

UNDER PEER REVIEW