31

Data-driven approach to COVID-19 infection forecast for Nigeria using negative binomial regression model

Chollette C. Olisah[1], Olusoji O. Ilori[2], Kunle Adelaja[3], Patience U. Usip[4], Lazarus O. Uzoechi[5], Ibrahim A. Adeyanju[6], Victor T. Odumuyiwa[7]

[1]DEPARTMENT OF COMPUTER SCIENCE, BAZE UNIVERSITY, ABUJA, NIGERIA; [2]DEPARTMENT OF ELECTRICAL/ELECTRONIC ENGINEERING, OBAFEMI AWOLOWO UNIVERSITY, ILE-IFE, NIGERIA; [3]DEPARTMENT OF MECHANICAL ENGINEERING, UNIVERSITY OF LAGOS, AKOKA, NIGERIA; [4]DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF UYO, UYO, NIGERIA; [5]DEPARTMENT OF ELECTRICAL/ELECTRONIC ENGINEERING, FEDERAL UNIVERSITY OF TECHNOLOGY OWERRI, OWERRI, NIGERIA; [6]DEPARTMENT OF COMPUTER ENGINEERING, FEDERAL UNIVERSITY OYE-EKITI, EKITI, NIGERIA; [7]DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF LAGOS, AKOKA, NIGERIA

1. Introduction

Since the first confirmed case of Coronavirus, SARs-CoV-2 (COVID-19), in China's city of Wuhan in December 2019, the pandemic has spread across the globe with over 190 countries affected. In Nigeria, the first confirmed case of COVID-19 was detected on 25th February 2020, in the city of Lagos. Since the index case, the outbreak has spread to 32 states of the country. The first set of confirmed cases was reported to be from foreigners and nationals who visited Nigeria from different countries mainly Italy, the United Kingdom, and the United States of America. Though there were several interventions the governments put in place to curb the spread of the virus, however, as of 28th of April 2020, Nigeria had recorded 1532 cases, 44 deaths, and 255 recoveries from the virus [1]. Though these statistics are minimal compared to the record of most countries in Europe and America, some experts believe that there is still going to be an exponential rise of the infected cases in Nigeria if the measures put in place by the governments are not adequately coordinated. It is, therefore, imperative to bring about scientific efforts toward its control. One of such efforts is the prediction of the future infection pattern of the virus to enable the governments at the Federal and State levels to make informed health-related decisions.

Data Science for COVID-19. https://doi.org/10.1016/B978-0-12-824536-1.00002-2

Typically, in epidemiological cases, from the first instances, a high health-risk disease is identified, researchers often sort the use of mathematical models to predict the course of the disease over time [2—8]. Mathematical models can be broadly classified into mechanistic models and empirical models [9]. Mechanistic models require detailed knowledge and data on the underlying problem.

In places where precise data and computational resources are available, mechanistic models may suffice for predicting coronavirus infection patterns and modeling the impact of various intervention strategies more accurately to inform policymakers and health workers [8]. In Nigeria, detailed and comprehensive data are not available. The data available are not precise enough for accurate mechanistic modeling. There are constraints inherent in the collection of the data because of scarce resources such as testing kits, and inadequate sampling strategies. Also, there is the problem of delayed infection case reports and under-reporting. As a result, the data are noisy [10]. Empirical models can be appropriate for extracting patterns in the available data and forecasting coronavirus transmission dynamics. However, the accuracy of any mathematical and statistical models depends heavily on assumptions, parameters, and theory. To state: how good is the assumptions on which the model is based [11—13], how significant are the estimated parameters in modeling a given infectious disease and within a geographical region [14], and is the model formulated based on theory. Alternatively, with an increase in real data points as the disease progresses, computational models, which are also grounded in mathematical models, can be explored to predict the future growth pattern of disease.

The use of computational models in epidemiology dates back to the 1980s [15], and it is still prevalent in modern epidemiology. Some commonly used in modern epidemiology are linear regression (LR) [3,16], poisson regression (PR), negative binomial regression (NBR) [17,18], exponential smoothing (ES) [19], autoregressive integrated moving average [20], support vector machines. These models can be adapted to a given problem. However, there is not a fit-all model, each problem's outcome data differ by type and distribution. By outcome data, we mean the data of the dependent variable and the term will be used as stated throughout this paper except where otherwise specified.

For the current prediction trend of COVID-19, Petropoulos and Makridakis [21] adopted a computational model from the ES family for the prediction of the global cumulative confirmed cases of COVID-19. Roosa et al. [22] employed and compared the generalized logistic growth model, the Richards growth model, and a subepidemic wave model capability to objectively forecast future global cases of COVID-19. Jia et al. [23] employed and analyzed the logistic model, Bertalanffy model, and Gompertz model to fit and analyze the situation of COVID-19. Anastassopoulou et al. [3] compared the predictions of LR along-side the susceptible-infectious-recovered-dead (SIRD) model for COVID-19 future spread. Considering that the approaches used to adapt a computational model to the COVID-19 data differs across these models, their prediction results will not be compared in this work. Rather, we take the time to acknowledge the quick prediction response they have made and present further, our prediction methodology as it contributes to knowledge, and facilitates health-care interventions.

In this work, we present an epidemiological prediction that is uniquely fashioned for Nigeria, though can be adapted for other countries. The reason for the focus on Nigeria is the fact that there is a gradual growing spread of the virus as reported by the Nigerian Center for Disease Control (NCDC) [1] and because of the heedless attitude of Nigerians about the virus. Frequently, people cluster to purchase from malls, the market, receive food aid, and without protective measures in place for preventing the spread of the virus in such gatherings. Yet, of her approx. 195 million people, only about 10,861 persons, constituting about 0.00557% of the population has so far been tested as at 28-APR-2020. Based on the presented case, we make the following assumptions:

> *There are a lot more people who are carriers of COVID-19 but are not showing symptoms because they are self-medicating or illegally treated at unauthorized private hospitals, which is likely to suppress the symptoms and make carriers go unnoticed. This is because Nigeria isn't running as many tests as possible, given the uncontainable interaction of people in gatherings, carriers further transmit to more persons.*

Having stated these, we hypothesize that there is a causal relationship between testing and identification of COVID-19 carriers in Nigeria. Therefore, we make the following statement:

> *"With an increase in COVID-19 testing relative to the suspected percentage of carriers in Nigeria, the number of infected cases will increase significantly"*

Subsequently, we will identify the predictor variables meaningful to the given case for the prediction of the infection pattern of the virus over some time. Unlike previous approaches [3,21–23], we explore a prediction model from the family of generalized linear models (GLMs) for our prediction.

2. Material and methods

This section comprises data, preprocessing, and the prediction model.

2.1 Data

Before data collection, it is important to identify the outcome and predictor variables to avoid erroneous data collection. Therefore, based on our hypothesis, the outcome variable: *infected count*, and the predictor variables: *number tested*, *number of deaths*, *time*, were identified with no consideration to "best predictors" given. Furthermore, these variables helped in the collection of the *number of tested* data[1] and the rest of the variable's data.[2] Since computational models for epidemiological prediction usually require historical data collected over a long period for prediction accuracy, we capitalized on collecting the daily incidence of COVID-19 from countries with a sufficient number of test counts to create baseline data. These countries are South Africa, Senegal, Slovenia, Australia, Belgium, and Israel.

As a result of the inclusion of the *number tested* predictor variables, the records from Nigeria as provided by the NCDC is omitted because there was no accurate daily record of the number of tested data from the period of March 9, 2020 to April 19, 2020. Also, guided by the trend of occurrence of infection in Nigeria, we consider records from the USA, UK, and Italy to be too extreme to be included in the data model.

2.2 Preprocessing

First, the data collected from the various countries earlier mentioned in Section 2.1 are merged into a single data. To enable each sample to still represent an infection count, two indexes are set; one for index and the other for the date entry. Second, a few *null values* appeared in the data. To generate the missing values, we adopted the linear interpolation method which estimates the null values from known values closest to it. A sample of the data after the processing is shown in Fig. 31.1. Third, based on our earlier assumption that the "percentage of infected" will be a significant factor for identifying new carriers of the virus, we employed feature engineering approach by creating a *percentage-suspected* of COVID-19 carriers as an additional predictor variable. This is to

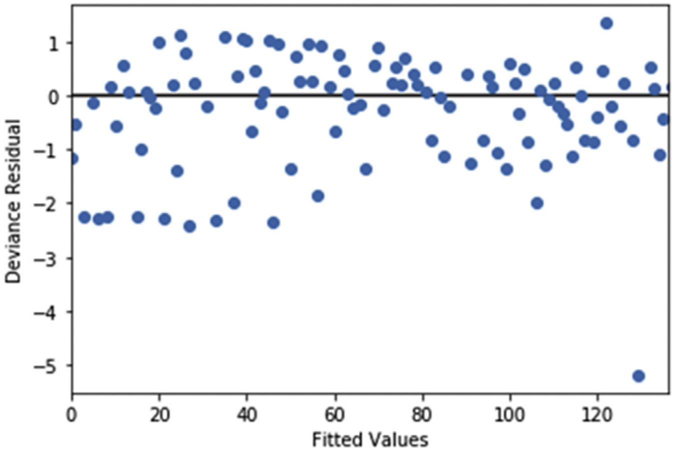

FIGURE 31.1 Deviance residual plot for the proposed model for the COVID-19 data.

[1] www.worldometers.info/coronavirus/.

[2] www.ourworldindata.org/covid-testing.

Table 31.1 Current COVID-19 baseline data from South Africa, Senegal, Slovenia, Australia, Belgium, Israel on infection and death cases.

DATE	Infected	number_of_deaths	number_tested	perc_suspected
March 09, 2020	4	0	70	5.7
March 10, 2020	0	0	101	0
March 11, 2020	6	0	90	6.7
March 12, 2020	3	0	203	1.5
March 13, 2020	8	0	76	10.5
March 14, 2020	14	0	93	15.1
March 15, 2020	23	0	459	5
March 16, 2020	3	0	867	0.3
March 17, 2020	21	0	568	3.7
March 18, 2020	31	0	159	19.5
March 19, 2020	34	0	1762	1.9
March 20, 2020	52	0	1606	3.2
March 21, 2020	38	0	987	3.9

enable us to test whether there is an effect of increasing testing capacity relative to the growing number of people suspected to be carriers of the virus and the prediction of the outcome.

After the data has been prepared, it is split into training and testing sets with a distribution of 80% and 20%, respectively, that is, 136 observations for the train set to 44 observations for the test set. The data sample is provided in Table 31.1. Different from other predictions of COVID-19 in the literature [22,23], our predictive model is trained on global data, of the countries aforementioned, to extract meaningful cues of the virus spread pattern and subsequently adapt the model to forecasting the future infection spread of the virus for Nigeria. This approach is inspired by transfer learning used when there is a limited number of samples for recognition tasks like [24] and basically because there are numerous missing values in the number of test data from the Nigeria COVID-19 reports. It should be noted that the aspect of transfer learning we refer to is the act using prediction patterns gained from a different problem to deduce the pattern of a different but related problem. Therefore, in predicting the future of COVID-19 infection count in Nigeria, we generated seven-test data, of which 3–7 are as stipulated[3]

(1) data that fit the current daily testing capacity in Nigeria (see Table 31.2).
(2) data that follow a 300 increase in a testing capacity.
(3) data that meet an expected 1500 daily testing capacity of Nigeria.
(4) data that meet an expected 2000 daily testing capacity of Nigeria.
(5) data that meet an expected 2500 daily testing capacity of Nigeria.
(6) data that meet an expected 3500 daily testing capacity of Nigeria.
(7) data that meet an expected 5000 daily testing capacity of Nigeria

[3] https://covid19.ncdc.gov.ng/media/files/COVID19TestingStrategy_Lz3ZVsT.pdf.

Table 31.2 Generated COVID-19 data for Nigeria infected and death cases.

Date	Infected	number_of_deaths	number_tested	perc_suspected
April 20, 2020	1	1	584	7
April 21, 2020	1	3	347	25
April 22, 2020	1	3	588	20
April 23, 2020	1	3	539	22
April 24, 2020	1	0	370	23
April 25, 2020	1	1	487	22
April 26, 2020	1	2	450	20
April 27, 2020	1	1	500	20
April 28, 2020	1	0	550	19
April 29, 2020	1	0	550	20
April 30, 2020	1	2	421	26
May 01, 2020	1	0	433	24
May 02, 2020	1	0	584	20
May 03, 2020	1	1	418	25

2.3 Prediction model

We employ a type of GLMs. This family of models is chosen because they are known for their powerful application to prediction problems of count data. Also, for the fact that they can be used to validate the relationship between variables to judge the contribution each variable makes to the model performance.

From the initial descriptive analysis, the population distribution is observed to be skewed and to approximate the Poisson distribution. However, the respective means of the outcome data show to greatly deviate from the variance. In statistics, this is termed over dispersion. By definition, overdispersion can be described as when data variance is greater than its statistical mean [25]. This characteristic of the data violates fitting the data to the PR model, a commonly used model for fitting epidemiological count data. Therefore, we explore the NBR model for fitting the count data. From the literature, the NBR is more appropriate for fitting overdispersed count data [26] and very much adopted in solving epidemiological problems like in Refs. [15,17,18,26–28]bib28.

With the NBR model, we consider the goal of predicting the outcome variable y_i, which is the number of infected cases at observation i, given the exposure time t_i, and a set of predictor variables $x_{1i}, x_{2i}, \ldots, x_{ki}$ at observation i. Thus, the model is formulated as:

$$P(\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \tag{31.1}$$

where α is the dispersion parameter, Γ is the gamma function, and μ_i is the expected mean value of y_i per t_i, which is given as:

$$\mu_i = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}), \quad (i = 1, 2, \ldots, k) \tag{31.2}$$

where β_0 is the intercept and the unknown parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are regression coefficients estimated using the maximum likelihood method [25,29]. Consequently, the future observations of the infection pattern of the virus at time $t+n, n>0$ can be predicted.

In applying Eq. (31.1) to COVID-19 data, the full model representation of μ_i can be specified as:

$$\mu_i = exp(intercept + percentage\_suspected + number\_tested + number\_of\_deaths + day) \tag{31.3}$$

The implementation of the prediction model was achieved using the Statsmodel v0.12.0.dev0 (v207) application programmer interface in Python Environment. The prediction model result and analysis will be presented and discussed in subsequent sections.

3. Results and discussion

The results will be presented as follows: (1) goodness of fit of the model to the data, (2) testing the effect of predictor variables on the outcome variable, (3) model prediction performance on unknown data, the Nigeria data, for one-month.

3.1 Goodness of fit test

In accessing the fit of the NBR model to the data, we used the Chi-square goodness of fit statistical measure as proposed in Ref. [30]. Based on the author's recommendations, we evaluate the deviance value with the model degree of freedom (see Table 31.3) at a 5% significance level using the following formula.

$$\chi^2 = \left[\frac{(O-E)^2}{E}\right] \tag{31.4}$$

where χ^2 is Chi-square goodness of fit, O is observed value, and E expected value.

The P-value determined from the Chi-square distribution calculator, $P(\chi^2) = 0.41144$, suggests that the Chi-square test is not statistically significant. Therefore, we conclude that the NBR model fits the data well.

Furthermore, our claim of the fitness of the proposed model can also be verified using a plot of the deviance residual and the fitted value, which is illustrated in Fig. 31.1. As expected, the deviance of the proposed model lies along the zero point and shows no evidence of one-directional bias, either of overestimation or underestimation given the

Table 31.3 Statistical values for verifying model goodness of fit.

	Value	Df	χ^2	*P*-value
Deviance	133.93	131	0.0227	0.41144
Log-likelihood	−778.72	–	–	–

closeness of the median of the residuals, 0.03912, to zero. Even though an outlier observation is identified, the median value suggests that it does not statistically differ from others.

3.2 Model effect and statistical analysis of predictor variables

Here, we report the model effect of the predictor variables given in Eq. (31.3) and their statistical significance to predicting the outcome count. These reports are tabulated in Table 31.4. To interpret the result of Table 31.4, we draw the attention of the reader to the coefficient estimates for the model effect and the *P*-value for statistical significance.

3.2.1 Statistical significance of predictor variables

The statistical analysis of the predictor variables using the *P*-values at 5% significance value reveals that the predictor variables except *death* are statistically significant to the model outcome. While the *number_tested* and *percentage_suspected* variables show to be of very high significance given their 0.000 *P*-values which are way below the 0.01 significance level. The *day* variable with a *P*-value of 0.030 reveals a high significance. As expected, the *number_of_deaths* variable does not influence the pattern of infection spread per t_i. Since the NBR model performance is in line with our assumptions, it clearly expresses the prediction power of the model for solving the given problem.

3.2.2 Model effect of predictor variables

The coefficient estimate is not informative by itself. So, we adopted an interpretative strategy for a coefficient estimate as provided in Ref. [31], it is given as:

$$H = 100 * [exp \; exp(\beta * \Delta) - 1] \tag{31.5}$$

where H is the percentage change in the expected mean of y_i, $\Delta \equiv 1$, which represents the one-unit change, β is the regression coefficient.

Applying Eq. (31.5) to the predictor variables *number_tested*, *percentage_suspected*, *day*, *number_of_deaths* gives the values 0.05%, 7.9%, 2.2%, 0.08%, respectively. By interpretation, the exponentiated value of each predictor regression coefficient indicates how much the mean of y_i, that is, μ_i changes with every one-unit increase in X, while holding other predictors constant. For instance, the *percentage_suspected* predictor with

Table 31.4 Model outcome and predictor effect and statistical significance verification.

	Estimates	**Std Error**	***P*-value**
Intercept	2.2290	0.202	0.000
Number_tested	0.0005	2.45e-05	0.000
Percentage_suspected	0.0763	0.008	0.000
Day	0.0218	0.010	0.030
Number_of_deaths	0.0008	0.002	0.659

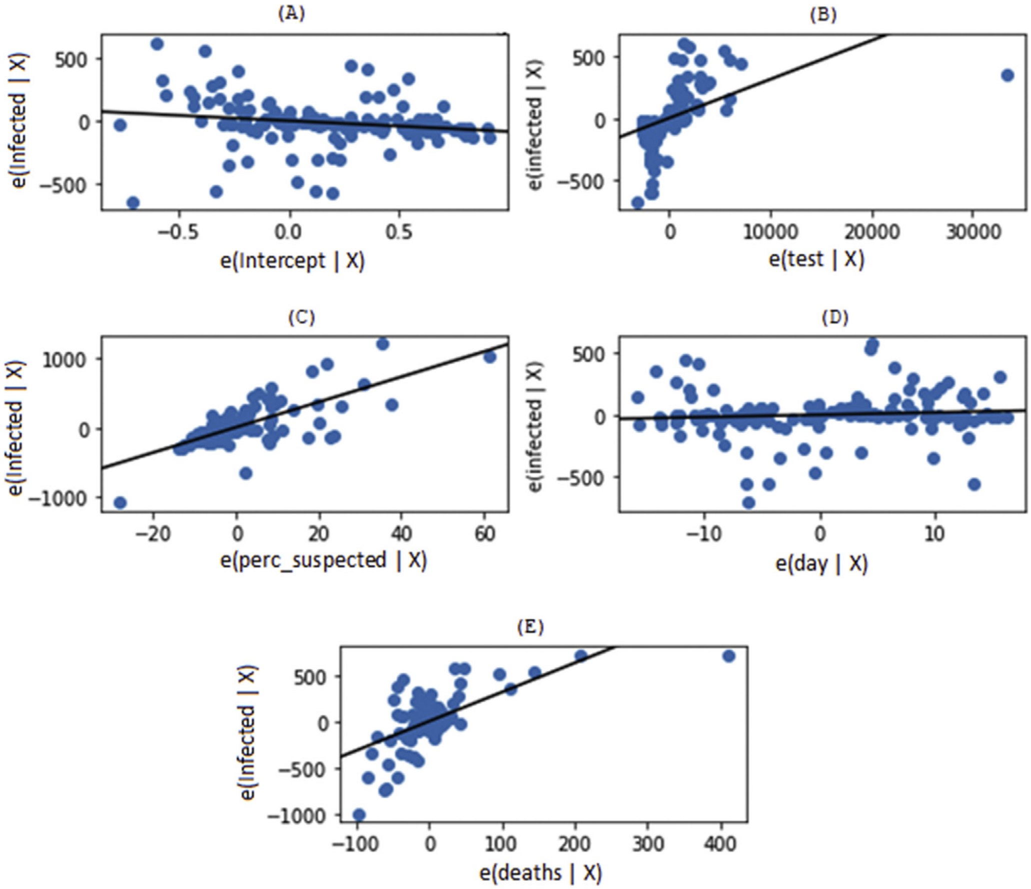

FIGURE 31.2 Partial regression plots of the effect of the predictor variables on the outcome variable. (A) outcome variable against the predictor variable, (B)–(E) outcome variable against variables X_kX omitting $X_{\sim k}$.

$H = 7.9\%$ means that for each one-unit increase in the percentage of the number of people infected, the mean count of the number of infected persons will increase by 7.9%, assuming all other variables have a zero value.

An alternative approach to evaluating the relationship between the outcome variable y and a predictor variable X_k, conditional on other predictor variables $X_{\sim k}$, is through a visual plot of the residuals retrieved by regressing the outcome variable against X_k. The partial regression plot available in the Statsmodel v0.12.0. dev0 (v207) is used to achieve this goal. From Fig. 31.2, it is obvious that all observations for the predictor variables were consistently close to the trend line, though a compact spread along the trend line is seen for *deaths* and *test* predictors which results from the presence of the outlier. However, there is a lack of trend for the *day* predictor which illustrates that it is not as explanatory as the regression coefficients suggested.

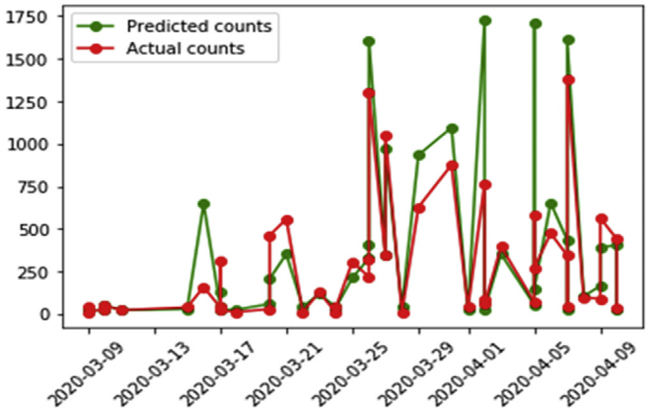

FIGURE 31.3 Predicted $\hat{y}_i$ and actual y_i observations of infected cases of COVID-19.

3.3 Prediction accuracy

The capability of the predictive model can be seen in Fig. 31.3 which illustrates the plot of the predicted observation of infected cases versus the actual observation of infected cases over the timestamps of dates from 09-03-2020 to 09-04-2020. These dates represent the periods most countries began recording numerous cases. It is interesting to see that the predictive distribution for predicted resembles the actual, though at some observations (13-03-2020 to 15-03-2020, 30-03-2020 to 04-04-2020) the model over predicted. However, the predicted observation of infected cases closely resembles the actual observation.

Of particular interest in epidemiological predictions is the ability to project into the future the spread pattern a disease might take over a duration of time to help facilitate health-care decisions. Henceforth, we term this phenomenon, forecasting. We consider the forecast distribution of future observations for infection pattern of the virus at time $t + n, n > 0$ to be of great significance to public health in Nigeria. This is mainly because a clear picture of the infection threats of COVID-19 is still not well-understood.

Using the generated data [1–7] discussed in subsection 2.2, which represents Eq. (31.3) variables, for testing the predictive power of the model to unseen data. The future cumulative numbers of infected cases for data [1–7] are illustrated in Fig. 31.4D and Fig. 31.5D. This is a report on a thirty-days-ahead forecast and illustration of the effectiveness of the proposed model when applied to COVID-19.

We assume that there are more COVID-19 infected cases than is reported and if being the case then, the Government must increase its COVID-19 testing capacity. As observed from Figs. 31.4 and 31.5, an increase in the testing capacity increased the number of infected cases. Though Refs. [1–7] of the generated data is created from the actual COVID-19 Nigeria data, it is interesting to observe the pattern of the spike on day 30 of the trained model reoccur on the daily and cumulative plot of the infected cases of COVID-19 forecast. Also, we observe that the forecast errors between February 20, 2020

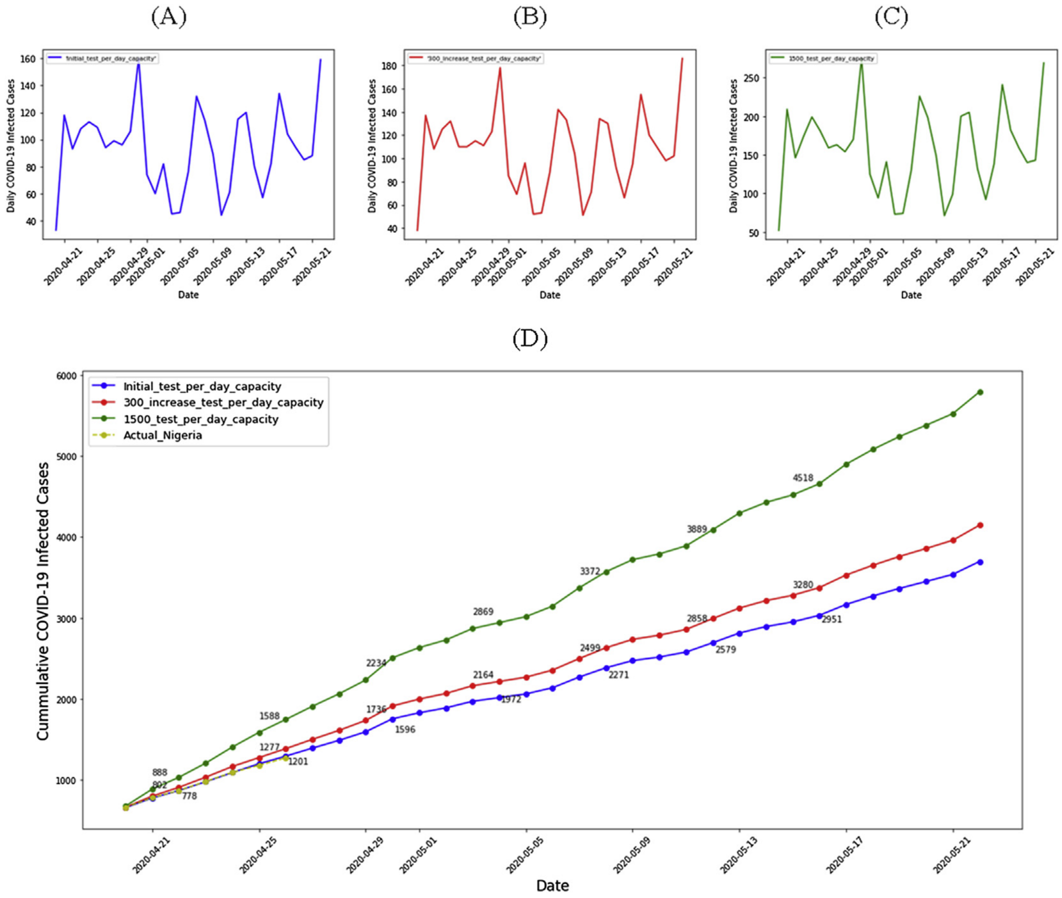

FIGURE 31.4 A 30-day forecast of infected cases of COVID-19. (A)–(C) the daily forecast of confirmed cases for generated data [1–3], (D) the cumulative COVID-19 forecast for generated data [1–3] together with actual COVID-19 Nigeria data.

to February 26, 2020, as can be seen from Fig. 31.4, are within the median of the trained model residuals (median Error = 0.03912). Days 21, 22, and 26 have negative percentage errors of −0.01%, −0.02%, and −0.03, respectively; while for days 20, 23, 24, and 25 have positive percentage errors of 0.08, 0, 0.01, and 0.22, respectively.

Interestingly, the cumulative predicted number of infected cases in Nigeria is expected to continue to rise in the coming weeks as seen from Figs. 31.4D and 31.5D. The growth level expected on the 30-04-2020 for the three-scenarios of testing capacities are: (1) scenario of gradual increase in testing capacity as is currently practiced in Nigeria which is labeled “initial_test_per_day_capacity” is 1756, (2) scenario of a 300 increase in the current testing capacity labeled as “300_increase_test_per_day_capacity” is 1914, (3) scenario of achieved 1500 daily test labeled as “1500_test_per_data_capacity” is 2509.

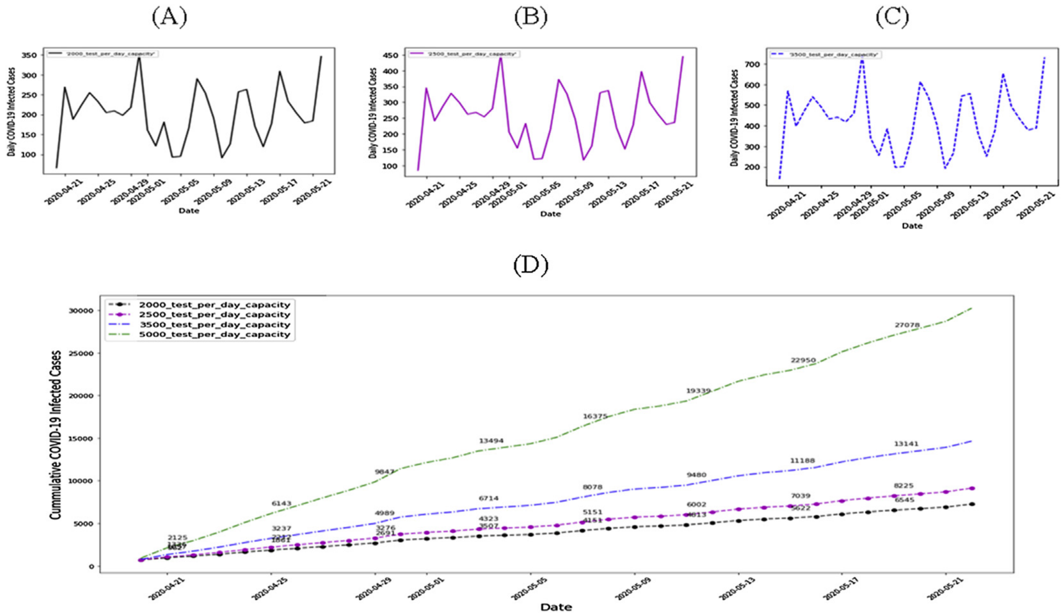

FIGURE 31.5 A 30-day forecast of infected cases of COVID-19. (A)–(C) the daily forecast of confirmed cases for generated data [4–7], (D) the cumulative COVID-19 forecast for generated data [4–7] together with actual COVID-19 Nigeria data.

By 15-05-2020 and 22-05-2020, the infected count is expected to rise to 2951 and 3697, 3280 and 4145, 4518 and 5790 for scenario 1, scenario 2, and scenario 3, respectively.

Furthermore, the worst infected number of COVID-19 cases, in Nigeria, can be observed for testing capacity from 2000 up to 5000. If Nigeria eventually carries out more tests as projected for the coming days, there will be more and more persons in need of health care facilities. As predicted, precisely about 7254, 9135, 14,639, 30,244 infected number of COVID-19 cases by the 22-05-2020 might be identified for the scenarios of 2000-test-per-day-capacity, 2500-test-per-day-capacity, 3500-test-per-day-capacity, and 5000-test-per-day-capacity, respectively. Therefore, care should be taken as Nigeria currently considers relaxing lockdown in the coming weeks without careful deliberations on the potential risk and ways to mitigate against it.

While the benefits of the lockdown can be observed through the gradual rise of the COVID-19 infected cases, we should be wary of the uncontainable interaction of people in markets, malls, shops around people's homes, and the contagiousness of the act of gathering a cluster of people to give aids to people by noble Nigerian philanthropist. If these gatherings are not contained, Nigeria should expect a spike that is by far more than the worst case of the 1500-testing-capacity infected number of cases. The predicted cases from 2000 up to 5000 testing capacity cases reveal so.

4. Conclusion

This paper explored the NBR model from the family of GLM for the prediction of the future infection pattern of COVID-19 in Nigeria. We approached the prediction from a whole new perspective that is inspired by transfer learning and feature engineering approaches widely used in machine learning. We trained the model to learn COVID-19 pattern cues of countries such as South Africa, Senegal, Slovenia, Australia, Belgium, and Israel with sufficient and recorded infection cases and test count as baseline data for forecasting infection trends in Nigeria. The experimental results showed the effectiveness of the proposed approach to predict the test set of the trained data and forecast a rise in the infected number of COVID-19 cases in Nigeria, which closely resembles the actual number of infected cases in Nigeria.

Acknowledgments

We would like to express our sincere gratitude to the Association of Massachusetts Institute of Technology Trained African Universities Lecturers (AMTAUL) for providing the platform for this research collaboration. We acknowledge the role of Professor Akintayo Akinwande (Program Coordinator) and TOTAL Nigeria for supporting the program.

References

[1] Nigerian Center for Disease Control, COVID-19 Case Update, 2020. Retrieved from covid19.ncdc.gov.ng [Accessed 28th April 2020].

[2] F.A. Hamzah, C.H. Lau, H. Nazri, D.V. Ligot, G. Lee, C.L. Tan, et al., CoronaTracker: World-wide COVID-19 outbreak data analysis and prediction, Bull. World Health Organ. (2020), https://doi.org/10.2471/BLT.20.255695.

[3] C. Anastassopoulou, L. Russo, A. Tsakris, C. Siettos, Data-based analysis, modelling and forecasting of the COVID-19 outbreak, PLoS One 15 (3) (2020).

[4] M. Arti, K. Bhatnagar, Modeling and predictions for COVID 19 spread in India, Researchgate (2020), https://doi.org/10.13140/RG.2.2.11427.81444.

[5] V.A. Okhuese, Mathematical Predictions for COVID-19 as a Global Pandemic. medRxiv, 2020, https://doi.org/10.1101/2020.03.19.20038794.

[6] I. Nesteruk, Statistics-based predictions of coronavirus epidemic spreading in mainland China, Innov. Biosyst. Bioeng. 4 (1) (2020) 13—18, https://doi.org/10.20535/ibb.2020.4.1.195074.

[7] L. Peng, W. Yang, D. Zhang, C. Zhuge, L. Hong, Epidemic Analysis of COVID-19 in China by Dynamical Modeling. arXiv.

[8] N.M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, et al., Imperial College COVID-19 Response Team. Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand, 2020, https://doi.org/10.25561/77482.

[9] Y. Pawitan, In All Likelihood: Statistical Modelling and Inference Using Likelihood, Clarendon Press; Oxford University Press, Oxford: New York, 2001, p. 528.

[10] S. Callaghan, COVID-19 is a data science issue, Patterns 1 (2) (2020).

[11] N.M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, et al., Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand. Imperial College COVID-19 Response Team, 2020 [Accessed 13th May 2020].

[12] M.J. Keeling, L. Danon, Mathematical modelling of infectious diseases, Br. Med. Bull. 92 (2009) 33–42.

[13] H. Wearing, P. Rohani, M. Keeling, Correction: appropriate models for the management of infectious diseases, PLoS Med. 2 (8) (2005) e320.

[14] M. Tizzoni, P. Bajardi, C. Poletto, J.J. Ramasco, D. Balcan, B. Gonçalves, et al., Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm, BMC Med. 10 (2012) 165.

[15] M.B. Bennett, On the use of the negative binomial in epidemiology, Biom. J. (1981) 69–72.

[16] A.N. Varaksin, V.G. Panov, Linear Regression Models in Epidemiology. Institute of Industrial Ecology, the Urals Branch of the Russian Academy of Sciences.

[17] E. Amene, L.A. Hanson, E.A. Zahn, S.R. Wild, D. Döpfer, Variable selection and regression analysis for the prediction of mortality rates associated with foodborne diseases, Epidemiol. Infect. (2016) 1959–1973. https://doi.org/10.1017/S0950268815003234.

[18] N. Charkha, A. Ghatge, P. Sharma, V.Z. Attar, P. AB, Estimating risk of mortality from cardiovascular diseases using negative, Epidemiol. Open Access 3 (2) (2013). https://doi.org/10.4172/2161-1165.1000127.

[19] X. Zhang, T. Zhang, A.A. Young, X. Li, Applications and comparisons of four time series models in epidemiological surveillance data, PLoS One 9 (2) (2014).

[20] A. Earnest, M.I. Chen, D. Ng, L.Y. Sin, Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, BMC Health Serv. Res. 5 (36) (2005).

[21] F. Petropoulos, S. Makridakis, Forecasting the novel coronavirus COVID-19, PloS One 15 (3) (2020) e0231236, 110.1371/journal.pone.0231236.

[22] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J.M. Hyman, et al., Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020, J. Clin. Med. 9 (596) (2020).

[23] L. Jia, K. Li, Y. Jiang, X. Guo, T. Zhao, Prediction and Analysis of Coronavirus Disease 2019, 2020.

[24] H. Hassanzadeh, A. Nguyen, S. Karimi, K. Chu, Transferability of artificial neural networks for clinical document classification across hospitals: a case study on abnormality detection from radiology reports, J. Biomed. Inf. 85 (2018) 68–79.

[25] J.W. Hardin, J.M. Hilbe, Regression models for count data based on the negative binomial(p) distribution, STATA J. 14 (2) (2014) 280–291.

[26] J.H. Lee, G. Han, W.J. Fulp, A.R. Giuliano, Analysis of overdispersed count data: application to the human papillomavirus infection in men (HIM) study, Epidemiol. Infect. 140 (6) (2012) 1087–1094.

[27] Q. An, J. Wu, X. Fan, L. Pan, W. Sun, Using a negative binomial regression model for early warning at the start of a hand foot mouth disease epidemic in dalian, liaoning province, China, PLoS One 11 (6) (2016) e0157815. https://doi.org/10.1371/journal.pone.0157815.

[28] A.L. Byers, H. Allore, T.M. Gill, P.N. Peduzzi, Application of negative binomial modeling for discrete outcomes: a case study in aging research, J. Clin. Epidemiol. 56 (6) (2003) 559–564.

[29] N.R. Draper, H. Smith, Applied Regression Analysis, vol. 326, John Wiley & Sons, 1998.

[30] J.H. McDonald, Handbook of Biological Statistics, third ed., Sparky House Publishing, Baltimore, Maryland, 2014.

[31] A.A. Beaujean, M.B. Grant, Tutorial on using regression models with count outcomes using R, Practical Assess. Res. Eval. 21 (1) (2016) 2.