STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–20.
Published online in International Academic Press (www.IAPress.org)

Estimation of extreme quantiles of confirmed COVID-19 cases using South African data

Claris Shoko [1,*], Caston Sigauke [2]

[1] *Department of Statistics, University of Botswana, plot 4775 Notwane Rd Gaborone Botswana*
[2] *Department of Mathematical and Computational Sciences, University of Venda*
Private Bag X5050, Thohoyandou 0950, Limpopo, South Africa

Abstract **Introduction**: Forecasting is important in any scientific field, including COVID-19 epidemiology. Daily confirmed COVID-19 cases are in different phases characterised by peaks, making it difficult for most mathematical models to handle. In such cases, extreme value theory plays a critical role because values of interest are usually far away from the mean. In this paper, we develop mathematical models using Extreme values to capture uncertainties of forecasts associated with the COVID-19 pandemic using real-time data. **Materials and Methods**: A three-stage approach to probabilistic forecasting is used in this study. The stochastic gradient boosting, generalised additive model, additive quantile regression, and the nonlinear quantile regression are used to predict extremely high quantiles, i.e. 0.95-, 0.99- and 0.995-quantiles. The second stage combines each model's predicted extremely high quantiles using the weighted mean and median methods. The pinball loss and coverage probabilities are used to evaluate the accuracy of the predictions in the third stage. **Results**: For all the extreme quantiles, i.e. the 0.95-, 0.99- and 0.995-quantiles, the cubic spline regression method gives the best predictions based on the lowest pinball losses, which are 171.41, 563.49 and 115.28, respectively. The weighted mean average model dominated by the mean is the second best based on the pinball losses but the best based on the coverage probability. **Conclusion**: This study provides insights into the strengths and weaknesses of different models for short-term extreme quantile prediction of COVID-19. Estimating extreme quantiles of daily COVID-19 using models with high predictive capabilities, such as the weighted mean-median model dominated by the mean, is important to public health officials and policymakers for planning and preparing for potential surges in CoVID-19 cases and similar pandemics in the future.

Keywords COVID-19, Extreme quantiles, Pin ball loss, weighted simple average median model, cubic spline regression model.

AMS 2010 subject classifications ???, ???

DOI: 10.19139/soic-2310-5070-2079

1. Introduction

COVID-19, just like influenza, is a contagious disease that is highly transmissible and spreads around the world with high morbidity and mortality, presenting a huge burden on worldwide public health [1]. The typical observations are usually far from the mean (or not normally distributed) [2]. This is caused by extremely large increases (or peaks) in daily COVID-19 cases and has become a challenge as most mathematical models cannot be used on non-normal data and with heavy tails. In epidemiology, forecasting of time series plays an important role. Researchers' main concern is to develop the most accurate forecasting techniques. Accurate assessment of the level and frequency of future extreme day-to-day increases in COVID-19 cases are subject to a wide range of uncertainties, including deaths, shortages of health equipment, and new infections, to mention a few. The

[*] Correspondence to: Claris Shoko (Email: shokoc@ub.ac.bw

ISSN 2310-5070 (online) ISSN 2311-004X (print)

application of Extreme Value Theory in forecasting the uncertainties associated with super-spreading COVID-19 is of paramount importance. However, the application of Extreme Value Theory in the study of COVID-19 forecasting remains relatively unexplored.

Extreme Value Theory (EVT) has shown promising results in estimating the probability of extreme events for a relatively short period of observations, and its history dates back to the early work by Fisher and Tippert (1928) [3]. EVT has been successfully applied in many multidisciplinary areas, including influenza [1] and value-at-risk application in finance [4]. This study develops Extreme Value models using different approaches to forecast the probability of the outbreak of the highly pathogenic COVID-19.

1.1. Literature review

Several studies have been conducted to forecast the incidence of new COVID-19 cases daily. Sciannameo et al., [5] used a deep learning approach for the spatial-temporal forecast of new cases and hospitalisation. Aljaaf et al. [6] used the Monte-Carlo sampling methods and Bayesian inference to reformat neural network parameters to estimate the uncertainty associated with COVID-19 data to forecast daily deaths, daily infections and daily recovered using COVID-19 data for Iraq. Safari et al. [7] also forecasted the incidence of new cases, recovery cases, and mortality rates, but they used a novel deep interval type-2 LSTM model.

Probabilistic forecasts provide quantitative uncertainty information associated with any pandemic. Probabilistic forecasts, in the form of intervals, densities, or quantiles, can provide more comprehensive information about uncertainties of the future COVID-19 pandemic. One area of interest is accurately predicting extremely high quantiles of confirmed daily COVID-19 cases. Several methods of estimating extreme quantiles are discussed in the literature.

Sun et al. [8] developed a two-step probabilistic wind forecast method based on pinball loss optimisation. In the first stage, they generated deterministic forecasts using a machine learning-based multi-model forecasting framework. In the second stage, a set of unknown parameters in the predictive distribution were optimised by minimising the pinball loss. A shortcoming of this study is that it only used one probabilistic evaluation metric, the pinball loss function.

Wang et al. [9] used a probabilistic combination approach to load forecasting. Their study contributes to the probabilistic load forecasting literature by proposing a constrained quantile regression averaging (CQRA) method for quantile forecast combinations. They formulated the proposed CQRA method as a linear programming (LP) problem for parameter estimation that minimises the pinball loss. The solution includes the optimal weights for the individual probabilistic forecasts. Similar to the previous study, this was also done in two stages. The developed models were used to produce forecasts combined in the second stage. Wang et al. [9] further argued that no individual forecasting method is the best for all data sets. Thus, combining forecasts reduces the risk of making a poor model selection. Swaraj et al. [10] also support this; they combined residuals from the autoregressive integrated moving average model with the artificial neural network model to model short-term forecasting of COVID-19 data. In this study, forecasts from competing models are also combined.

Application of the quantile regression based models in estimating extremely high quantiles is discussed in literature. Maswanganyi et al. [11] used additive quantile regression (AQR), extreme mixture (EM), and nonlinear quantile regression (NLQR) models to estimate high and low quantiles of electricity demand. Their findings showed that the AQR model produces the most accurate prediction at high and low quantile levels.

There are a few articles in the literature on estimating the extreme quantiles of daily confirmed cases of COVID-19. One notable contribution is that of Enriquez et al. [12], who used the generalised extreme value distribution type 1-Gumbel and Exponential (1, 2 parameters) to analyse the probability of new daily confirmed cases. Furthermore, the data set included information on events in terms of how they occur within a specific time frame measured in days. Another important contribution is the work of Liu and Zheng [13], who provided a systematic statistical analysis of the features of the empirical geographical distribution for confirmed COVID-19 cases and deaths at cumulative and daily counts, county, city, and state levels from January 2020 to June 2022. In a related study Daouia et al. [2] suggested that the major driver of overall transmission in the case of SARS-CoV-2 is superspreading. Hence, statistically investigating extreme tail events is important to better understand virus propagation and control. They used discrete Generalised Pareto models. This is also supported by Wong and Collins [14], who provided evidence that coronavirus superspreading is fat-tailed; thus, extreme value theory offers a framework for modelling superspreaders. Hence, this study considers the estimation of extremely high quantiles.

This study uses a two-stage probabilistic COVID-19 forecasting approach. In the first stage, we use the following models: cubic regression spline, generalised additive models, stochastic gradient boosting and additive quantile regression. Forecasts from these models are combined by considering their average and median point forecasts. The weighted average-median model is then formulated by combining the average and median models based on weights. A mean-dominated model is formulated by assigning more weight to the average forecasts and less weight to the median forecasts. The median-dominated model is formulated by assigning more weight to the median forecasts and less weight to the average model. In the second stage, the pinball loss function determines the best model. An improved and current version of the pinball loss is used in the study. This study focuses on extremely high quantiles. Thus, to make recommendations to the health policy and decision-makers, the maximum number of COVID-19 cases is an issue. Hence, instead of using extremely high and low quantiles as in [8] and [11], we only consider extremely high quantiles (0.95-, 0.99- and 0.995-quantiles). The largest possible number of cases of life-threatening diseases like COVID-19 is an important issue in epidemiology.

1.2. Contribution and Research highlights

To our knowledge, the literature does not discuss the estimation of extremely high quantiles of confirmed COVID-19 cases. Unlike previous literature, we use the percentiles method, including the nonlinear and additive quantile regression models, and the stochastic gradient boosting method in predicting extreme COVID-19 cases using South Africa. Based on the literature review discussed in Section 1.1, the highlights and key findings of this study are

- The percentiles method outperformed the other models and had the highest predictive accuracy.
- On combining forecasts, the mean approach gave the best results and was the second best at all quantile levels.
- The study provided a robust modelling framework for estimating high quantiles of daily COVID-19 cases.

The rest of the paper is organised as follows: Section 2 presents a discussion of the methods used in the study, followed by the empirical results in Section 3. A detailed discussion of the results is given in Section 4 and Section 5 concludes.

2. Methods

2.1. Data and regional setting

In this study, we used an openly available daily number of confirmed cases of COVID-19 data set reported by “Our World in Data” and is available on (`https://github.com/csigauke/Data-for-article-Estimation-of-extreme-quantiles`). The data is from 7 March 2020 to 27 September 2021. Since the beginning of the pandemic in South Africa, 17 588 025 tests have been conducted, with an average of 31 514 tests per day. COVID-19 vaccination started on the 18th of February 2021, and by the 27th of September 2021, an average of 20 644 vaccinations per day was administered.

Modelling and predicting the spread of COVID-19 in South Africa are done using the R packages: ‘gbm’ developed by Greenwell et al. [15] for fitting the stochastic gradient boosting model, ‘quantreg’ developed by Koenker et al. [16] and used for fitting the nonlinear quantile regression models, ‘qgam’ by Fasiolo et al. [17] used for fitting additive quantile regression models and ‘gefcom2017’ developed by Roach [18] for calculating the pinball losses.

2.2. Extreme Value Theory: Quantile regression approach

Quantile regression is an appealing method for analysing high dimensional data because of its ability to model heteroskedasticity relationships correctly; it is robust to outliers in the response, sparsity levels change with quantiles, and can provide a thorough analysis of the conditional distribution of the response variable [19]. From

7 March 2020 to 27 September 2021, the historical time series data consist of 569 data points. These data points are split into two sets: the training set and the test set. We fit the single forecast models for each training set using the cubic regression spline(referred to in this paper as the percentile method) model, the stochastic gradient boosting method, the generalised additive quantile regression model, and the nonlinear quantile regression model. Forecasts from the models are combined using the simple mean and median approaches. We then used the pinball loss function and the coverage probability as the probabilistic evaluation metrics based on the extreme 0.95-, 0.99- and 0.995-quantiles. Figure 1 shows the conceptual framework for this study. A detailed description of each of the methods then follows.

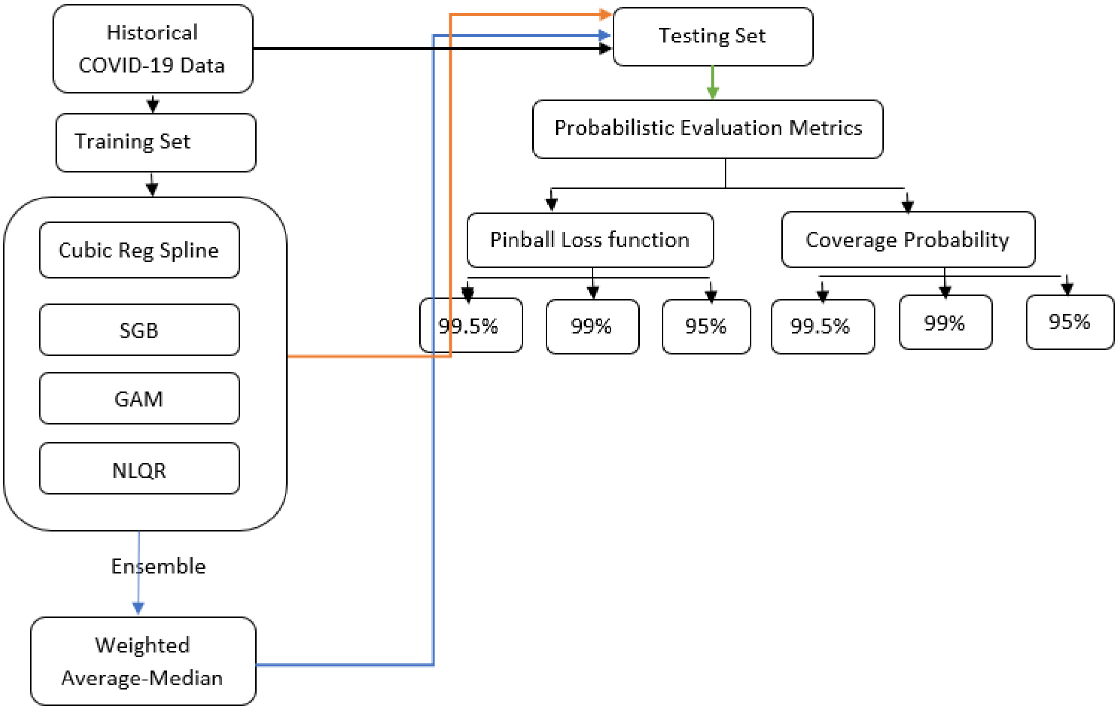

Figure 1. Modelling framework.

2.2.1. *Cubic regression splines model* We develop a parsimonious model based on cubic regression splines with a positive shift factor for estimating high quantiles of confirmed daily COVID-19 cases. The model is given in equation (1).

$$\pi_1(t) = \sum_{t=1}^{n} \left(y_t - f(x_t)\right)^2 + \lambda \int \left(f^{''}(x)\right)^2 dx + \varepsilon_t, \tag{1}$$

where y_t denotes our NCSA on day t, λ is a smoothing parameter and ε_t is the error term. We then extract residuals as shown in equation (2).

$$\varepsilon_t = \pi_1(t) - (\sum_{t=1}^{n} \left(y_t - f(x_t)\right)^2 + \lambda \int \left(f^{''}(x)\right)^2 dx) \tag{2}$$

Let $\hat{\pi}_1(t) = (\sum_{t=1}^{n} \left(y_t - f(x_t)\right)^2 + \lambda \int \left(f^{''}(x)\right)^2 dx)$. To estimate extreme quantiles of NCSA, i.e. 0.95-,0.99- and 0.995-quantiles, we use the distribution of the residuals to estimate the residual quantiles, $\tau = \{0.95, 0.99, 0.995\}$. Then, extreme quantiles of NCSA will be estimated using equation (3).

$$\pi_2(t) = \hat{\pi}_1(t) + q(\tau) \tag{3}$$

That is

$$q(0.95) = \hat{\pi}_1(t) + \tau(0.95),$$

where $\tau(0.95)$ is the 95^{th} percentile of the error distribution. Similarly

$$q(0.99) = \hat{\pi}_1(t) + \tau(0.99)$$

$$q(0.995) = \hat{\pi}_1(t) + \tau(0.995)$$

2.2.2. Additive quantile regression The Additive modelling approach helps reduce dimension in a multivariate nonparametric mean or quantile regression. This is done by ranking variables in terms of their relative influence. Gradient boosting (GB) is a machine-learning technique that fits an additive model stage-wise ([20]). The additive model can take the form given in equation (4) ([21]).

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m), \tag{4}$$

where $b(x; \gamma_m) \in \mathbb{R}$ 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 ([21]). Stochastic gradient boosting (SGB) is a slight modification of GB in which a subsample of the training data set is randomly drawn without replacement from the full training set. The base function is then fitted from the subsample. A detailed algorithm for the SGB model is given in ([22]).

The additive quantile regression (AQR) model is formed by combining the generalised additive model (GAM) and the quantile regression model (QRM); hence it is called a hybrid model. Gaillard et al. [23] were the first to use the AQR models in short-term load forecasting. The model was further extended by Fasiolo et al. [24]. Let y_t denote hourly electricity demand where $t = 1, \ldots, n$, n is the number of observations and let the number of days be denoted by n_d. Then, $n = 24n_d$, where 24 is the number of hours in a day and the corresponding p covariates, $x_{t1}, x_{t2}, \ldots, x_{tp}$. The AQR model is given in equation (5).

$$y_{t,\tau} = \sum_{j=1}^{p} s_{j,\tau}(x_{tj}) + \varepsilon_{t,\tau}; \quad \tau \in (0,1), \tag{5}$$

where $s_{j,\tau}$ are smooth functions and $\varepsilon_{t,\tau}$ is the error term. The smooth function, s, is given in equation (6)

$$s_j(x) = \sum_{k=1}^{q} \beta_{kj} b_{kj}(x_{tj}), \tag{6}$$

where β_j denotes the j^{th} parameter, and $b_j(x)$ represents the j^{th} basis function with the dimension of the basis being denoted by q. The parameter estimates of Equation (5) are obtained by minimising the function given in equation (7).

$$q_{Y|X}(\tau) = \sum_{t=1}^{n} \rho_\tau \left(y_{t,\tau} - \sum_{j=1}^{p} s_{j,\tau}(x_{tj}) \right), \tag{7}$$

where ρ_τ is the pinball loss function. The AQR models are given in equation (8).

$$y_{t,\tau} = \sum_{j=1}^{p} s_{j,\tau}(x_{tj}) + \sum_{k=1}^{K} \sum_{j=1}^{j} \alpha_{jk} s_j(x_{tj}) s_k(x_{tk}) + \varepsilon_{t,\tau}, \tag{8}$$

In this study, we considered three quantile values, which were 0.95, 0.99 and 0.995, respectively.

Table 1. Model comparisons.

Models	**Strengths**	**Weaknesses**
M1 (CRS)	1. Can model complex, non-linear relationships in data. 2. Provides a smooth estimate of the quantile function. 3. The spline coefficients can often be interpreted in the context of the data.	1. Can overfit, especially with a large number of knots. 2. More computationally expensive than linear methods. 3. The choice of knots and degree of the spline requires careful tuning, which can be challenging.
M2 (SGBM)	1. Generally provides excellent predictive performance, especially for complex datasets. 2. Effectively captures complex and non-linear relationships in the data. 3. Tends to be robust to overfitting, especially with large datasets.	1. Models can be difficult to interpret due to their complexity. 2. Training can be time-consuming and resource-intensive. 3. Model performance is highly sensitive to hyperparameters, requiring extensive tuning.
M3 (AQR)	1. Provides clear, interpretable models as it breaks down the relationship into additive components. 2. Can model non-linear relationships through the use of non-linear functions. 3. Directly estimates quantiles, which is useful for extreme quantile estimation.	1. Can become complex when including many interactions or non-linear terms 2. There is a risk of overfitting with a large number of terms or complex basis functions. 3. Computationally more intensive than simple linear models.
M4 (NLQR)	1. Can capture nonlinear relationships between predictors and the response variable. 2. Directly models the quantiles of interest, which is crucial for extreme quantile estimation. 3. Allows for flexible modelling of quantiles without assuming a linear relationship.	1. Setting up and interpreting nonlinear quantile regression models can be complex. 2. More computationally intensive than linear quantile regression. 3. Prone to overfitting, particularly with small datasets or complex models.

2.2.3. Nonlinear quantile regression Nonlinear quantile regression is a powerful tool for understanding the relationship between variables across the entire response variable distribution. Its flexibility in modelling nonlinear relationships makes it applicable to a wide range of fields where understanding the tails or specific quantiles of the distribution is crucial.

In contrast to the linear quantile regression model, a nonlinear quantile regression (NLQR) model is one in which the model is nonlinear in its parameters ([25]). The NLQR model (nonlinear in parameters) is given in equation (9) and discussed in detail in Koenker [25]:

$$q_{Y|X}(\tau) = g(X, \beta(\tau)), \tag{9}$$

where Y denotes the response variable, X represents the predictor variables and they are modelled using a nonlinear function $g(X, \beta(\tau)$ (where $0 < \tau < 1$), $\beta(\tau)$ represents the parameters of the model at a given quantile τ and $q_{Y|X}(\tau)$ is the conditional $\tau^{\text{th}}-$quantile of Y given X.

The parameters $\beta(\tau)$ are estimated by minimising an objective function based on the quantile loss function. The estimator of the parameters is given equation (10).

$$\hat{\beta}(\tau) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - g(x_i, \beta) \right), \tag{10}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ and I is the indicator function, $g(x_i, b)$ is a function with unknown parameters, and $\hat{\beta}(\tau)$ is the unknown regression coefficient for the τ^{th} quantile.

2.2.4. Model comparisons based on strengths and weaknesses A comparative summary o the models in terms of their strengths and weaknesses is given in Table 1.

2.3. Modelling Framework: Forecast combinations

Single forecast models may not be competent enough to capture all the patterns in the dataset. Combining forecasts from different models with different variability improves the models' performance.

Let $Y = \left[y_1, y_2, ..., y_n\right]^T$ be the actual testing dataset of a time series forecasted from k different models. For each i^{th} model forecasts of Y are given by:

$$f^{(i)} = \left[\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, ..., \hat{y}_n^{(i)}\right]^T, i = 1, 2, ..., k. \tag{11}$$

In particular for $i = (CRS, SGB, AQR, NLQR)$, we have a set of forecasts $(f^{CRS}, f^{SGB}, f^{AQR}, f^{NLQR})$, for each of the fitted models. In epidemiological studies, extreme quantiles are interesting because they are associated with disease outbreaks. Thus, we considered extreme quantiles for each fitted model, $\tau = 0.95, 0.99, 0.995$. For each model, we have the following:

$$f_\tau^{(i)} = \left[\hat{y}_{1(\tau)}^{(i)}, \hat{y}_{2(\tau)}^{(i)}, ..., \hat{y}_{n(\tau)}^{(i)}\right]^T, i = 1, 2, ..., 4; \tau = 0.95, 0.99, 0.995. \tag{12}$$

For each extreme quantile, $\tau = 0.95, 0.99, 0.995$, forecasts from the $k = 4$ models are combined using the simple average (Av) and median (Md). The weighted average-median model is developed by assigning weights to the average (Av) and the median (Md) and then combining the two. The selection of weights is based on the robustness of Av and Md.

The weighted average-median model This method takes the simple arithmetic means of the prediction limits given in equation (13).

$$f_\tau^{Av} = \frac{1}{k}\sum_{i=1}^{k} f_\tau^k \tag{13}$$

This approach is fairly simple and produces robust intervals ([26]). Compared to complicated forecasting techniques, the average method produces better results. The average method has its shortcomings when the data contains outliers.

The median method is also fairly easy to use and is not sensitive to outliers. It is given is given in equation (14).

$$f_\tau^{Md} = \text{median}\left(f_\tau^{CRS}, f_\tau^{SGB}, f_\tau^{AQR}, f_\tau^{NLQR}\right) \tag{14}$$

We identify the model with the smallest KPIs from the simple mean and median models: RMSE, MAE, and pinball losses. Weights are assigned to the fitted models based on equation (15).

$$F_\tau^i = \alpha f_\tau^{Md} + (1-\alpha) f_\tau^{Av} \tag{15}$$

The value of α represents the weight assigned to each fitted model, f_τ^{Av} are the forecasts from the mean or average model, and f_τ^{Md} are forecasted from the median model. If $0 < \alpha < 0.5$, then the model is mean-dominated (MeanDom) and if $0.5 < \alpha < 1$ is median-dominated (MedDom).

2.4. Probabilistic evaluation metrics

Probabilistic forecasting plays a vital role in forecasting complex systems like the spread of COVID-19, where uncertainty is always irreducible. With the spread of COVID-19, probabilistic forecasts are essential to produce robust decisions against uncertain future conditions. Uncertainties about the spread of COVID-19 are a threat to human life. This study uses two probabilistic evaluation metrics: the pinball loss function and the coverage probability.

2.4.1. Pinball loss function Quantile regression is traditionally based on the pinball loss. The pinball loss is one of the most popular metrics for evaluating the performance of probabilistic forecasting. The pinball loss (PL) function is relatively easy to use and is given in equation (16)

$$PL(q_{\tau,t}) = \begin{cases} 2(1-\tau)|y_t - q_{\tau,t}|, & \text{if } y_t < q_{\tau,t}, \\ 2\tau|y_t - q_{\tau,t}|, & \text{if } y_t \geq q_{\tau,t}, \end{cases} \tag{16}$$

where $q_{\tau,t}$ is the quantile forecast at time t and y_t is the observed value of the confirmed cases t. The interpretation of $PL(q_{\tau,t})$ is made easier by the inclusion of the multiplier number 2 ([27]). When $\tau = 0.5, PL_{0.5,t} = |y_t - q_{\tau,t}|$, which is the same as the absolute error. Hence, $PL(_{\tau,t})$ is generally interpreted as an absolute error.

2.4.2. Coverage probability Coverage probability is the probability that a procedure for constructing a region will give an interval that covers the true population parameter. Coverage probability is a way to evaluate the performance of a confidence interval estimator; ideally, the confidence interval should have the highest possible coverage probability. The coverage probability (CP) is computed as follows: we count the number of observations below the $\tau-$quantile for $\tau = 0.95, 0.99, 0.995$ and divide it by the total number of observations, n. The formula is given in equation (17).

$$\text{CP} = \frac{n_\tau}{n}, \tag{17}$$

where n_τ is the number of observations below the $\tau-$quantile.

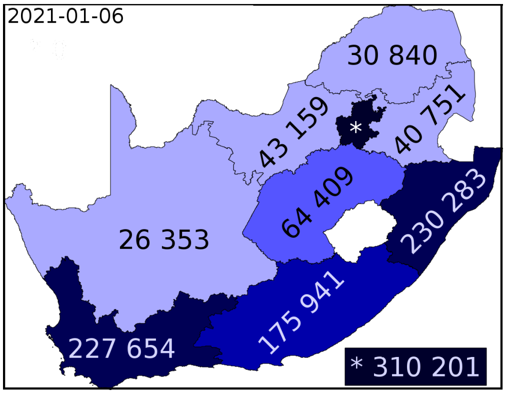

Figure 2. CoViD-19 pandemic cases in South Africa on 2021-01-06. Source: https://en.wikipedia.org/wiki/File:CoViD-19_pandemic_cases_in_South_Africa.svg

3. Empirical Results

In this section, we fit the percentile model (Model 1), the Stochastic Gradient boosting method (Model 2), the Quantile Regression model (Model 3) and the Nonlinear Quantile Regression model (Model 4)on daily COVID-19 cases in South Africa. Training and testing of the data set are done by taking the training set up to the 400th observation, and the testing set is from the 401st observation to the 558th observation. After testing the four models, a point forecast combination is done using the mean and median of the fitted models. Probabilistic forecasting for the six models uses the Pinball loss and coverage probability.

3.1. Exploratory Data Analysis

The distribution of COVID-19 in South Africa for each province by the 6th of January 2021 is given in Figure 2. Mpumalanga has the highest total number of cases.

Table 2 shows summary statistics of the daily confirmed COVID-19 cases. We carried out a formal test to check whether the data was stationary or not using the Kwiatkowski- Phillips-Schmidt-Shin (KPSS) test. The results of the test showed that the data was not stationary. We then went on to difference the data before computing the skewness and the kurtosis. The mean and median values from Table 2 are far apart, showing that the data is not approximately distributed. This is supported by the skewness value, which suggests that the data is skewed to the right. The kurtosis is very high at 6.9398, suggesting that the data distribution is leptokurtic. So, appropriate distributions that best fit the data are fat or heavy-tailed distributions.

Table 2. Summary statistics of daily COVID-19 cases in South Africa

Variable	Statistics
Minimum	0
Q1	1287
Median (Q2)	2574
Mean	5192
Q3	8387
Maximum	26485
Skewness	0.9084
Kurtosis	6.9298

The trend and distribution of daily COVID-19 cases in South Africa are plotted in Figure 3.

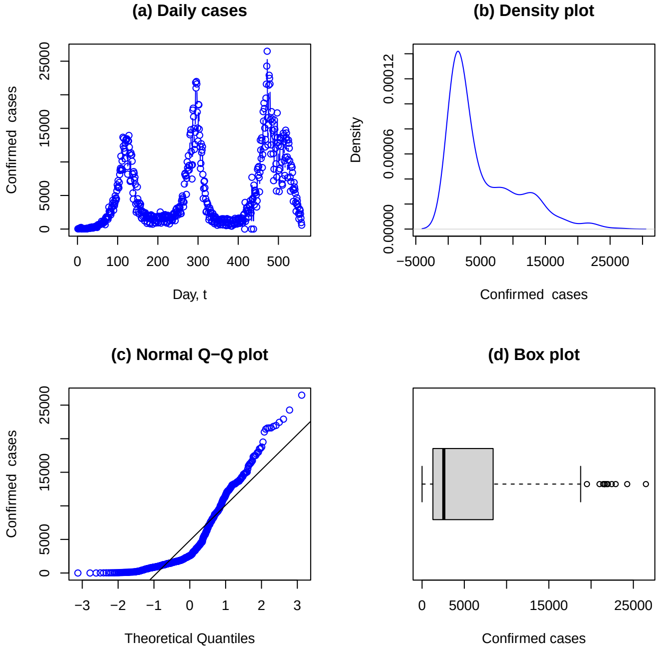

Figure 3. Reported daily COVID-19 cases: Time series plot (a), density plot (b), Normal Q-Q plot (c), and Box plot (d).

Figure 3(a) shows that the number of daily reported COVID-19 cases follows a nonlinear trend characterised by peaks. The density plot of the daily cases has a long right tail. This indicates the existence of a heavy tail in the data, which is a characteristic of super-spreading events. The Box-plot in Figure 2d also confirms the heavy tail in the distribution of daily COVID-19 cases [2]. This has led to modelling daily COVID-19 cases using Extreme value analysis based on extremely high quantiles, that is $95\%, 99\%, 99.5\%$ quantiles for different models.

Figure 4 shows box plots of the monthly distribution of confirmed COVID-19 daily cases from 2020 to 2021. It can be seen that the incidences of COVID-19 are high during the winter and summer seasons and its values are low during the autumn and spring seasons.

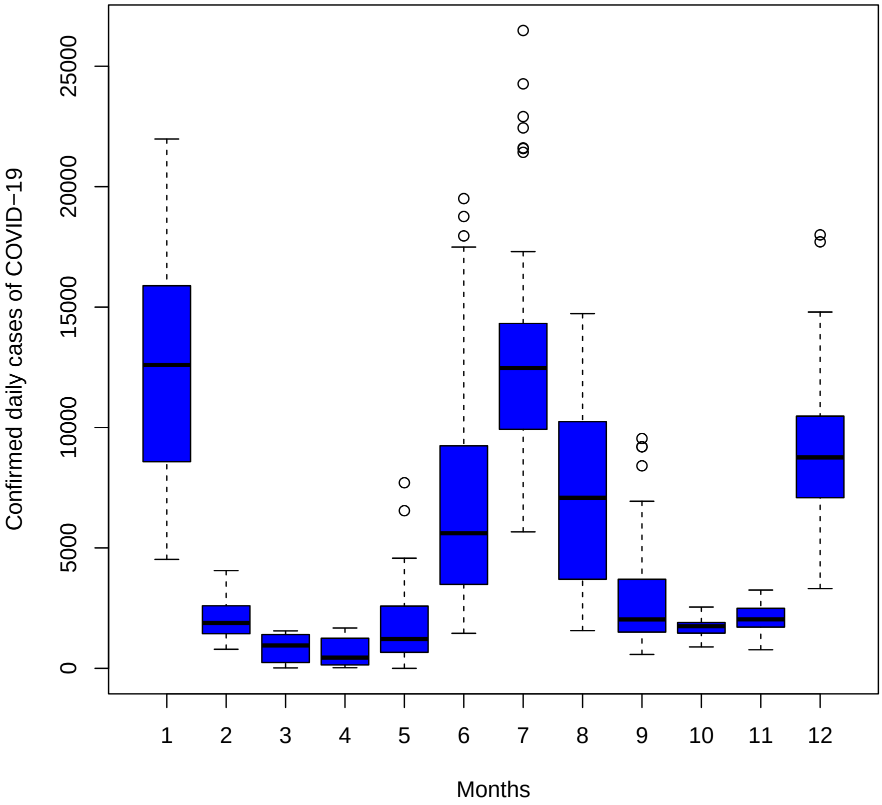

Figure 4. Distribution of monthly confirmed COVID-19 cases for the sampling period 19 March 2020 to 27 September 2021.

3.2. *First stage of analysis*

3.2.1. Cubic regression spline model In fitting the cubic regression spline model, two variables are used: the number of new COVID-19 cases (response) and the time index (independent variable). Residuals from the fitted model are then extracted, and their density plot is given in Figure 5.

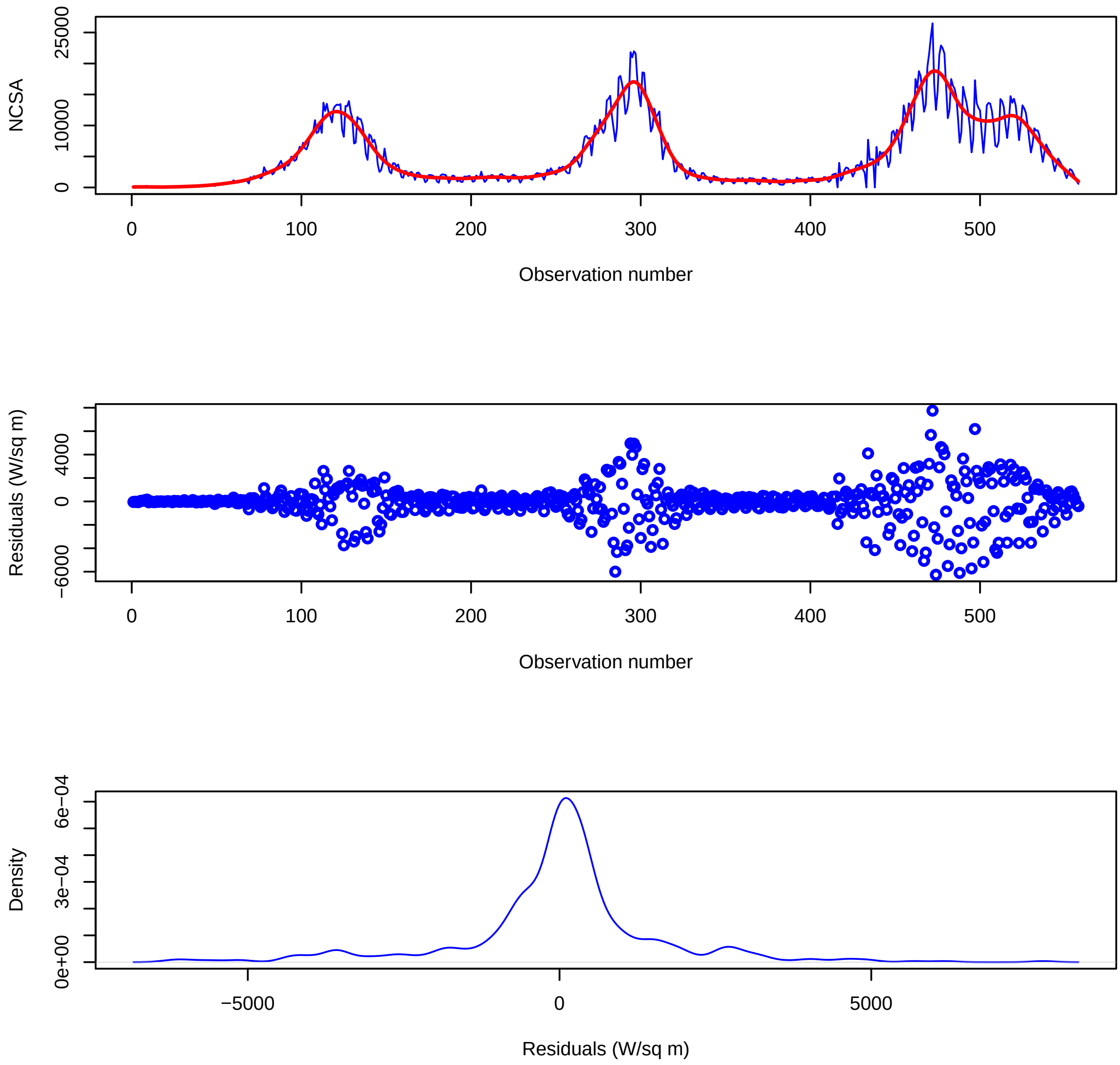

Figure 5. Model M1 cubic regression splines with residuals and density plot for residuals.

Figure 5 (top panel) shows the fitted cubic regression splines model. Like median regression, the model divides the observations in half, that is, 50% of the observations up and the other 50% below the fitted model. The smoothing parameter used for fitting the model is $\lambda = 0.4514152$. Figure 5 (middle panel) presents the pot of residuals. The residuals are scattered around periods where there are peak COVID-19 cases. Figure 5 (bottom panel) presents the density plot for the residuals. From the residuals, the extreme quantiles, that is, 0.99-, 0.95- and 0.995-quantiles, are estimated and added to the residuals as shifting parameters. The 0.95-quantile is the residual value of 2700, the 99-quantile is the residual value of 4626, and the 99.5% is the residual value of 4953. The fitted cubic regression splines model is shifted using each quantile residual value for the same smoothing parameter $\lambda = 0.4514152$. The shifted models are used for training and testing sets. The predicted values versus the whole data are presented in Figure 6. Predictions for the test set are extracted, and their performance on the test set is presented in Figure 7, respectively.

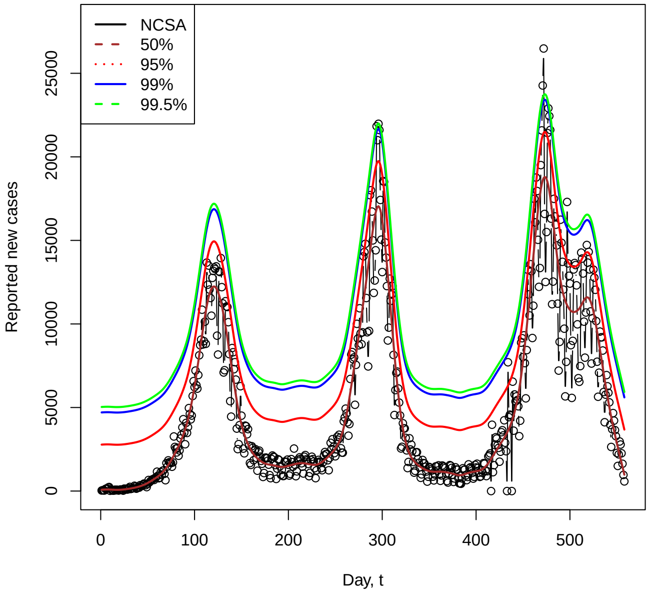

Figure 6. Model M1 cubic regression splines.

Figure 6 presents the cubic regression splines model and the extreme quantiles for the training set.

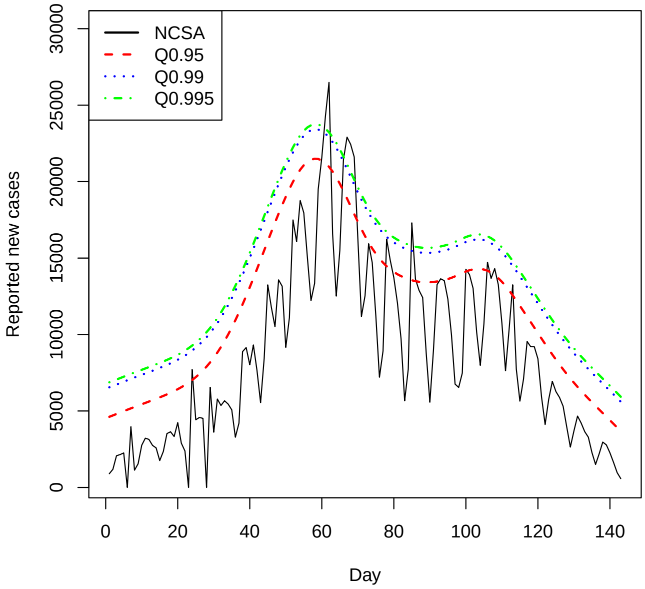

Figure 7. Model M1 cubic regression splines.

Figure 7 presents the cubic regression splines model and the extreme quantiles for the test set.

3.2.2. Stochastic gradient boosting model SGBM is used to select variables that have a relative influence on the spread of COVID-19. The fitted values from the cubic regression splines model indicated as nlifit in Table 3 are extracted and included as an additional covariate. The relative influence of the variables is ranked in Table 3.

Table 3. Relative influence of independent variables on new COVID-19 cases.

var	rel.inf
nlifit	68.2%
NTSA	19.0%
PRSA	9.30%
lag2	1.08%
new tests per thousand	1.07%
month	0.369%
lag1	0.329%
TTSA	0.315%
day	0.190%
RRSA	0.105%
trim	0.006.61%
quard	0.000924%
total tests per thousand	0%
PFVSA	0%
NVSA	0%

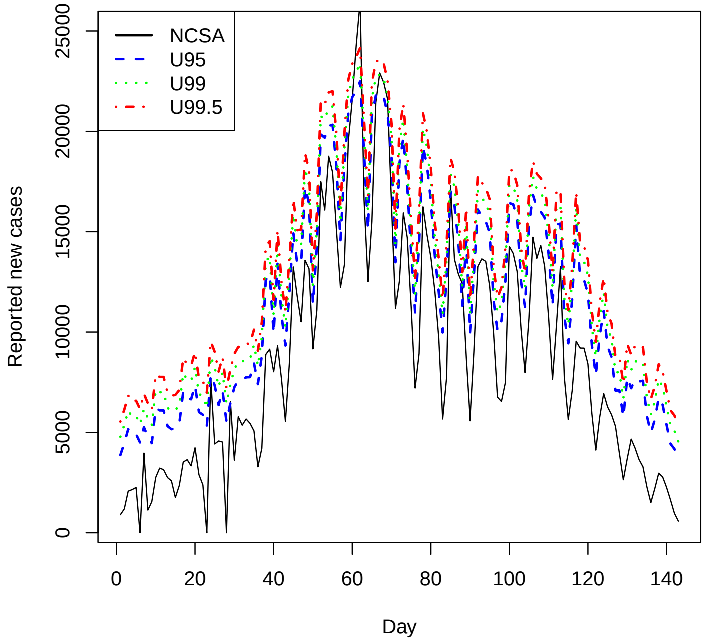

Figure 8. Model M2 SGBM.

Results from Table 3 show that estimated values from the nonlinear cubic regression splines fit (nlifit) significantly influence the number of new daily COVID-19 cases. The second influential variable is the number of new tests (NTSA) with a percentage influence of 19%. The other influential variables are lag 2, new tests per thousand, month, lag 1, total tests (TTSA), day, and reproductive rate (RRSA). However, the variable total tests per thousand, PFVSA, and NVSA have zero influence on the number of new daily COVID-19 cases.

The SGBM model is now used to train and test the data set. The SGBM is also fitted for each extreme quantile (0.99-, 0.95- and 0.995-quantiles). This is done by calculating the standard error =1477.618 of the testing set. For the 0.95-quantile, the fitted SGBM is shifted by $1.96 \times se$, 0.99-quantile. The SGBM is shifted by $2.576 \times se$, 0.995-quantile the SGBM is shifted by $3.090 \times se$. The fitted models are presented in Figure 8.

Results from the forecasts using the four models presented in Figure 8 show that the SGBM has a better memory than the cubic regression splines model. The model managed to memorise the behaviour of the test data.

3.2.3. Additive quantile regression Variables with non-zero relative influence selected using the SGBM are used to fit the average quantile regression (AQR) model. Three AQR models were fitted using the extreme quantiles (0.95-, 0.99- and 0.995-quantiles). For the AQR model with 0.995-quantiles, the deviance explained is 100%, the 0.95 quantiles the deviance explained is 99.3%, and for the 0.99-quantiles, the deviance explained is 99.9%. In all the fitted AQR models, all variables were highly significant except for the new tests (NTSA) and lag 1. The predicted daily COVID-19 cases for each extreme quantile are estimated. The plots are presented in Figure 9.

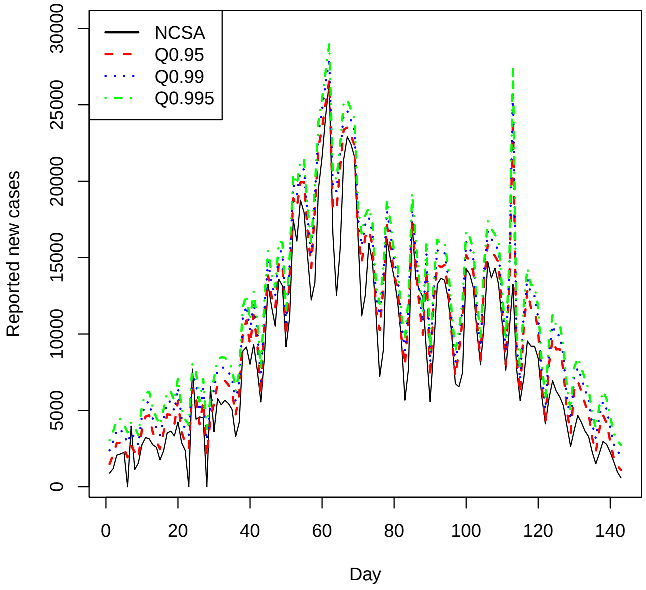

Figure 9. Model M3 AQR.

Figure 9 shows that the predictions from all the quantiles generally behave better than the other cubic regression splines model and the SGBM. The AQR has proved to have a better memory except on days 124 and 125, where the AQR for the 0.95 quantiles misbehaved.

Table 4. Comparative evaluation of models. Above UL = number of forecasts above the upper prediction limit.

PINC	Model	Pinball Loss (%)	MAE
95%	M1(CRS)	**563.49**	-
	M2(SGBM)	1541.29	1099.04
	M3(AQR)	2397.02	2020.81
	M4(NLQR)	2429.58	2011.70
	Median	1569.84	1131.16
	Mean	1521.67	1096.27
	MedDom	1553.523	1118.73
	MeanDom	1432.185	1004.77
99%	M1(CRS)	**171.41**	394.19
	M2(SGBM)	999.53	923.39
	M3(AQR)	1744.03	1624.68
	M4(NLQR)	1633.95	1523.97
	Median	1051.98	988.50
	Mean	905.74	833.77
	MedDom	1006.74	943.12
	MeanDom	847.0783	781.98
99.5%	M1(CRS)	**115.28**	449.04
	M2(SGBM)	738.69	754.03
	M3(AQR)	1355.03	1315.00
	M4(NLQR)	1382.63	1362.90
	Median	811.78	844.24
	Mean	691.69	715.72
	MedDom	775.165	806.27
	Mean1Dom	626.3497	655.64

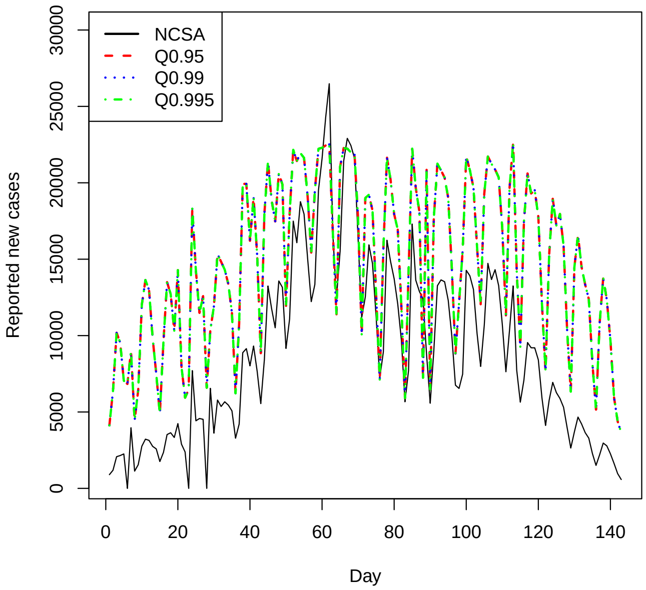

Figure 10. Model M4 NLQR.

3.2.4. Nonlinear quantile regression In fitting the nonlinear quantile regression (NLQR) model, the same covariates as in the AQR were used: estimates from the cubic regression splines models, productivity rate, new COVID-19 tests, lag1, lag 2, new tests per thousand, and month. Three NLQR models are used with extreme quantiles 99%, 95%, and 99.5%). The plots of the predictions from the three NLQR models are presented in Figure 10. A plot of the testing set is included for comparison.

The NLQR models for the extreme quantiles show fairly good behaviour, although not as good as the AQR models.

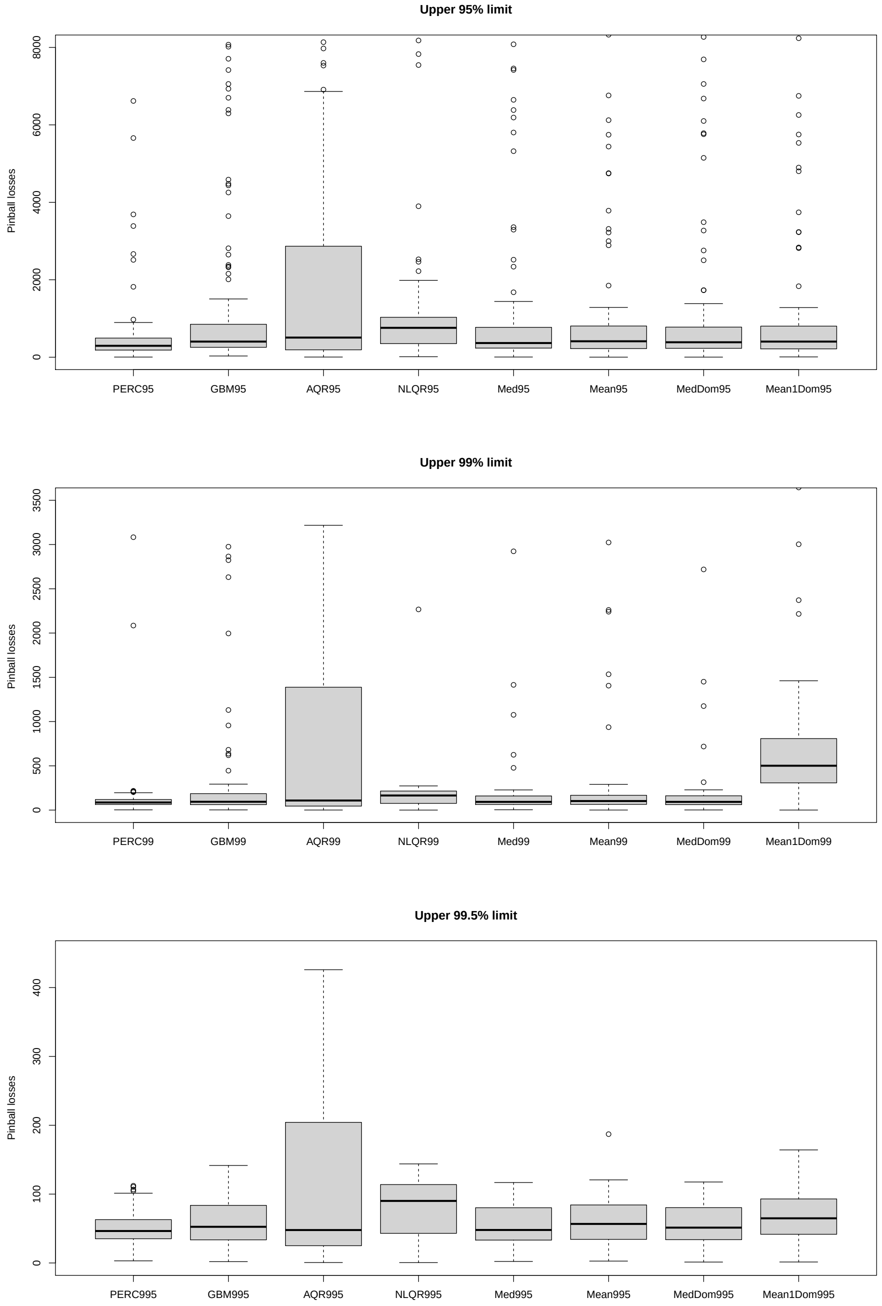

Figure 11. Box plots of pinball losses at 95%, 99% and 99.5%.

3.3. *Forecast evaluation of the extreme quantiles*

Two probabilistic forecasting evaluation metrics are used: the pinball loss and the coverage probability.

Pinball loss and coverage probability: Point forecasts from cubic regression splines, SGBM, AQR, and NLQR models are combined by computing the average, mean, and median of the four models at each point. The pinball loss of each model is computed, and the model with the smallest pinball loss is considered the best. For each model, the coverage probability is estimated. Table 4 presents the pinball losses and the corresponding coverage probabilities.

Table 4 shows that the cubic regression splines model with pinball loss optimisation (M1(CRS)) has the smallest pinball loss value at all extreme values (99%, 95%, 99.5%). The smallest pinball loss at 99.5%. The second best model is the combined forecast model using the weighted average median model with the mean dominating (MeanDom). For all the models, the smallest pinball loss is at 99.5%. Thus, the cubic regression spline model performs better than the single forecast models. From the ensemble methods, the simple average median dominated by the mean is the best.

We further diagnose our fitted models by plotting each quantile's Box and Whiskers plots. Figure 11 presents the Box plots for the pinball losses at all quantile levels for each fitted model.

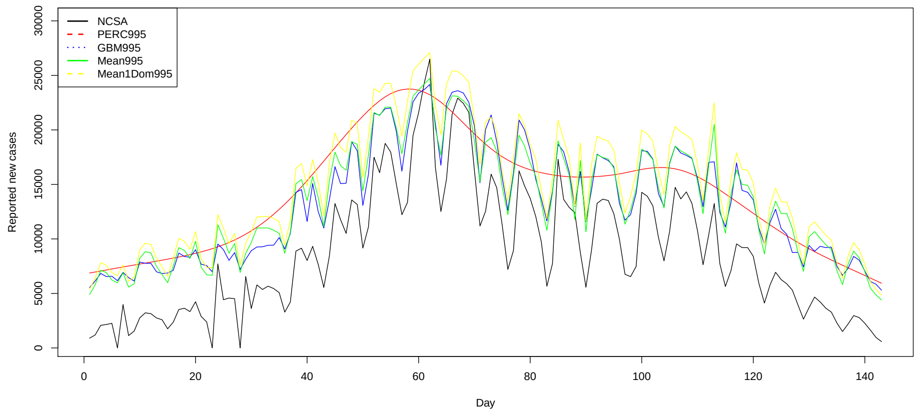

Figure 12. Forecast performance for the best 4 models.

At the 95% quantile level, all models have a longer upper whisker than the lower whisker, with the AQR model having the longest upper whisker. The median values from the SGBM, AQR, Median, and Mean models are fairly equal. The AQR model has fewer outliers compared to the other models. At the 99% quantile level, the Box plots for all the models have their upper whiskers fairly equal to their lower whiskers except for the AQR model, which has its upper whisker longer than its lower whisker. The medial values are almost the same for all models. The AQR model has no outliers, followed by the NLQR model, which has only one outlier. At the 99% quantile level, the upper and lower whiskers for all the models are fairly the same, with the exemption of the AQR model. Generally, the median values are not the same. Most models do not have outliers except for the cubic regression splines and mean models. Overall, the box plots have outliers at the extreme upper tail of the distribution. This shows that the distributions of the pinball losses are positively skewed, indicating extremely high quantiles.

The key performance indicators shown in Table 4 and the Box and Whiskers plots in Figure 11 show that the cubic regression spline model, the gradient boosting method, the mean (average) model, and the simple average-median dominated by the mean model are the best-performing models at the 99.5% quantile in particular. Figure 12 compares the performance of the four models on the test set.

Figure 12 shows that the cubic spline regression model performs well at all the other points except on day 60, around days 65 and 67, and on day 84. These are the days with the highest peaks. The simple average-median model dominated by the mean is the best. It managed to capture all the peaks in the test set. Thus the simple average-median model dominated by the mean has a coverage probability of 100%.

4. Discussion

4.1. General discussion

In this study, we investigated the dataset of South African-reported COVID-19 cases. The data is characterised by a nonlinear trend with extremely high peaks. The distribution of the dataset has a heavy right tail. This motivated the choice of extreme value analysis using extremely high quantiles for all the models. This study proposes a weighted nonlinear approach to combining forecasts from multiple points. This study started with a comparative analysis of the cubic regressions spline (CRS) model, stochastic gradient boosting (SGBM), additive quantile regression (AQR), and nonlinear quantile regression (NLQR) models in predicting extremely high quantiles (0.95-, 0.99- and 0.995-quantiles) of daily COVID-19 cases. A study by Nadaraji and Ojo [29] on extreme value analysis of COVID-19 in West African countries also used the same quantile levels. The cubic spline regression model is based on the residuals from the percentile model. Swaraj and others also used the residuals from the ARIMA model and fed them into the neural network model before forecasting [10].

For each quantile level, point forecasts for the cubic spline regression model, AQR, NLQR, and SGB, were used to build the ensemble algorithm by taking the median from the four models at each forecast point and calculating the mean (average). The accuracy of the median and average models was measured, and the results showed that the mean model performs better than the median model. We then assigned some weights to the mean and median models, bearing in mind that the mean model is more robust. In our weighted mean-median model, a weight of 0.8 on the mean and 0.2 on the median improves the accuracy performance of the model in predicting the extreme values. Thus, we ended up with a mean-dominated model. The prediction performance of the models was based on probabilistic forecasting using the pinball loss (PL), Box and Whiskers plots for each model. The four best-performing models were the cubic spline regression model, the weighted mean-dominated model, the mean model, and the GBM at 99.5% quantile level in that order. The coverage probability of these models was considered, and only the mean-dominated model at the 99.5% had a 100% coverage probability. This study further confirms that combined forecasting from conceptually different models reduces, effectively, prediction error, resulting in improved accuracy [[10], [28], [30], [31]]. The study further proposes the weighted mean-dominated model to forecast daily accurately confirmed COVID-19 cases characterised by extreme values. Thus, the mean-dominated method provides significantly better accuracy than each model. A study by [32] which combines forecasts from neural network models and autoregressive integrated moving average model using nonlinear weighted approaches also supports the efficiency of forecasting using the weighted approaches.

This study shows that the application of EVT provides a promising approach for COVID-19 forecasting, particularly the mean-dominated model. Future work can explore more sophisticated approaches in EVT. Since the appropriate threshold was determined by trial and error, there is a need to explore the models' performance on various quantile levels. In addition, this study was limited to COVID-19-reported cases in South Africa; the findings can be applied in other countries to investigate further findings that lead to generalisability in other regions.

4.2. Practical implications

Extreme quantile estimates provide valuable insights for health authorities by informing them of worst-case scenarios, which helps better plan hospital resources and medical supplies to avoid undersupply. These estimates help design effective surge capacity strategies, ensuring healthcare systems are not overwhelmed during peak periods. The findings also guide the timing and intensity of interventions like lockdowns or social distancing to prevent healthcare systems from being overrun. Public health officials can prioritize vaccination campaigns and enhance compliance with health guidelines by identifying high-risk periods. Furthermore, these insights justify investments in health infrastructure for long-term pandemic preparedness, with adaptable models to update risk estimates as new virus variants emerge.

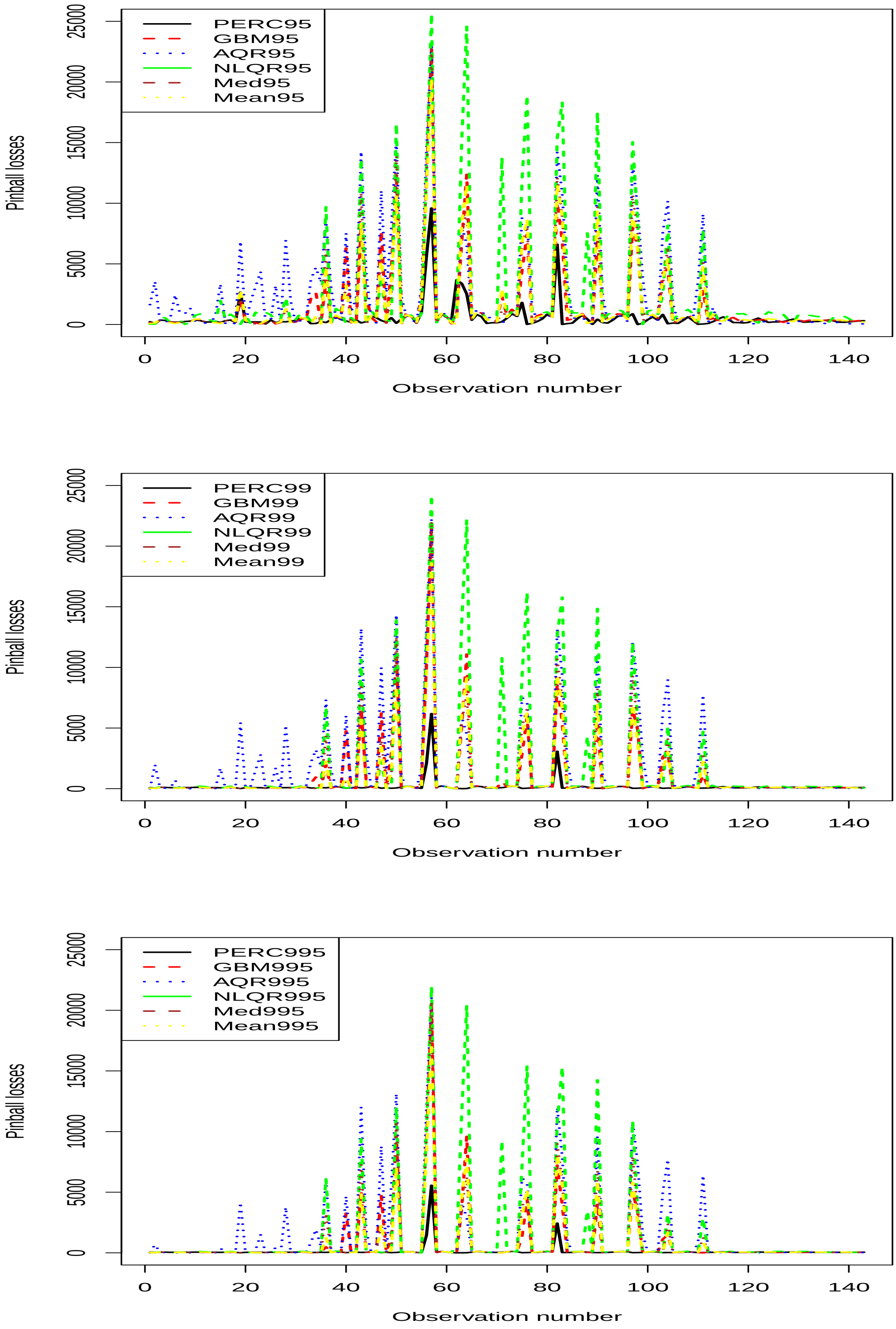

Figure 13. Time series plots of pinball losses at 95%, 99% and 99.5%.

4.3. Limitations

A limitation of the study is the focus on data from South Africa, which is not easily generalizable to other contexts. The unique socio-economic, cultural, and political features of South Africa may bias the findings, making them relevant only to South Africa. Thus, one immediate next step for this line of research would be to conduct the same study in a different setting. Replicating the study to a few more diverse countries will improve the generalizability of the research.

The idea that models remain fixed throughout time is another limitation of our research since it may not apply to changing pandemics marked by dynamic shifts in population behaviour, public health actions, and transmission rates. Future studies will examine how to increase the adaptability and accuracy of models by incorporating time-varying parameters, state-space models, or non-stationary time series techniques.

5. Conclusion

This paper proposes the ensemble method based on the weighted mean-median model dominated by the mean for predicting extremely high quantiles of COVID-19 using South African data. Although this model is the second best from the cubic regression spline model based on the pinball loss, it is the best in terms of coverage probability. The mean-dominated model successfully predicts the peaks in the daily confirmed COVID-19 cases. This study is crucial in the public health sector because it alerts the interested parties on the days on which peaks are likely to occur and helps them to prepare in advance in terms of treatment kits for the infected, protective clothing for the staff in attendance, enough bedding for hospitalised cases, vehicles to ferry the seriously infected patients. Thus, the study recommends using the model to anticipate potential outbreaks and implement potential mitigation measures by the Public Health personnel.

Data Availability Statement

The data that support the findings of this study are openly available at https://github.com/csigauke/Data-for-article-Estimation-of-extreme-quantiles

Authorship Contribution

Claris Shoko: Conceptualization, Methodology, Visualization, Investigation, Validation, Programming, Writing-Original draft preparation. Caston Sigauke: Supervision, Methodology, Validation, Programming, Writing preparation, Reviewing, Editing.

Acknowledgments

The authors would like to acknowledge the University of Botswana and the University of Venda for giving us time to research.

Appendix: Time series plots of pinball losses.

REFERENCES

1. Chen, J., Lei, X., Zhang, L. and Peng, B, (2015). *Using Extreme Value Theory Approaches to Forecast the Probability of Outbreak of Highly Pathogenic Influenza in Zhejiang, China*. PLoS ONE, 10(2): e0118521. doi:10.1371/journal.pone.0118521
2. Daouia, A., Stupfler, G., Usseglio-Carleve, A. (2023). *Extreme value modelling of SARS-CoV-2 community transmission using discrete generalized Pareto distributions*. Royal Society Open Science, 10, 220977. https://doi.org/10.1098/rsos.220977
3. Fisher, R.A. and Tippett, L.H.C. (1928). *Limiting forms of the frequency distribution of a sample's largest or smallest member*. Math Proc Cambridge, 24, 180–190. https://doi.org/10.1017/S0305004100015681
4. Zhao, X. (2010). *Extreme value modelling with application in finance and neonatal research*. PhD Thesis, The University of Canterbury. Available: http://ir.canterbury.ac.nz/bitstream/10092/4024/1/thesis_fulltext.pdf
5. Sciannameo, V., Goffi, A., Maffeis, G., Gianfreda, R., Pagliari, D.J., Filippini, T., Mancuso, P., Giorgi-Rossi, P., Dal Zovo, L.A., Corbari, A., Vinceti, M.H. and Berchialla, P. (2022). *A deep learning approach for Spatio-Temporal forecasting of new cases and*

new hospital admissions of COVID-19 spread in Reggio Emilia, Northern Italy. Journal of Biomedical Informatics, 132, 104132. https://doi.org/10.1016/j.jbi.2022.104132.

6. Aljaaf, A.J., Mohsin, T.M., Al-Jumeily, D. and Alloghani, M. (2021). *A fusion of data science and feed-forward neural network-based modelling of COVID-19 outbreak forecasting in IRAQ*. Journal of Biomedical Informatics, 118, 103766. https://doi.org/10.1016/j.jbi.2021.103766
7. Safari, A., Hosseini, R. and Mazinani, M. (2021). *A novel deep interval type-2 fuzzy LSTM (DIT2FLSTM) model applied to COVID-19 pandemic time-series prediction*. Journal of Biomedical Informatics, 123, 103920. https://doi.org/10.1016/j.jbi.2021.103920
8. Sun, X., Wang, Z. and Hu, J. (2017). *Prediction interval construction for byproduct gas flow forecasting using optimised twin extreme learning machine*. Mathematical Problems in Engineering, vol. 2017, 1–12. https://doi.org/10.1155/2017/5120704
9. Wang, Y., Zhang, N., Tan, Y., Hang, T., Kirschen, D.S. and Kang, C. (2018). *Combining probabilistic load forecasts*. arXiv, stats., 1–10. https://doi.org/10.48550/arXiv.1803.06730
10. Swaraj, A., Verma, K., Kaur, A., Singh, G. and Kumar A. (2021). *Leandro Melo de SalesImplementation of stacking based ARIMA model for predicting Covid-19 cases in India*. Journal of Biomedical Informatics, 121, 103887. https://doi.org/10.1016/j.jbi.2021.103887
11. Maswanganyi, N., Sigauke, C. and Rangani, E. (2021). *Prediction of Extreme Conditional Quantiles of Electricity Demand: An Application Using South African Data*. Energies, 14(20), 6704 https://doi.org/10.3390/en14206704
12. Enriquez, D.C., Niembro-Ceceña, J.A., Mandujano, M.M., Alarcon, D., Guerrero, J.A., Garcia, I.G., Gutierrez, A.A.M. and Gutierrez-Lopez, A. (2022). *Application of probabilistic models for extreme values to the COVID-2019 epidemic daily dataset*. Data in Brief, 40, 107783, 1–10. https://doi.org/10.1016/j.dib.2021.107783
13. Liu, P. and Zheng, Y. (2023) *Heavy-tailed distributions of confirmed COVID-19 cases and deaths in spatiotemporal space*. PLoS ONE, 18(11), e0294445. https://doi.org/10.1371/journal.pone.0294445
14. Wong, F. and Collins, J.J. (2020). *Evidence that Coronavirus superspreading is fat-tailed*. PNAS, 117(47), 29416–29418. doi:10.1073/pnas.2018490117/-/DCSupplemental
15. Greenwell, B., Boehmke, B. and Cunningham, J. (2022). *Generalised Boosted Regression Models*. GBM R package, Version 2.1.8.1 https://cran.r-project.org/web/packages/gbm/index.html
16. Koenker, R., Portnoy, S., Ng, P.T., Melly, B., Zeileis, A., Grosjean, P., Moler, C., Saad, Y., Chernozhukov, V. and Fernandez-Val., I. (2022). *quantreg: Quantile Regression*. R package, version 5.94. https://cran.r-project.org/web/packages/quantreg/index.html
17. Fasiolo, M., Wood, S.N., Zaffran, M., Goude, Y. and Nedellec, R. (2022). *Smooth Additive Quantile Regression Models*. Available online: https://cran.r-project.org/web/packages/qgam/index.html (Accessed on 17 January 2024).
18. Roach, C. (2017). *Functions for GEFCOM 2017*. R package Version 0.3.0 Available online: https://rdrr.io/github/camroach87/gefcom2017/ (Accessed on 25 January 2024).
19. Sherwood, B. and Maidman, A. (2022). *Additive nonlinear quantile regression in ultra-high dimension*. Journal of Machine Learning Research, 23, 1–47. https://www.jmlr.org/papers/volume23/19-697/19-697.pdf
20. Friedman, J.H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
21. Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition, Springer: Berlin/Heidelberg, Germany.
22. Friedman, J.H. (2002). *Stochastic gradient boosting*. Comput. Stat. Data Anal., 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2
23. Gaillard, P., Goude, Y. and Nedellec, R. (2016). *Additive models and robust aggregation for GEFcom2014 probabilistic electric load and electricity price forecasting*. International Journal of Forecasting, 32, 1038–1050. https://doi.org/10.1016/j.ijforecast.2015.12.001
24. Fasiolo, M., Wood, S.N., Zaffran, M., Nedellec, R. and Goude, Y. (2021). *QGAM: Bayesian Nonparametric Quantile Regression Modeling in R*. Journal of Statistical Software, 100(9), 1–31. https://doi.org/10.18637/jss.v100.i09
25. Koenker, R. (2017). *Quantile regression: 40 years on*. Annual Review of Economics, 9, 155–176. https://doi.org/10.1146/annurev-economics-063016-103651
26. Gaba, A., Tsetlin, I. and Winkler, R.L. (2017). *Combining interval forecasts*. Decision Analysis, 14, 1–20. https://doi.org/10.1287/deca.2016.0340
27. Hyndman, R.J. (2020). *Quantile Forecasting With Ensembles and Combinations*. Accessed on Aug. 20, 2023. [Online]. Available: https://robjhyndman.com/publications/quantile-ensembles/
28. Thorey, J., Chaussin, C. and Mallet, V. (2018). *Ensemble forecast of photovoltaic power with online CRPS learning*. International Journal of Forecasting, 34(4), 762–773. https://doi.org/10.1016/j.ijforecast.2018.05.007. hal-01909827v2
29. Nadarajah, S. and Ojo, O.O. (2023). *An extreme value analysis of daily new cases of COVID-19 for sixteen countries in West Africa*. Scientific Report, 13, 10814. https://doi.org/10.1038/s41598-023-37722-9
30. Shoko, C. and Sigauke, C. (2023). *Short-term forecasting of COVID-19 using support vector regression: An application using Zimbabwean data*. American Journal of Infection Control, 51(10), 1095–1107. https://doi.org/10.1016/j.ajic.2023.03.010
31. Shoko, C., Sigauke, C. and Njuho, P. (2022). *Short-term forecasting of confirmed daily COVID-19 cases in the Southern African Development Community region*. Afrian Health Sciences, 22(4), 534–550. https://dx.doi.org/10.4314/ahs.v22i4.60
32. Adhikari, R. and Agrawal, R.K. (2012). *A Novel Weighted Ensemble Technique for Time Series Forecasting*. In: Tan, PN., Chawla, S., Ho, C.K., Bailey, J. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2012. Lecture Notes in Computer Science(2012), vol 7301. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30217-6_4