Short-term forecasting of confirmed daily COVID-19 cases in the Southern African Development Community region

Claris Shoko[1], Caston Sigauke[2], Peter Njuho[3]

1. Department of Mathematics and Computer Sciences, Great Zimbabwe University. Private Bag 1235, Masvingo.
2. Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou, 0950, South Africa.
3. Department of Statistics, University of South Africa, South Africa.

Abstract

Background: The coronavirus pandemic has resulted in complex challenges worldwide, and the Southern African Development Community (SADC) region has not been spared. The region has become the epicentre for coronavirus in the African continent. Combining forecasting techniques can help capture other attributes of the series, thus providing crucial information to address the problem.

Objective: To formulate an effective model that timely predicts the spread of COVID-19 in the SADC region.

Methods: Using the Quantile regression approaches; linear quantile regression averaging (LQRA), monotone composite quantile regression neural network (MCQRNN), partial additive quantile regression averaging (PAQRA), among others, we combine point forecasts from four candidate models namely, the ARIMA (p, d, q) model, TBATS, Generalized additive model (GAM) and a Gradient Boosting machine (GBM).

Results: Among the single forecast models, the GAM provides the best model for predicting the spread of COVID-19 in the SADC region. However, it did not perform well in some periods. Combined forecasts models performed significantly better with the MCQRNN being the best (Theil's U statistic=0.000000278).

Conclusion: The findings present an insightful approach in monitoring the spread of COVID-19 in the SADC region. The spread of COVID-19 can best be predicted using combined forecasts models, particularly the MCQRNN approach.

Keywords: Combined Forecasts, LQRA, PLAQR, OPERA, Quantile Regression Neural Networks, COVID-19.

DOI: https://dx.doi.org/10.4314/ahs.v22i4.60

Cite as: *Shoko C, Sigauke C, Njuho P. Short-term forecasting of confirmed daily COVID-19 cases in the Southern African Development Community region. Afri Health Sci.2022;22(4). 534-550. https://dx.doi.org/10.4314/ahs.v22i4.60*

Introduction

Coronaviruses, a large family of viruses, can cause illnesses that range from the common colds to much more severe diseases like SARS, Middle East respiratory syndrome, and COVID-19[1]. Signs of the COVID-19 disease may include fever, cough, shortness of breath and general breathing difficulties, organ failure, and even death. Some Chinese health authorities stated that coronavirus is likely to be transmitted from one person to another even before any symptoms (spread during the incubation period), making the epidemic difficult to prevent and control. This poses a severe threat to society as a whole.

Corresponding author:
Claris Shoko,
Department of Mathematics and Computer Sciences, Great Zimbabwe University. Private Bag 1235, Masvingo.
Email: cshoko@gzu.ac.zw

The Southern Africa region has been hit hardest by the COVID-19 pandemic in Africa, thus the epicentre of the coronavirus in the African continent[2]. Sixteen countries in the southern part of Africa constitute the SADC region namely Angola, Botswana, Eswatini, Comoros, Democratic Republic of Congo (DRC), Lesotho, Madagascar, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Tanzania, Zambia, and Zimbabwe. By February 2021 the SADC region had accounted for half of the reported cases in Africa. Of the five African countries accounting for close to 76% of new infections, three are members of the SADC, namely South Africa, Zambia, and Namibia[3].

Forecasting is a part of statistical modelling widely used in various fields because of its benefits in decision making[4]. Forecasting is related to the formulation of models and methods that can be used to predict the future trend of uncertain situations. In most cases, one model is selected based on selection criteria, for example, the AICc, hypothesis testing and/or graphical inspection[5]. The model is considered to have the best performance accuracy forecast future values. However, this concept is only true if the model's premises are valid when applying it to the data. following Martinez et al.[6]: forecast models are based on the assumption that "the most reliable way to predict the future is to understand the present," and, for this reason, these models do not say what will happen in the future, but say what can happen if the conditions observed in the present do not change. Thus, a bad model may casually predict the future better than a good model if the observed conditions in the present change radically in the future. A single technique cannot efficiently use a great deal of information due to the complexity of some time series. According to Bates and Granger,[7] forecasting techniques have high accuracy when performing combination is achieved. Individual forecasting techniques based on different approaches capture distinctive characteristics of the series and allow for the combination to benefit from such characteristics[8]. A combined forecast allows for gathering available information, hence increasing the accuracy of the final forecast[9].

ARIMA models have commonly been used in time series data analysis and forecasting and in predicting COVID-19 spread in particular[10,11]. Even though the ARIMA model is useful and powerful in time series analysis, sometimes it is difficult or rather cumbersome to identify the appropriate model for the data[12]. Recent results in machine learning show an improved performance of the final model not by choosing the model structure expected to predict the best but by creating a model whose results is the combination of the output of models having different formats. The various machine learning techniques applied are:

- **Generalized Additive Model (GAM):** These models assume that the mean of the response variable depends on an additive predictor through a link function. GAMs permit the response probability distribution to be any member of the exponential family of distributions.
- **Gradient Boosting Machine (GBM):** A decision tree model is chosen typically as a base model; however, an ensemble of such prediction model is chosen
- **Quantile Regression (QR) Models:** Standard linear regression focuses on finding a conditional mean function describing a linear relationship between the predictor and the independent variable(s). QR models look at different quantiles of the response defined by the conditional quantile function.

i. *Linear Quantile Regression (LQR) model:* The quantile regression model was introduced by Koenker and Bassett[39], which models the relationship between predictor X and the conditional quantiles of Y given $X = x$. The linear quantile regression model complements the linear mean regression model if the error terms in the mean regression model are heteroscedastic.

ii. *Quantile Regression Neural Network (QRNN) Model:* the theoretical support for the use of quantile regression within an Artificial Neural Network to estimate potentially nonlinear quantile models.

iii. *Monotone Composite Quantile Regression Neural Network (MCQRNN) model:* estimates simultaneously multiple non-crossing quantile functions and allows optional monotonicity constraints

iv. Partial Additive Quantile Regression (PLAQR) averaging: Estimation, prediction, thresholding, transformation, and plotting for partial linear additive quantile regression

.

- **Online Prediction by ExpeRt Aggregation (OPERA):** Considers a sequence of observations from a bounded time series to be predicted step by step. At each instant t, a finite set of experts, provides predictions x of the next observation in y.

Forecast combination methods exist and previous studies on forecasting show that combining forecasts generated from different models can considerably improve forecasting performance over single forecast models[13]. According to Zou and Yang5 combined forecasting improves accuracy performance. A fact confirmed by Adhikari and Agrawal14 is that combined forecasts lower forecast errors than individual models[14]. To the best of our knowledge. There is relatively no evidence of forecast combination in the context of COVID-19. This study introduces an efficient, flexible nonlinear quantile regression model, the monotone composite quantile regression neural network model to the modelling of the spread of COVID-19 in the SADC region.

In the next section we outline the methodology used in the study as well as formulation of the model. This is followed by the Results section where we explore the COVID-19 data for the SADC region and interpret results from fitted models. Lastly, we discuss and make conclusion based on the findings.

Methods

Data

In this study, we use an openly available daily number of confirmed cases of COVID-19 reported by Our World in Data (https://www.ourworldindata/coronavirus-source-data) from 7 March 2020 to 25 August 2021. We extract data from the daily confirmed cases for the SADC region. The SADC region is presented in Figure 1 below.

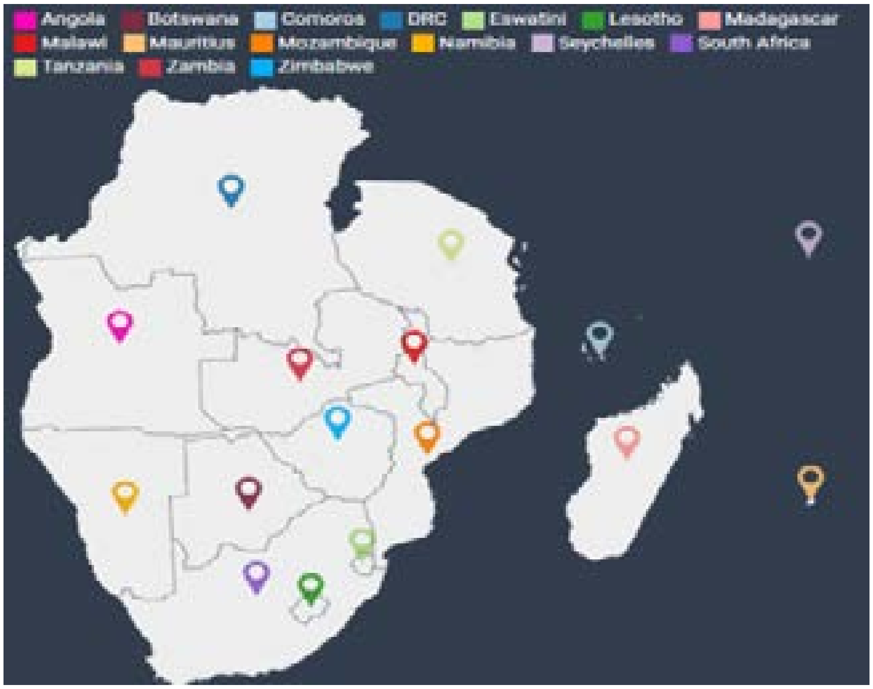

Figure 1: The spread of COVID-19 in the SADC region

Modelling and prediction of the spread of COVID-19 in the SADC region are done using the R packages: 'forecast'[15] for fitting the ARIMA and TBATS models, 'gam'[16] for fitting the generalized additive models, 'gbm'[17] for fitting the stochastic gradient boosting model, 'qrnn'[18] for fitting the linear quantile regression averaging and monotone composite quantile regression neural network model, 'plaqr'[19] and 'opera'[20]. Figure 2 provides a schematic summary of the analysis procedure for predicting the spread of COVID-19 in the SADC region.

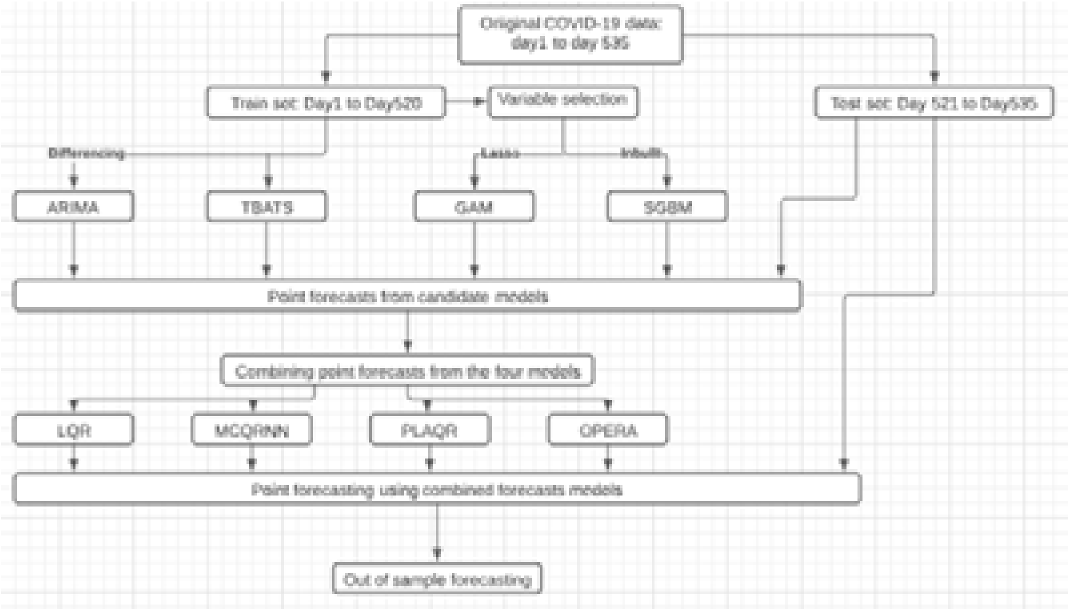
Original COVID-19 data: day1 to day 535
Train set: Day1 to Day520
Variable selection
Test set: Day 521 to Day535
Differencing
Lasso
Inbuilt
ARIMA
TBATS
GAM
SGBM
Point forecasts from candidate models
Combining point forecasts from the four models
LQR
MCQRNN
PLAQR
OPERA
Point forecasting using combined forecasts models
Out of sample forecasting

Figure 2: Schema for predicting the spread of COVID-19 in the SADC region.

Single Forecasting Methods

Non-Seasonal Autoregressive Integrated Moving Average (ARIMA) models

The growth of daily COVID-19 disease cases for the SADC region falls into the category of time series data, easily captured by an integrated model such as the ARIMA[21]. ARIMA models describe series that exhibit a trend that differencing can remove.

SARIMA Model

We have the general SARIMA model represented analytically as:

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D y_t = \theta(B)\Theta(B^s)a_t,\ a_t \sim N0,\sigma^2), \tag{1}$$

where y_t represents the SADC confirmed daily cases on day t, $t = 1,\dots,n$, $a_t \sim N(0,\sigma^2)$ is the error term at time t, s is the seasonal length, B is a backshift operator ($Bz_t = z_{t-1}$). $\phi(B) = (1-\phi_1 B - \dots - \phi_p B^p)$ is the non-seasonal autoregressive (AR) operator, $\Theta(B) = (1-\Theta_1 B - \dots - \Theta_q B^q)$ is the seasonal AR operator, $\theta(B^s) = (1-\theta_1 B^s - \dots - \theta_Q B^{Qs})$ is the non-seasonal moving average (MA) operator, $\theta(B^s) = (1-\theta_1 B^s - \dots - \theta_Q B^{Qs})$ is the seasonal MA operator. ∇^d and ∇_s^D are the non-seasonal and seasonal difference operators of order d and D respectively, where $\nabla^d = (1-B)^d$ and $\nabla_s^D = (1-B^s)^D$.

TBATS model

The TBATS model uses the Box-Cox transformation, exponential smoothing, trigonometric seasonality and ARMA errors[1]. It is generally used for forecasting time series with complex seasonal patterns. The components of the model are:

(i) The Box-Cox transformation

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega}; & \omega \neq 0 \\ \log y_t & \omega = 0 \end{cases} \tag{2}$$

where y_t is the confirmed daily cases on day t, ω is the transformation parameter and denotes the natural logarithm.

(ii) Deterministic and stochastic trend

$$y_t^{(\omega)} = \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-1}^{(i)} + d_t$$

$$\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = (1-\phi)b + \phi b_{t-1} + \beta d_t \tag{3}$$

where T denotes the number of seasonal patterns ℓ_t is the local trend in period t, b represents the long-run trend, b_t denotes the short-run trend in period $s_{t-1}^{(i)}$, represents the i^{th} seasonal component at time $t-1$, d_t denotes the ARMA (p, q) process and α, β and ф are smoothing parameters.

(iii) Trigonometric seasonality

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t$$

$$s_{j,t}^{*(i)} = -s_{j,t-1} \sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{4}$$

wher $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$ are smoothing parameters and $\lambda_j^{(i)} = \frac{2\pi_j}{m_i}$ with m_i representing the period of the seasonal cycle.

(iv) ARMA errors

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t, \tag{5}$$

where φ_i, θ_i denote the autoregressive and moving average parameters, respectively and ε_t is a white noise process.

The components (i) – (iv) put together give the TBATS model.

Generalized additive models

Let y_t denotes the SADC confirmed daily cases on day t, $t = 1,\dots,n$ with the corresponding covariates $x_{t1}, x_{t2}, \dots, x_{tp}$, where p represent the number of variables. The generalized additive model is then written as:

$$y_t = \beta_0 + \sum_{j-1}^{p} s_j(x_{tj}) + \varepsilon_t, \tag{6}$$

where β_0 is a constant parameter, s_j are smooth functions and ε_t are independent and identically distributed ($i.i.d$) error terms. Equation (1) is estimated using penalized cubic splines[22,23] given as:

$$\min_{s_j}\left[\sum_{t=1}^{n}\left(y_t-\beta_0-\sum_{j=1}^{p}s_j(x_{tj})\right)^2+\sum_{j=1}^{p}\lambda_j\left(\int\left(f''(x)\right)^2dx\right)\right]$$

(7)

The penalty parameter controls the degree of smoothness $\Lambda=(\lambda_j,j=1,\ldots,p)$ which is optimized using the generalized cross-validation criterion (GCV)[23]. The smooth function, $b_i(x)$, is a sum of basis functions, , together with their regression coefficients β_i and is given by $s_j(x)=\sum_{i=1}^{q}b_i(x)\,\beta_i$, where q denotes the basis dimension.

Variable Selection

To reduce the problem of multicollinearity amongst the predictor variables we use the least absolute shrinkage and selection operator (Lasso). Lasso formulation is given as[24,25]:

$$\hat{\beta}_j=\underset{\beta_j}{\operatorname{argmin}}\left[\sum_{t=1}^{n}\left(y_t-\beta_0-\sum_{j=1}^{p}\beta_j x_{tj}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right]$$

(8)

where λ is the shrinkage factor. The shrinkage factor, which lies between 0 and 1, is given by

$$\lambda=\frac{t}{\sum_{j=1}^{p}|\beta_{tj}|}$$

. See Tibshirani[24] and Friedman et al.[25] for a detailed discussion of Lasso.

Stochastic Gradient Boosting Method (SGBM)

Gradient boosting (GB) is a machine learning technique that fits an additive model in a
stage-wise way. The additive model can take the form given in Equation (8)[26].

$$f(x)=\sum_{m=1}^{M}\beta_m b(x;\gamma_m),$$

(9)

where $b(x;\gamma_m)\in\mathfrak{R}$ are functions of x which are characterised by the expansion parameters γ_m,β_m . The parameters β_m and γ_m are fitted in a stage-wise way, a process which slows down over-fitting[26]. Stochastic gradient boosting (SGB) is an extension of GB in which a random sample of the training data set is taken without replacement. See Friedman for a detailed discussion of the gradient boosting method[27].

Combining Forecasts

Combining forecasts was first developed by Bates and Granger[7], who argued that combined forecasts improve forecast over the single model forecast. Suppose the point forecasts from the ARIMA, TBATS, GAM, and SGBM models are combined so that we have a vector

$$\hat{\boldsymbol{y}}_{t+h}=\left(\hat{y}_{t+h}^{(ARIMA)},\hat{y}_{t+h}^{(TBATS)},\hat{y}_{t+h}^{(GAM)},\hat{y}_{t+h}^{(SGBM)}\right)'$$

(10)

Then, the combined forecasts for this vector are obtained using the LQRA, MCQRNN, PLAQR, and OPERA approaches. The accuracy of the performance of these models is checked by comparing their respective RMSE, MAPE and Theil's U statistics.

Quantile Regression Averaging (QRA)

In the standard QR setting, individual point forecasts are used as independent variables and the corresponding target variable as the dependent variable[28]. The relationship between the predictor and the independent variable(s) is not described with a single slope parameter just like in linear regression models, but a set of parametersβ_τ dependent on the quantile τ must be estimated. We define the τ^{th}regression quantile ( ) as any solution, to the quantile regression minimization problem[29]:

$$\min_{\beta_\tau\in\mathbb{R}}\sum_{i=1}^{n}\rho_\tau\left(y_i-\xi_\tau(x_i,\beta_\tau)\right),$$

(11)

where$\rho_\tau(y_i-\xi(x_i,\beta_\tau))$ is a function of τ and $y_i-\xi_\tau(x_i,\beta_\tau)$. This kind of loss function is most often called check or pinball loss function and is defined as follows:

$$\rho_\tau(y_i-\xi(x_i,\beta_\tau))=\begin{cases}\tau(y_i-\xi(x_i,\beta_\tau)) & \forall y_i\geq\xi(x_i,\beta_\tau)\\(\tau-1)(y_i-\xi(x_i,\beta_\tau)) & \forall y_i<\xi(x_i,\beta_\tau)\end{cases},$$

(12)

where$\{x_i:i=1,\ldots,n\}$ denotes a sequence of explanatory variable and $\xi_\tau(x_i,\hat{\beta}_\tau)$ is formulated as a linear function of parameters. The LQRA model is given by

$$Y_{t+h}=\beta_0(\rho)+\boldsymbol{X}'_{t+h}\beta_1(\rho)+U_{t+h},$$

(13)

where $\Pr(U_{t+h}\leq 0\mid X_{t+h}=x)=\rho\in(0,1),E(U_{t+h}^2\mid X_{t+h}=x)=\sigma_\rho^2(x)$ and ρ is the probability mass of interest. $\boldsymbol{X}_{t+h}$ is a vector of covariates for the $t+h$ forecasted value from the fitted ARIMA model, TBATS model, GAM, and SGBM, i.e., multivariate quantile regression model. The unknown parameter vectors appearing in the above equation can be solved from the following optimization problem:

$$\min_{\beta\in\theta}\sum_{t=1}^{n}\rho_\tau(y_{t+h}-\beta_0-X_{t+h}\beta_1),$$

(14)

where $\rho_\tau(x)=\tau xI(x\geq 0)-(1-\tau)xI(x<0)$ is the check function. Figure 3 presents a schematic expression of the quantile regression average. Figure 3 shows the link between the individual point forecasts through the quantile regression to the combined interval forecast.

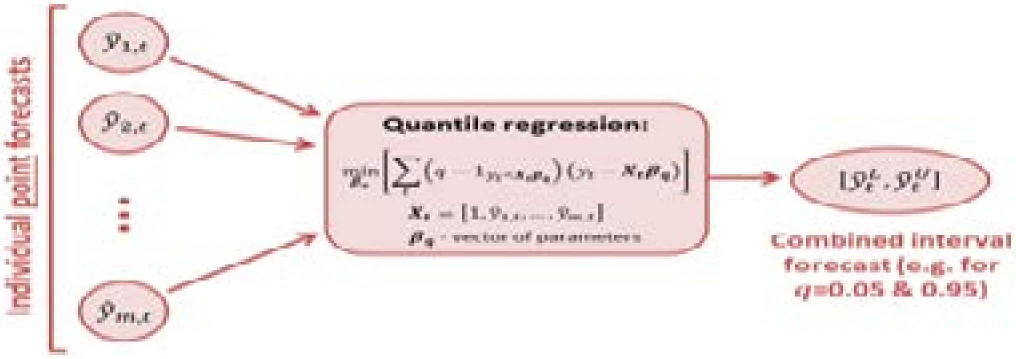
Individual point forecasts
$y_{1,t}$
$\hat{y}_{2,t}$
$\hat{y}_{m,t}$
Quantile regression:
$\min_{\beta_q}\left[\sum_t (q - 1_{y_t < X_t\beta_q})(y_t - X_t\beta_q)\right]$
$X_t = [1, \hat{y}_{1,t}, \ldots, \hat{y}_{m,t}]$
β_q - vector of parameters
$[\hat{y}_t^L, \hat{y}_t^U]$
Combined interval forecast (e.g. for q=0.05 & 0.95)

Figure 3: Quantile regression averaging.

Monotone Composite Quantile Regression Neural Network (MCQRNN)

MCQRNN is a novel form of quantile regression that can be used to simultaneously estimate multiple non-crossing. It combines elements drawn from the QRNN model[30,31], the monotone multilayer perception (MMPP)[32], the composite QRNN[33], the expectable regression network[34] and the generalized additive neural network[35]. Cannon18 gives an elaborate explanation on the formulation of the MCQRNN.

Combining prediction intervals

Robust prediction intervals are known to be produced from combining prediction limits from various models[36,37,38]. We use the simple average and median methods for combining the prediction limits. The simple average method uses the arithmetic means of the prediction limits from the forecasting models. Thus, expressed as

$$L_{Av} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{L}_{ij}, \quad U_{Av} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{U}_{ij}\,. \tag{15}$$

The median method is known to be less sensitive to outliers and is given in Equation 16

.

$$L_{Md} = \mathrm{median}(L_{1,\ldots,}L_m), \quad U_{Md} = \mathrm{median}(U_{1,\ldots,}U_m) \tag{16}$$

For each of the models, $M_j, j = 1,\ldots,k$, we compute the prediction interval widths (PIWs),
which we denote by $\mathrm{PIW}_{ij}, i = 1,\ldots,m; j = 1,\ldots,k$ and calculate as

$$\mathrm{PIW}_{ij} = \mathrm{U}_{ij} - \mathrm{L}_{ij}$$

where U_{ij} and L_{ij} are the upper and lower limits of the prediction interval, respectively. Various indices are used to evaluate the reliability of prediction intervals (PIs). In this study we use the prediction interval normalised average width (PINAW). We express the PINAW, an index that check if the required value is covered by the prediction interval as

$$PINAW = \frac{1}{m(\max(\mathrm{PIW}_{ij}) - \min(\mathrm{PIW}_{ij}))}\sum_{i-1}^{m}(\mathrm{PIW}_{ij}) \tag{17}$$

Using PINAW we compare different models and then determine the one that possesses the smallest percentage value.

Empirical Results

Exploratory data analysis

We use an openly available daily number of confirmed cases of COVID- 19 reported by Our World in Data (www.ourworldindata/coronavirus-source-data) from 7 March 2020 to 25 August 2021. The number of daily reported cases for the SADC region ranged from 0 to 32 321. From 7 March 2020 to 3 August 2021, the average number of reported cases was 6 581 per day.

We further perform a univariate data analysis for the reported daily COVID-19 cases by plotting the time series data and the density plot, normal Q-Q plot and the Box plot as shown in Figure 4. The plots check for the normality assumption in the time series data.

Figure 4(a) presents the time series trend for the daily COVID-19 cases in three phases. The first phase has the lowest peak and the peaks increase with time, with the third phase having the highest peak. Figure 4(b) presents the density plot, which shows that the series is positively skewed, thus, not normally distributed. Figure 4(c) is the Q-Q plot. The deviations from the diagonal line in the normal Q-Q plot imply that the data extend farther out than expected under normality. A correlation matrix showed some highly correlated variables (see the correlation matrix given in the supplementary material). We use Lasso (discussed in Section 2.2.4) to reduce the multicollinearity problem in variable selection.

Predictive modelling for the reported daily COVID-19 cases in the SADC region

The series is relatively long and can be divided into train and test sets. The training set constitutes the first 520 observations and 521 to 535 represent the test set. In the next section, we fit the ARIMA model, TBATS, Generalized Additive Model and Stochastic Gradient Boosting for the training set and use the fitted models to check if they fit the test set well.

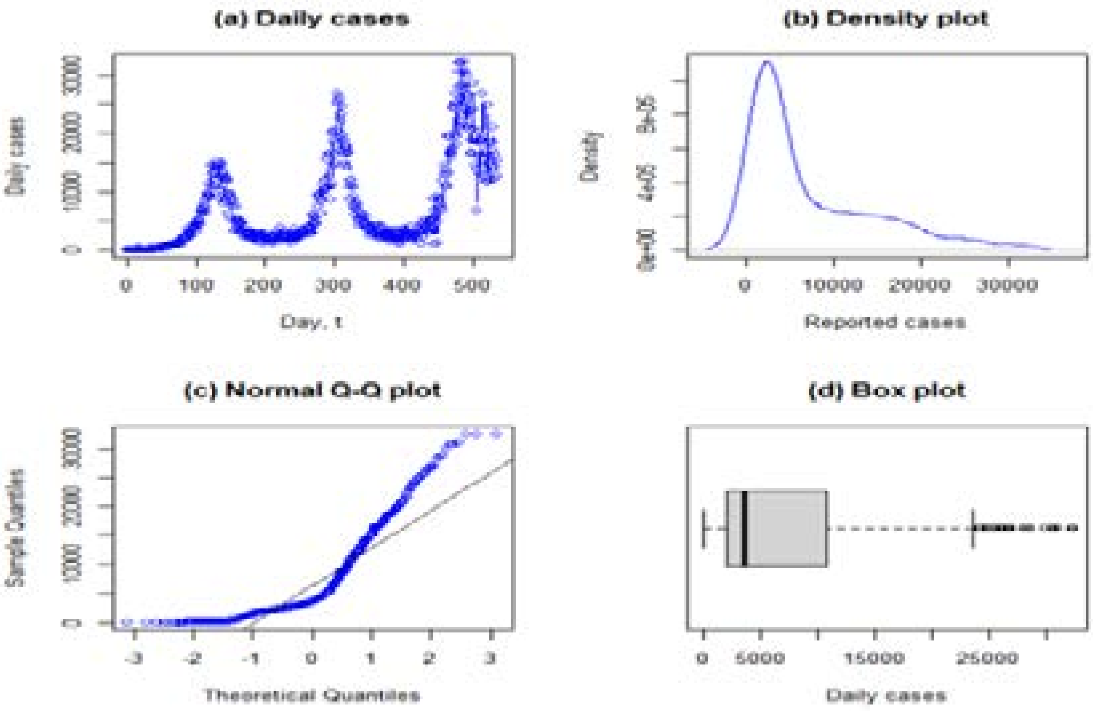

Figure 4: Normality checks for the daily COVID-19 series

Time series ARIMA model

We start by testing for the stationarity of the original time series data and that of the differenced time series data. This is done using the augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test at a 5% level of significance. ADF test: the null hypothesis is that the data are non-stationary and non-seasonal. KPSS test: the null hypothesis is that the data are stationary and non-seasonal. A plot of the residuals autocorrelation function (ACF) is also used to investigate the stationarity of the original time series. Figure 5 present the results.

The ACF plot shows that all autocorrelations are outside the threshold limit. This indicates that the original series is not white noise. A Scatter plot of residuals shows that yt is correlated to yt-1.

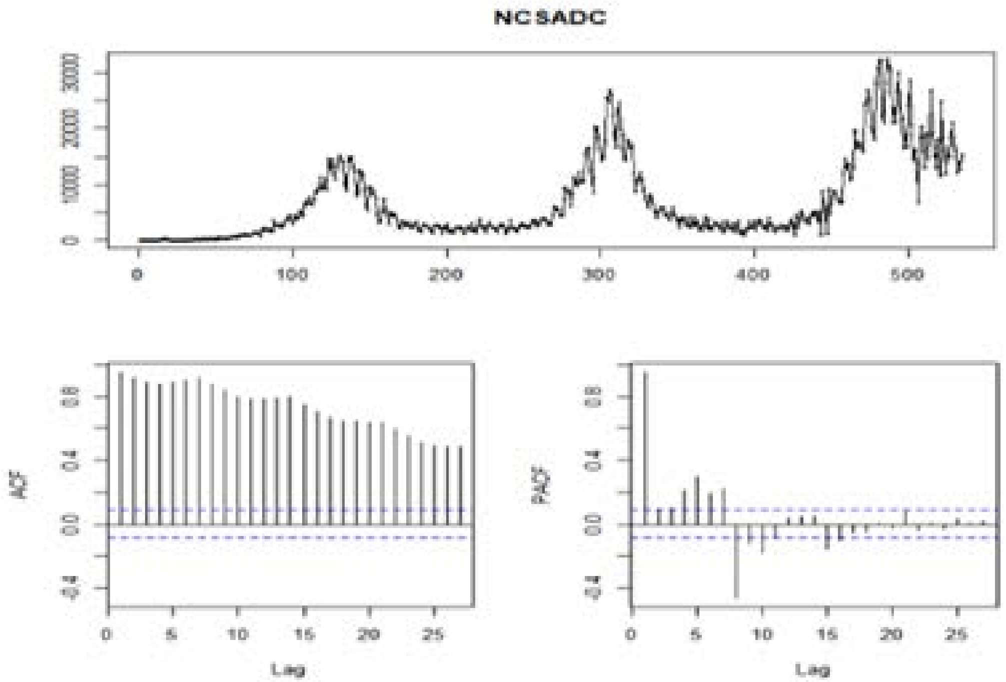

Figure 5: Display of the time series for the daily COVID-19 cases in the SADC region.

Further diagnostic of residuals using the Box-Pierce test and the Box-Ljung test return small p-values <2.2e-16, suggesting that the original series is not white noise. The ADF (-2.2052, p-value=0.4915) and the KPSS tests (1.9668, p-value=0.01) show that the original time series is not stationary. After the first difference of the data, both the ADF and KPSS test show that the differenced series is stationary in its mean and variance at 5% level, p-values= 0.01 and 0.1 respectively. Therefore, we adopt d = 1 for ARIMA (p, d, q) model.

The ACF and PACF charts for the differenced time series, though not shown, were used to help select the candidate ARIMA models by observing the spikes in the ACF and PACF. The spikes in the PACF plot suggest an AR[7] and ACF suggest a MA[7]. Thus, the initial candidate model takes the form of ARIMA[7,1,7]. We consider several ARIMA models, including the auto-selected ARIMA model and assess the accuracy of their performance, based on the AICc. The ARIMA[14,1,8] with the lowest AICc (AICc=9102.03) compared to all the other ARIMA models is considered the best ARIMA model for predicting the spread of COVID-19 in the SADC region. At 5% significance level, the Box-Ljung test (-squared = 8.9064, df = 20, p-value = 0.984) shows that the residuals for the fitted ARIMA[14,1,8] model are stationary.

TBATS

The best TBATS model for the confirmed daily COVID cases for the SADC region is a BATS (1, {3,2}, 0.886, -) where Box-Cox transformation is 1 (doing nothing).

Generalized additive model

Before fitting the GAM and the GBM we created some covariates where

$$t = 3,4,5,\ldots,515;\ tsq = t^2 = 9,16,25,\ldots,515^2;\ tcub = t^3 = 27,\ldots,515^3;\ day = 1,2,3,\ldots,7,1,2,\ldots,7,\ldots,12;\ tday = t * day;\ month;\ tmonth = t * month;\ tsqmonth = t^2 * month;\ tcubmonth = t^3 * month;\ lag1_i = x_i - x_{i-1};\ lag2_i = lag1_i + lag1_{i-1},\ \text{where } i = i\text{th observation};\ daymonth = day * month$$

The model for the GAM and GBM is:

$$y_t = \beta_0 + \beta_1 s(t) + \beta_2 s(tsq) + \beta_3 s(tcub) + \beta_4 s(day) + \beta_5 s(month) + \beta_6 s(t * day) + \beta_7 s(t * month) + \beta_8 s(tsq * month) + \beta_9 s(tcube * month) + \beta_{10} s(day * month) + \beta_{11} s(lag1) + \beta_{12} s(lag2) + \epsilon_t$$

(18)

Before fitting using Lasso shrinkage approach discussed in Section 2.2.4, we fitted the GAM because it does not have an inbuilt mechanism for variable selection. Table 1 presents the results for the fitted model (18).

Table 1

The results in Table 1 show that the variables $t * cub$, day, $tsq * month$ and $day * month$ do not contribute significantly to the GAM and hence exclude them in the building of the GAM model. Figure 6 presents plots

Table 1: Selection of variables for the GAM via the Lasso approach

Variable	Regression coefficient
Intercept	3360
t	178
tsq	-0.498
tcub	.
day	.
month	-1880
*t * day*	0.961
*t * month*	-2.24
*tsq * month*	.
*tcub * month*	0.00008
*day * month*	.
*lag*1	0.242
*lag*2	0.461

that check for the normality assumptions in the fitting of GAM.

The graphs in Figure 6 theoretical quantile plot's tails. This suggests a rather heavy tail distribution, different from a normal distribution. In addition, the plot of residuals shows strong heteroscedasticity.

Stochastic Gradient boosting method (SGBM)

Unlike the GAM, the SGBM has an inbuilt mechanism for selecting variables. Table 2 shows the influence of the variables on the fitted SGBM. The variables are ranked

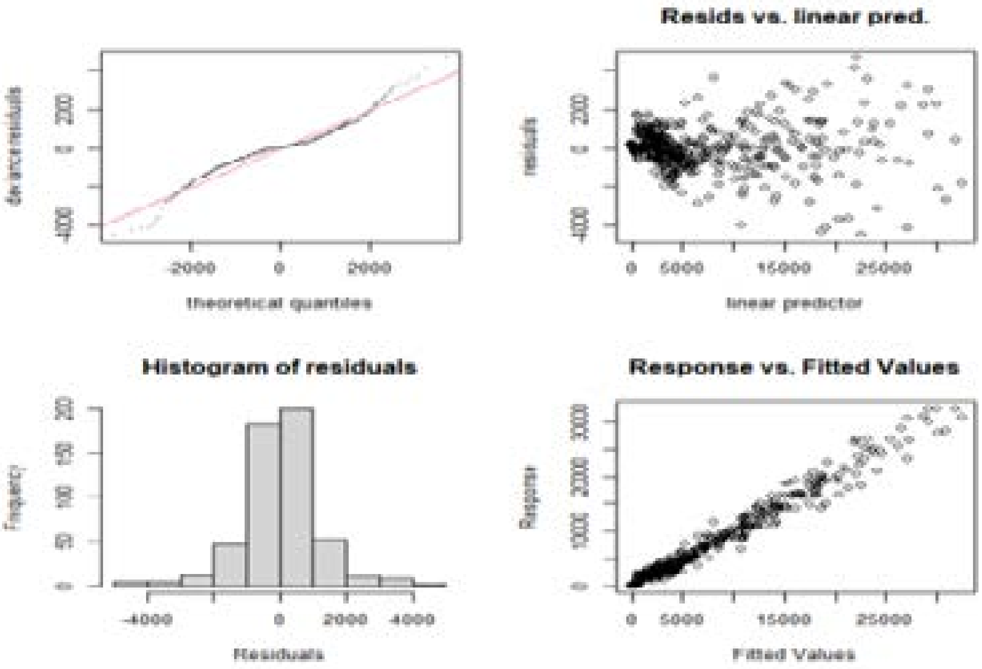

Figure 6: Normality checks for the fitted Generalized Additive Model

from the most influential to those that do not influence the fitted model.

Table 2 results show that the variables , and   have a zero influence on the fitted model.

We display the predicting outputs for all the fitted models namely the ARIMA (14,1,8), TBATS, GAM and Stochastic gradient boosting model (SGBM) in Table 3. Included in the table are the forecast performance measures namely the RMSE, MAE, MAPE and Theil's U results.

Table 2: Selection of variables for the SGBM

Variable	Relative Influence
t	29.46118
$tsq * month$	19.03578
$month$	16.43531
$lag2$	15.35461
$tcub * month$	8.49682
$t * month$	5.958118
$lag1$	2.630775
$t * day$	1.43863
$day * month$	0.679078
day	0.509696
tsq	0
$tcub$	0

Among the four models, the GAM performed best in the prospective forecasting of daily COVID-19 cases over the following 15 days, with the smallest values of RMSE (1459.164), MAE (1158.378) and MAPE (7.81968). The SGBM showed better goodness of fit than the ARIMA^{14,1,8} and TBATS models. For the forecast accuracy, the ARIMA^{14,1,8} showed a greater RMSE (2602.141) than the GBM (1906.75), as well as a greater MAE (2372.979 vs. 1658.871) and MAPE (16.2012 vs. 10.54381).

Table 3: Forecast performance measures for the single forecast models

Model	RMSE	MAE	MAPE	Theil's U
SGBM	1906.75	1658.871	10.54381	0.3356
GAM	1459.164	1158.378	7.81968	0.2387
TBATS	2610.837	2091.848	14.19489	0.3880
ARIMA (14,1,8)	2602.141	2372.979	16.2012	0.4747

Table 4 presents the forecasting results from the fitted ARIMA^{14,1,8}, TBATS, GAM and SGBM models.

The results in Table 4 show that the GAM, the model that performs the best compared to the rest, predicts the lowest number of COVID-19 cases than all the other models.

Figure 7 (a-d) displays comparison plots of the 15-days forecast from the training set and the test set (observed series) of the fitted models. The black line represents observed/actual values of the test set. Figure 7(a) pres-

Table 4: Forecasting from the ARIMA (14,1,8), TBATS, GAM and SGBM

		$\hat{y}_{t+h}^{(ARIMA)}$	$\hat{y}_{t+h}^{(TBATS)}$	$\hat{y}_{t+h}^{(SGBM)}$	$\hat{y}_{t+h}^{(GAM)}$
Forecasts(*h*)	1	15011.18	18366.55	17254.98	14445.34
	2	24344.25	20869.12	24742.36	21662.12
	3	20593.48	20632.58	23556.69	18847.04
	4	18784.38	18209.17	17652.61	13026.52
	5	15970.79	15357.42	17374.73	11163.43
	6	18708.81	14073.74	17915.55	14706.16
	7	15575.50	15146.23	20528.35	16241.85
	8	15556.11	17594.13	21731.04	17235.09
	9	23111.45	19438.31	24227.89	18920.46
	10	19832.68	19204.77	17110.47	14820.07
	11	18226.37	17019.44	17652.61	13729.31
	12	14903.75	14480.82	17374.73	10198.84
	13	17213.17	13430.18	18457.68	12865.33
	14	15787.95	14563.86	16915.97	13778.62
	15	15048.45	16932.37	18629.08	15226.46

ents the forecast from the training set for the ARIMA^{14,1,8} model and the test set, (the observed series). Figure 7(b) presents the prediction from the TBATS model. Figure 7(c) presents the prediction from the SGBM and Figure 7(d) represents the prediction from the GAM.

Figure 7(a-d) plots show that prediction from the GAM gives a better fit of the test set (observed set) followed by the SGBM, the ARIMA^{14,1,8} and lastly the TBATS model. Although the GAM is the best of the four fitted models, it does not perform well in certain periods while other models perform better in other periods. Therefore, we suggest combining forecasts from the ARIMA^{14,1,8} model, TBATS, GAM, and the SGBM to improve forecasts over individual models.

Combining forecasts

Let the vector of forecasters from the ARIMA (14,1,8) model, TBATS model, GAM and SGBM be

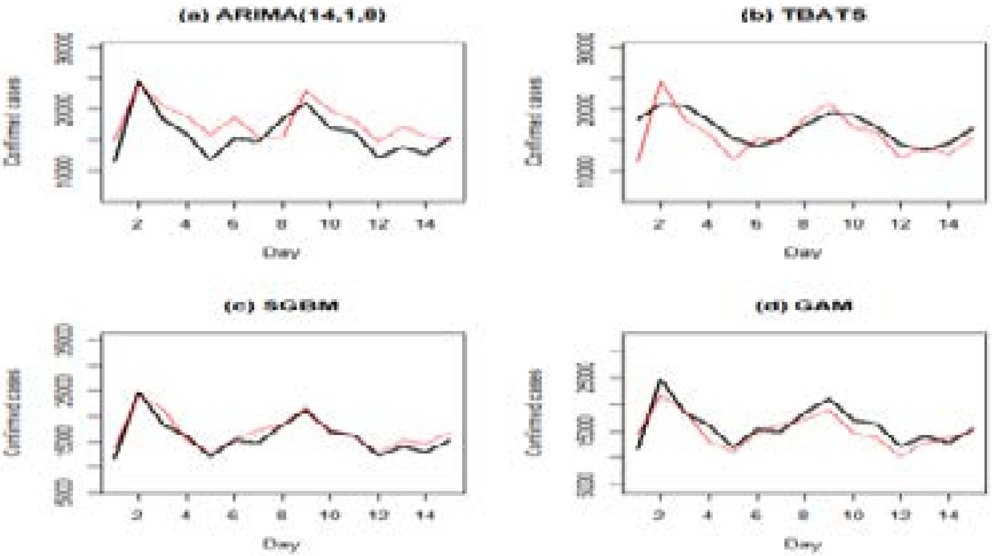

Figure 7: Plot of the 15-day forecast from the training set (predicted values) and the test set, the black line represents observed/actual values of the test set

$$\hat{y}_{t+h} = \left(\hat{y}^{(ARIMA)}_{t+h}\hat{y}^{(TBATS)}_{t+h}, \hat{y}^{(GAM)}_{t+h}, \hat{y}^{(SGBM)}_{t+h}\right)'$$

We combine the forecasters using four different methods namely the Linear Quantile Regression (LQR) model, Monotone Composite Quantile Regression Neural Network (MCQRNN) model, Partial Linear Additive Quantile Regression (PLAQR) averaging, and Opera. For the first three (LQR, MCQRNN, and PLAQR) the value of the conditional quantile, gives better forecasts. Table 5 presents results of the comparison of the RMSE, MAE, MAPE, and Theil's U statistic for the fitted models, used to check the accuracy of the performance of the combined forecasts.

Table 5: Forecast performance measures for the combined forecast models

Forecast combination model	RMSE	MAE	MAPE	Theil's U
LQR	1196.701	703.655	4.014866	0.1896
MCQRNN	0.001323	0.001049		
PLAQR	351.5644	111.6311	0.72325	0.0611
OPERA	1244.245	936.4002	6.264443	0.1970

Results in Table 5 indicate that the RMSE and MAPE for combination forecast models (LQR, MCQRNN, PAQR, OPERA) are lower than the RMSE and MAPE for the single forecast models (ARIMA, TBATS, GAM, GBM). Thus, forecast combinations improve the accuracy over the single forecast models for the daily COVID-19 cases for the SADC region. The MCQRNN has the lowest RMSE=380.931 and MAPE=0.808865, compared to the other models. Theil's U statistic for the MCQRNN model is close to 0 suggesting a perfect fit for the forecast. Figure 8 shows a further comparison of the performance of the combination forecast models. We visualize how good forecasts from the training data set fits the testing set. The testing set represents the original series, which as explained earlier constitutes the last 15 of the observed data.

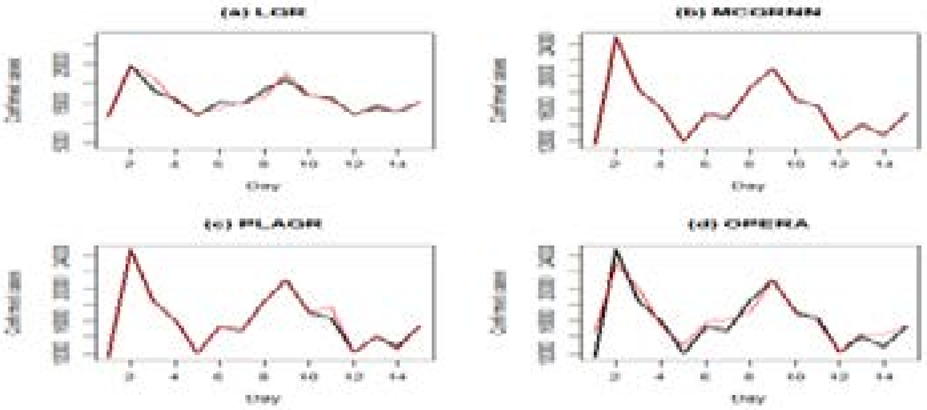

Figure 8: The comparison of forecasts from combined forecast models where the red line represents the predicted values from the training set and the black line represents observed/actual values of the test set

Figure 8 results reveal that the MCQRNN model fits the observed series (test set) well. The plot of the MCQRNN shown in black is closer to the plot of the test set. Although the PLAQR also gives a better fit, it does not provide a good prediction for 10-12 days. Thus, the MCQRNN model outperforms all the other combination models hence, the preferable model.

Out of sample forecasts

We use the developed models for out-of-sample prediction of the confirmed daily cases of COVID-19. Table 6 presents the instances of the predicted cases for the next 14 days, ranging from 26-08-2021 to 08-09-2021.

Results in Table 6 indicate that the number of new confirmed COVID-19 cases fluctuates between 1297 and 23000 for the next 14 days, that is from 26 August 2021 to 8 September 2021. A downward trend in the number of confirmed cases is occurring.

Evaluation of Prediction Intervals

We also assess the sharpness of the predictive distributions by calculating the prediction intervals normalized average width (PINAW) using the methods discussed in Section 2.3.2, i.e., from simple average and median. All the prediction intervals are at the 95% level. The computed PINAWs for the models are 2.5795 and 1.8285 for the simple average and median, respectively. The median has a narrower prediction interval than the average. Figure 9 presents plot of the confirmed cases including forecasted cases for the period 26-08-2021 to 08-09-2021 with the 95% prediction interval. The prediction intervals are from the median combination method for combining prediction limits.

Table 6: Predicted cases for the next 15 days (26-08-2021 to 08-09-2021)

$t+h$	$\hat{y}_{t+h}^{(ARIMA)}$	$\hat{y}_{t+h}^{(TBATS)}$	$\hat{y}_{t+h}^{(SGBM)}$	$\hat{y}_{t+h}^{(GAM)}$	**L.95_med**	$\hat{y}_{t+h}^{(combined)}$	**U.95_med**
536	19600.88	17649.11	24652.49	22166	14218.63	**22194.46**	26630.79
537	17038.81	17263.71	18486.34	15722	13427.25	**16764.91**	21100.17
538	18070.16	15091.17	16263.04	13946	10971	**15600.43**	19211.35
539	12547.15	12715.74	13962.07	12528	8436.59	**10831.03**	16994.88
540	13526.01	11844.08	15577.29	13104	7452.54	**12911.73**	16235.62
541	15958.47	13039.32	16688.39	15052	8495.5	**14541.09**	17583.14
542	15992.99	15326.47	16723.41	18052	10509.49	**13252.8**	20143.45
543	19571.18	16944.35	20527.88	21879	11727.81	**17673.14**	22160.89
544	16607.93	16664.65	17424.34	17328	11034.46	**14654.66**	22294.83
545	18057.55	14690.51	16490.99	16624	8746.95	**14684.7**	20634.07
546	13177.78	12477.85	14171.26	14983	6332.65	**10319.51**	18623.05
547	13531.66	11632.73	14617.19	15061	5334.77	**10981.04**	17930.69
548	15887.62	12719.35	14933.97	16719	6240.52	**11805.66**	19198.17
549	15984.92	14862.81	16581.5	17837	8109.93	**13189.7**	21615.69

Discussion

Improvement of time series forecasting accuracy through combining forecasts from multiple time series candidate

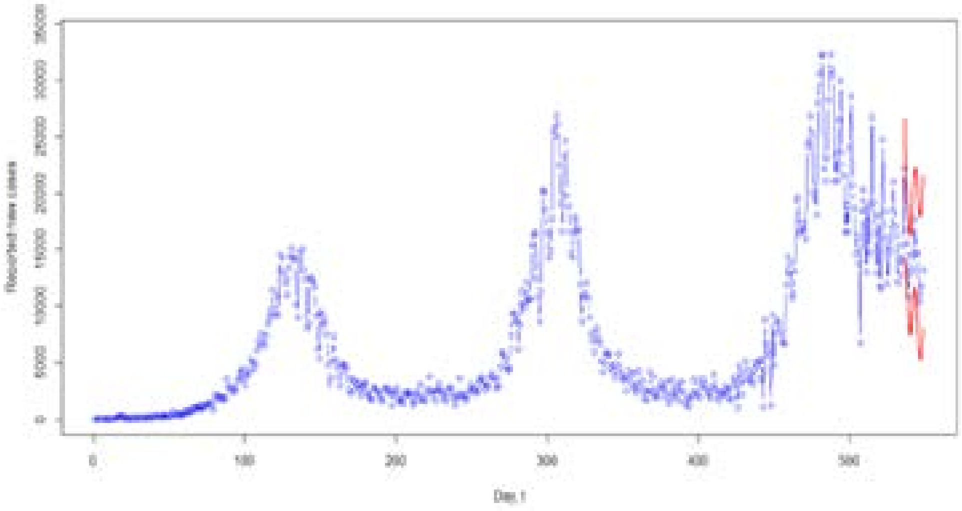

Figure 9: Plot of the confirmed cases including forecasted cases for the period 26-08-2021 to 08-09-2021 with the 95% prediction interval. The prediction intervals are from the median combination method for combining prediction limits

models is an important and dynamic area of research. In this study, we predict the spread of COVID-19 in the SADC region using confirmed daily cases from the 7th of March 2020 to the 25th of August 2021, yielding 535 observations. Since the data set is relatively large, training using the first 520 observations is done, then testing using the last 15 observations followed by a 14-forecast using the candidate models.

The single forecast models used in this study are the ARIMA models, TBATS model, GAM and the SGBM approaches. The GAM outperforms all the other models since has the lowest RMSE, MAPE and Theil's U statistic from these approaches. On testing the model's performance using plots, we discovered that the performance of the GAM was not outright. The GAM could not perform well from 18th August, 2021, to the 20th August, 2021. However, the SGBM and the TBATS models perform better in this interval. Thus, we decided to combine forecasts from the ARIMA, TBATS, GAM, and SGBM to ensure the final model's accuracy.

The forecasts from the single technique models are combined using the Quantile regression approaches, i.e., linear quantile regression averaging (LQR), Monotone Composite Quantile Regression Neural Network model (MCQRNN), PLAQR and the OPERA. The MCQRNN is a novel approach to nonlinear quantile regression modelling that: 1) simultaneously estimates multiple non-crossing, nonlinear conditional quantile functions, 2) allows for optional monotonicity, positivity and generalized additive model constraints, 3) can be adapted to estimate standard least-squares regression and non-crossing expectile regression functions[18].

The combined forecasts models show an increased performance accuracy compared to the performance accuracies for the single forecast models. This is in congruency with findings from studies on combining time series, which purports that combining forecasts from different models effectively reduces the prediction errors and provides considerably increased accuracy[9,14,39]. Cross-validation results suggest that MCQRNN is more robust than all the other models (RMSE=0.00132, MAPE=0.00000614, Theil's U=0.000000278). Its Theil's U statistic is close to zero, indicating a perfect fit. The closer the Theil inequality coefficient is to 0, the smaller the difference between the predicted value and the real value will be, which indicates the better fitting degree of the prediction model[40]. A study on non-crossing nonlinear regression quantiles by MCQRNN on rainfall extremes also confirms the robustness of the MCQRNN approach compared to other baseline models[18]

We developed a quantile regression average model to perform a 14-day out of sample forecast. The model predicted a fairly decreasing trend from 22194 on the 26th of August 2021 to 13189 on the 8th of September 2021, on the number of confirmed cases in the SADC region. We further investigated the sharpness of the fitted models using the PINAWs from simple average and median at 95% level. The median showed a narrower prediction interval. Considering this as the first study conducted using the combined forecast approach to predict the spread of COVID-19, our findings significantly predict the pandemic. The approach allows for the timely forecasting of the spread of COVID-19, hence informing of the introduction of effective interventions in the SADC region.

Conclusion

Forecasting plays an important role in decision making, particularly in this period where the COVID-19 pandemic is challenging the entire world. However, single forecasts techniques do not perform well in predicting the spread of COVID-19 in the SADC region. Combined forecasts models using quantile regression averaging increases accuracy in predicting COVID-19 cases. A prediction of a downward trend for the next 14 days in the COVID-19 cases is shown from the fitted combined forecast model. The findings present an insightful approach in monitoring the spread of COVID-19 in SADC region. The spread of COVID-19 in the SADC region can best be predicted using combined forecasts models, particularly the MCQRNN approach.

Acknowledgment

The authors thank the Universities (Great Zimbabwe University, University of Venda and University of South Africa) for their time in undertaking this work.

Abbreviations

ACF: Autocorrelation function
ADF: Augmented Dickey-Fuller
AIC: Akaike information criteria
ARIMA: Autoregressive integrated moving average
COVID-19: Coronavirus Infectious Disease 2019
CVC: Cross-validation criterion
GAM: Generalised additive models
KPSS: Kwiatkovski-Phillips-Schmidt-Shin
LQRA: Linear quantile regression averaging
MAE: Mean absolute error
MAPE: Mean absolute percentage error
MCQRNN: Monotone composite QRNN
OPERA: Online prediction by ExpeRt aggregation
PACF: Partial ACF
PAQRA: Partial additive quantile regression averaging
PINAW: Prediction interval normalized average width
PIW: Prediction interval width
QRNN: Quantile regression neural network
RMSE: Root means square error
SADC: Southern African Development Community
SARIMA: Seasonal ARIMA
SGBM: Stochastic gradient boosting machine
TBATS: Trigonometric seasonality, Box-Cox transformation, ARIMA errors, Trend and Seasonal components

Reference

1. Center for Health Security. Coronaviruses: SARS, MERS, and 2019-nCoV Updated April 14, 2020. Johns Hopkins. https://www.centerforhealthsecurity.org/resources/fact-sheets/pdfs/coronaviruses.pdf

2. Massinga Loembé, M., Tshangela, A., Salyer, S.J. et al. (2020). COVID-19 in Africa: the spread and response. *Nat Med* 26, 999–1003. https://doi.org/10.1038/s41591-020-0961-x

3. United Nations, Economic Commission for Africa (2020). Socio-Economic Impact of COVID-19 in Southern Africa. COVID-19 Response, 2020. https://www.uneca.org/sites/default/files/COVID-19/Presentations/socio-economic_impact_of_COVID-19_in_southern_africa_-_may_2020.pd

4. Alysha M. De Livera, Rob J. Hyndman & Ralph D. Snyder (2011) Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing, *Journal of the American Statistical Association*, 106:496, 1513-1527, https://doi.org/10.1198/jasa.2011.tm09771

5. Martinez E. Z., Aragan D.C., and Nunes A.A. Long-term forecasting of the COVID-19 epidemic a dangerous idea. *Journal of the Brazilian Society of Tropical Medicine*. Vol53: (e20200481): 2020. https://doi.org/10.1590/0037-8682-0481-2020

6. P Hendikawati, Subanar, Abdurakhman and Tarno. A survey of time series forecasting from stochastic method to soft computing 2020 J. Phys.: *Conf. Ser.* 1613 012019. https://doi.org/10.1088/1742-6596/1613/1/012019

7. Zou, H., Yang, Y. Combining time series models for forecasting. *International Journal of Forecasting* 20 (2004) 69-84. https://doi.org/10.1016/S0169-2070(03)00004-9

8. Bates, J. M.; Granger, C. W. J. The Combining of Forecasts. *Operational Research Quarterly* 1969, n. 20, p. 451-468.

9. Armstrong, J. S. Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers 2001.

10. Mancuso A. C. B. and Werner L. Review of the combining forecasts approaches. *Independent Journal of Management and Production* (IJM&P) 2013; 4(1), 248-277.

DOI:10.14807/ijmp. v4i1.59.

11. Lee, D.H.; Kim, Y.S.; Koh Y.Y.; Song, K.Y.; Chang, I.H. Forecasting COVID-19 Confirmed Cases Using Empirical Data Analysis in Korea. *Healthcare* 2021, 9, 254 https://doi.org/10.3390/healthcare 9030254.

12. Verma P., Khetan M., Dwivedi, S., and Dixit, T. Forecasting the COVID-19 outbreak: An application of ARIMA and Fuzzy time series models. June 2020. DOI:10.21203/rs.3.rs-36585/v1.

13. Gecili E., Ziady A., and Szczesniak R. D. Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modelling through novel applications for the USA and Italy. *PLoS ONE*. 2021; 16.

14. Chan, Y. L., Stock, J. H., and Watson, M. W. A Dynamic Factor Model Framework for Forecast Combination. *Spanish Economic Review* 1999; 1, 91-121.

15. Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, Farah Yasmeen, R Core Team, Ross Ihaka, Daniel Reid, David Shaub, Yuan Tang, Zhenyu Zhou. Forecasting Functions for Time Series and Linear Models. Version 8.15. 1 June 2021. https://CRAN.R-project.org/package=forecast

16. Brandon Greenwell, Bradley Boehmke, Jay Cunningham, GBM Developers. Generalized Boosted Regression Models. *Package "gbm"* Version 2.1.8. July 15, 2020. https://CRAN.R-project.org/package=gbm

17. Trevor Hastie. Generalized Additive Models. *Package 'gam'*. Version 1.20. July 5, 2020. https://CRAN.R-project.org/package=gam

18. Cannon, A.J., 2018. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32(11): 3207-3225. doi:10.1007/s00477-018-1573-6. https://CRAN.R-project.org/package=qrnn

19. Adam Maidman. Partially Linear Additive Quantile Regression. *Package 'plaqr'*. Version 2.0. August 8, 2017 https://CRAN.R-project.org/package=plaqr

20. Pierre Gaillard, Yannig Goude, Laurent Plagne, Thibaut Dubois, Benoit Thieurmel. Online Prediction by Expert Aggregation. *Package 'opera'*. Version 1.2.0. December 6, 2021. https://CRAN.R-project.org/package=opera

21. Adhikari, R., & Agrawal, R. An Introductory Study on Time Series Modeling and Forecasting. *ArXiv* 2013; abs/1302.6613.

22. Chatfield C. The analysis of time series: an introduction. *Chapman and Hall/CRC*; 2016.

23. Wood, S. N. *Generalized Additive Models: An Introduction with R, 2nd Edn*.

24. Boca Raton, FL: CRC Press 2017.

25. Wood, S.N., Goude, Y., and Shaw, S. Generalized additive models for large data sets. *Appl. Statist* 2015; 64; Part 1, pp139-155.

26. Tibshirani R. Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society. Series B (methodology)*, 1996, 58(1), 267-288. https://statweb.stanford.edu/~tibs/ftp/lasso-retro.pdf

27. Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K. and Simon, N. Lasso and Elastic-Net Regularized Generalized Linear Models: glmnet r package version 4.1-2, 2021. https://cran.r-project.org/web/packages/glmnet/glmnet.pdf (Accessed on 10 September 2021).

28. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.

29. Friedman, J.H. Stochastic gradient boosting. Comput. *Stat. Data Anal.* 2002, 38, 367–378.

30. White H (1992) Nonparametric estimation of conditional quantiles using neural networks. In: Page C, LePage R (eds) Computing science and statistics. *Springer*, pp 190–199. https://doi.org/10.1007/978-1-4612-2856-1_25

31. Cannon AJ (2011) Quantile regression neural networks: implementation in R and application to precipitation downscaling. *Comput Geosci* 37(9):1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005

32. Zhang H, Zhang Z (1999) Feedforward networks with monotone constraints, In: IJCNN'99, International joint conference on neural networks, vol 3. IEEE, pp 1820–1823. https://doi.org/10.1109/IJCNN.1999.832655

33. Xu Q, Deng K, Jiang C, Sun F, Huang X (2017) Composite quantile regression neural network with applications. *Expert Syst Appl* 76:129–139. https://doi.org/10.1016/j.eswa.2017.01.054

34. Jiang C, Jiang M, Xu Q, Huang X (2017) Expectile regression neural network model with applications. *Neurocomputing* 247:73–86. https://doi.org/10.1016/j.neucom.2017.03.040

35. Potts WJ (1999) Generalized additive neural networks. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. *ACM*, pp 194–200

36. Sun, X.; Wang, Z.; Hu, J. Prediction interval construction for byproduct gas flow forecasting using optimized twin extreme learning machine. *Math. Probl. Eng.* 2017.

37. Shen, Y.;Wang, X.; Chen, J.Wind power forecasting using multi-objective evolutionary algorithms for wavelet neural network-optimized prediction intervals. *Appl. Sci.* 2018, 8, 185.

38. Mpfumali, P.; Sigauke, C.; Bere, A.; Mulaudzi, S. Day Ahead Hourly Global Horizontal Irradiance Forecasting-Application to South African Data. *Energies* 2019, 12, 3569.

39. Koenker, Roger. "Quantile Regression This article has been prepared for the Statistical Theory and Methods section of the Encyclopedia of Environmetrics edited by Abdel El-Shaarawi and Walter Piegorsch. The research was partially supported by NSF grant SES-0850060". Quantile Regresssion. John Wiley & Sons 2005, Ltd. 10.1002/9780470057339.vnn091. ISBN 9780470057339.

40. Koenker, R., and G. S. Bassett. "Regression Quantiles," Econometrica, 46, 33–50. (1982): "Robust Tests for Heteroscedasticity based on Regression Quantiles," *Econometrica* 1978, 50, 43–61