Journal of Healthcare Informatics Research (2022) 6:72–90
https://doi.org/10.1007/s41666-021-00105-8

RESEARCH ARTICLE

Analyzing the Impact of Demographic Variables on Spreading and Forecasting COVID-19

Omar Sharif[1] · Md Rafiqul Islam[2] · Md Zobaer Hasan[3] · Muhammad Ashad Kabir[4] · Md Emran Hasan[5] · Salman A. AlQahtani[6] · Guandong Xu[2]

Received: 10 November 2020 / Revised: 13 July 2021 / Accepted: 26 August 2021 /
Published online: 16 September 2021

Abstract

The aim of this study is to analyse the coronavirus disease 2019 (COVID-19) outbreak in Bangladesh. This study investigates the impact of demographic variables on the spread of COVID-19 as well as tries to forecast the COVID-19 infected numbers. First of all, this study uses Fisher's Exact test to investigate the association between the infected groups of COVID-19 and demographical variables. Second, it exploits the ANOVA test to examine significant difference in the mean infected number of COVID-19 cases across the population density, literacy rate, and regions/divisions in Bangladesh. Third, this research predicts the number of infected cases in the epidemic peak region of Bangladesh for the year 2021. As a result, from the Fisher's Exact test, we find a very strong significant association between the population density groups and infected groups of COVID-19. And, from the ANOVA test, we observe a significant difference in the mean infected number of COVID-19 cases across the five different population density groups. Besides, the prediction model shows that the cumulative number of infected cases would be raised to around 500,000 in the most densely region of Bangladesh, Dhaka division.

Keywords Infected cases · COVID-19 · Fisher's Exact test · ANOVA test · Holt's method

This article belongs to the Topical Collection: *Special Collection on COVID-19*
Guest Editors: Christopher C. Yang, Julio Cesar Facelli, David Buckeridge, Fei Wang

✉ Md Rafiqul Islam
rafiqulislam.cse24@gmail.com; MdRafiqul.Islam-1@student.uts.edu.au

Extended author information available on the last page of the article.

1 Introduction

Bangladesh is one of the world's most densely populated country. According to the World Population Review 2020[1], it is placed in eight position of population rank and tenth position of density rank. Currently, the novel COVID-19 is rapidly spreading globally, and most of the cases, the number of infection is high in densest populated country. Due to one of the densest countries in the world, the virus has been spreading rapidly in Bangladesh. According to the Institute of Epidemiology, Disease Control and Research (IEDCR)[2] the first COVID-19 case in Bangladesh was found on March 8, 2020 [25]. Since then the number of infection is increasing rapidly. Therefore, it is imperative to analysis spreading of COVID-19 in Bangladesh and to predict future cases.

Most of the existing studies on spreading and forecasting COVID-19 [5, 10, 12, 20, 27, 36, 38, 40] have been performed solely on the basis of COVID-19 situation. However, there have been only a few research (e.g. [7, 22, 33]) that have studied the effect of population density on the spread of coronavirus infection. Thus, in this study, we aim to investigate the effects of population density, literacy rate, and division on spreading and forecasting COVID-19 infection in Bangladesh. We started our interest by looking through Bangladesh's COVID-19 data with the demographic variables where we used both Fisher's Exact test [29], and ANOVA test [6]. For forecasting analysis, we have used Holt's method [39] because it is a very effective for the trend data with no seasonality [14]. Furthermore, to reduce forecasting error, we applied Unreplicated Linear Functional Relationship (ULFR) model [41] to find best smoothing constant (a parameter used in forecasting process). The prediction value would help the government to take proper preparation to tackle the potential unprecedented situations in Bangladesh. The key contributions of this paper are threefold:

- We investigate the association between COVID-19 and three demographic variables such as population density, region/division and literacy rate through a Fisher's Exact test.
- We further examine the significant difference in the mean infected number of COVID-19 cases across the various combinations of three demographic variables through a ANOVA test.
- We adopt Holt's method to predict the number of infected cases in the epidemic peak region in Bangladesh. To reduce forecasting error, we have utilized a single level of smoothing constant of Holt's method.

The rest of the paper is organized as follows. Section 2 presents related work. Our study methodology is presented in Section 3. After reporting COVID-19 spreading analysis in Sections 4 and 5, we present our spreading results and forecasting discussion in Section 6. Finally, Section 7 concludes the paper and discuss the future work.

[1] https://worldpopulationreview.com/

[2] https://www.iedcr.gov.bd.

2 Related Work

Over the recent few months, a number of research have analysed the impact of various aspects such as climate and demographic variables on the spread of COVID-19 around the world.

There are different variables that have a significant impact on the spread of COVID-19 globally [4]. For instance, Ahmadi et al. [1] used the number of infected people in Iran with COVID-19, population density, average temperature, average precipitation, humidity, wind speed as the main parameters and tried to understand the effects of these parameters on spreading COVID-19 in Iran. Rader et al. [28] found that the accessibility to COVID-19 testing sites in USA is increased with high population density. Wang et al. [38] argued that the effect of climatic factors on spreading of COVID-19 can play an important role in the new COVID-19 outbreak. Therefore, it is clear that factors mentioned above have a significant and direct relationship with the number of infected people. For analysis spreading and forecasting COVID-19, a number of models have been used such as SIDR [9], DASS-21 [2], Fuzzy Clustering [22], and SEIR [26]. A summary of the existing work is presented in Table 1. Two studies [10] and [13] have used mathematical models to predict the number of infected cases. Some others used Genetic Programming and Regression model for short-term prediction [19, 30]. However, due to the small number of data or parameter selection problem, many of those models' outcomes have shown a wide range of dissimilarities. Therefore, selecting parameter value is a key for the model prediction.

While only a few research have studied COVID-19 situation in Bangladesh, there is no research work conducted by using more than one quantitative analysis. Thus, in this paper, we have conducted two quantitative analyses: Fisher's Exact test and ANOVA test. Furthermore, we have predicted the number of infected cases in the epidemic peak region of Bangladesh, Dhaka division.

3 Methodology

Figure 1 shows the methodology used in this study. First, we have collected the dataset. In this study, we have used two types of dataset: (1) daily region/division-wise COVID-19 infected cases in Bangladesh from 5th April to 6th June 2020, and (2) demographic data of Bangladesh. The dataset is freely available to download from the GitHub repository[3]. The daily COVID-19 infected cases for Bangladesh were collected from IEDCR[4], and the demographic data is collected from the Bangladesh Bureau of Statistics (BBS)[5].

After the data collection, we analyse the association between COVID-19 and demographic variables such as population density, literacy rate, and division, followed by the analysis of forecasting COVID-19 infected cases.

[3] https://github.com/rafiqulcse/Analysis-and-Forecasting-of-COVID-19-in-Bangladesh.git

[4] https://www.iedcr.gov.bd.

[5] http://www.bbs.gov.bd/

Table 1 Key studies on spreading and forecasting COVID-19

Reference	Objective	Variable	Country	Approach
Fokas et al. [10]	Compute the impact on the number of deaths.	Cumulative number of deaths	Globally	Computational approach
Fanelli and Piazza [9]	Analyse and forecast of COVID-19 spreading.	Susceptible, Infected, Recovered, Dead, Scheme	China, Italy, France	SIDR
Grasselli et al. [12]	To coordinate the critical care of COVID-19.	Positive, Negative	Italy	Linear model
Mahmoudi et al. [22]	Study the relation between spread of Covid-19 and population size	Population size	USA, Spain, Italy, Germany, UK, France, and Iran	Fuzzy clustering
Shen et al. [33]	Estimate the effective reproduction number of 2019-nCoV and to predict the epidemic peak time	Population size, time, infected and death cases	China	2019-nCov
Hasan and Siddik [13]	Examine the correlation between daily and total COVID-19 case	Temperature, humidity, and wind speed	Bangladesh	Linear association
Paul et al. [26]	Predict the disease burden with special emphasis on India, Bangladesh and Pakistan	Precautionaries such as maintain lockdown, social distancing, using of mask and hand wash	South Asia	SEIR
Petropoulos and Makridakis [27]	Predicting the continuation of the COVID-19	Confirmed cases, deaths and recoveries	Globally	Exponential smoothing models
Roosa et al. [30]	forecasting of the COVID-19 epidemic	daily confirmed case, provinces	China	phenomenological models
Md Hasinur Rahaman Khan and Ahmed Hossain [19]	analysing the COVID-19 outbreak situations, and predicting infections and deaths case.	temporal data of confirmed and death cases	Bangladesh	Infection Trajectory-Pathway Strategy (ITPS)

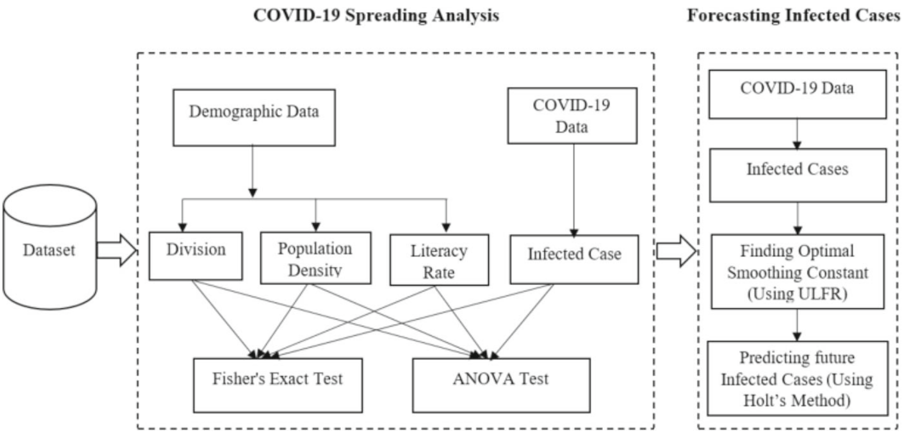
COVID-19 Spreading Analysis
Forecasting Infected Cases
Dataset
Demographic Data
COVID-19 Data
Division
Population Densitv
Literacy Rate
Infected Case
Fisher's Exact Test
ANOVA Test
COVID-19 Data
Infected Cases
Finding Optimal Smoothing Constant (Using ULFR)
Predicting future Infected Cases (Using Holt's Method)

Fig. 1 The study methodology

4 COVID-19 Spreading Analysis

A two-phase study was conducted that consisted of an initial Fisher's Exact test followed by a ANOVA test. In the following subsections, we present the association between the infected groups of COVID-19 and demographic variables which were identified by the Fisher's Exact test. During the ANOVA test, we then present the significant difference in the mean infected number of COVID-19 cases across the divisions, population density, and literacy rate in Bangladesh.

4.1 Fisher's Exact Test

The Fisher's Exact test is considered to investigate whether there is any association between the infected groups of COVID-19 and demographical variables (divisions, literacy rate classes, population density groups) in Bangladesh. The variables and their corresponding categories used in this phase is presented in Table 2. According to Starnes et al. [35], the assumptions of chi-square test [24] is: "No more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater". But in our study the assumptions of chi-square are not satisfied. Therefore, instead of chi-square test, this research uses Fisher's Exact test [29].

4.2 ANOVA Test

The purpose of Analysis of Variance (ANOVA) test [6] was to examine whether there is any significant deference in the mean infected number of COVID-19 cases across eight divisions, five population density groups and two literacy rate classes in Bangladesh. Thus, seven statistical hypotheses were considered to run ANOVA are as follows:

H_{01} : There is no difference in the mean infected number of COVID-19 cases among the eight divisions in Bangladesh.

Table 2 Categories of demographical variables for Fisher's Exact test as well as ANOVA test

Variables	Categories
Division	Dhaka
	Mymensing
	Barisal
	Chittagong
	Khulna
	Rajshahi
	Rangpur
	Sylhet
Population Density	Low ($\leq$ 686 people/sq km)
	Semi-low (687–1211 people/sq km)
	Medium (1212–1935 people/sq km)
	Semi-high (1936–4308 people/sq km)
	High ($\geq$ 4309 people/sq km)
Literacy Rate	Below average = 50.53
	Above average = 50.53
COVID-19 infected cases	Low = (0–25)
	Medium = (26–100)
	Seveare = ($\geq$ 101)

H_{02} : There is no difference in the mean infected number of COVID-19 cases among the five population density groups in Bangladesh.

H_{03} : There is no difference in the mean infected number of COVID-19 cases among the two literacy rate classes in Bangladesh.

H_{04} : There is no interaction effect of divisions and population density on the mean infected number of COVID-19 in Bangladesh.

H_{05} : There is no interaction effect of divisions and literacy rate on the mean infected number of COVID-19 in Bangladesh.

H_{06} : There is no interaction effect of population density and literacy rate on the mean infected number of COVID-19 in Bangladesh.

H_{07} : There is no combine effect of divisions, population density and literacy rate on the mean infected number of COVID-19 in Bangladesh.

4.3 Pairwise Comparison: Tukey Test

From the results of ANOVA test in Table 5, we found that there is a significant difference in the mean affected number of COVID-19 cases across the five different population density groups in Bangladesh. Therefore, we run the post hoc test (Tukey test) to investigate which pairs of the population density groups are different in terms of the mean infected number of COVID-19 cases in Bangladesh. So, we use ten statistical hypotheses to run Tukey test that are given below:

Hypothesis for Pair 1 $H_0 : \mu_{low} = \mu_{semi-low} vs. H_a : \mu_{low} \neq \mu_{semi-low}$

Hypothesis for Pair 2 $H_0 : \mu_{low} = \mu_{medium} vs. H_a : \mu_{low} \neq \mu_{medium}$

Hypothesis for Pair 3 $H_0 : \mu_{low} = \mu_{semi-high} vs. H_a : \mu_{low} \neq \mu_{semi-high}$

Hypothesis for Pair 4 $H_0 : \mu_{low} = \mu_{high} vs. H_a : \mu_{low} \neq \mu_{high}$

Hypothesis for Pair 5 $H_0 : \mu_{semi-low} = \mu_{medium} vs. H_a : \mu_{semi-low} \neq \mu_{medium}$

Hypothesis for Pair 6 $H_0 : \mu_{semi-low} = \mu_{semi-high} vs. H_a : \mu_{semi-low} \neq \mu_{semi-high}$

Hypothesis for Pair 7 $H_0 : \mu_{semi-low} = \mu_{high} vs. H_a : \mu_{semi-low} \neq \mu_{high}$

Hypothesis for Pair 8 $H_0 : \mu_{medium} = \mu_{semi-high} vs. H_a : \mu_{medium} \neq \mu_{semi-high}$

Hypothesis for Pair 9 $H_0 : \mu_{medium} = \mu_{high} vs. H_a : \mu_{medium} \neq \mu_{high}$

Hypothesis for Pair 10 $H_0 : \mu_{semi-high} = \mu_{high} vs. H_a : \mu_{semi-high} \neq \mu_{high}$

5 Forecasting Infected Cases

In the existing study, various models such as Ace-Mod (Australian Census-based Epidemic Model) [8], clustering method [22], SEIR models [26] and others ([9], [34] and [3]) have been employed to forecast future infected cases. However, these models are limited to analyse time series data. Therefore, this study uses Holt's method to predict the number of infected cases in the epidemic peak region. It utilizes a single level of smoothing constants to compare forecasting performance. Therefore, to find the optimal smoothing constant, this analysis uses unreplicated linear functional relationship model as most of the research uses rule of thumb.

5.1 Unreplicated Linear Functional Relationship (ULFR) Model

To measure the functionality between a continuous dependent variable and independent variable, linear regression model has been used. Sometimes, the functionality will become obscure because of random variations accompanying with variables. Therefore, Fuller [11] has figured out, it is unfeasible if it apply an independent variable in all conditions.

$$Y_i = \beta_a + \beta_f X_i \tag{1}$$

The functional model, where both dependent and independent variables are subject to errors. Suppose that Y_i and X_i are unobservable dependent and independent variables respectively which correspond to random variables y_i and x_i that are observed with errors, ϵ_i and δ_i respectively, such that,

$$\begin{aligned} y_i &= Y_i + \epsilon_i \\ x_i &= X_i + \delta_i \text{ where i = 1, 2, 3, ..., n.} \end{aligned} \tag{2}$$

Moreover, the following conditions are assumed

$$\begin{aligned} E(\delta_i) = E(\epsilon_i) = 0, Var(\delta_i) = \sigma_d^2, Var(\epsilon_i) = \sigma_e^2, \\ \forall i Cov(\delta_i, \delta_j) = Cov(\epsilon_i, \epsilon_j) = 0, i \neq j \\ Cov(\delta_i, \epsilon_j) = 0, \forall i, j \end{aligned} \tag{3}$$

Chang et al. [20] termed the model as mentioned in Eq. 1 as the unreplicated linear functional relationship (ULFR) model when there is only the variables X and Y, and where δ_i and ϵ_i are random variables that are mutually independent and normally distributed. The log-likelihood function is given by

$$\begin{aligned} L(\beta_a, \beta_f, \sigma_d^2, \sigma_e^2, X_i) = -nln(2\pi) - \frac{1}{2}n(ln\sigma_d^2 + ln\sigma_e^2) - \sum_{i-1}^{n} \frac{(x_i - X_i)^2}{2\sigma_d^2} \\ - \frac{1}{2\sigma_e^2} \sum_{i=1}^{n} (y_i - \beta_a - \beta_f X_i)^2 \end{aligned} \tag{4}$$

$$\begin{aligned} L(\beta_a, \beta_f, \sigma_d^2, \sigma_e^2, X_i) = -nln(2\pi) - \frac{1}{2}n(ln\sigma_d^2 + ln\sigma_e^2) - \sum_{i-1}^{n} \frac{(x_i - X_i)^2}{2\sigma_d^2} \\ - \frac{1}{2\sigma_e^2} \sum_{i=1}^{n} (y_i - \beta_a - \beta_f X_i)^2. \end{aligned} \tag{5}$$

When the ratio of the error variance is known, that is $\frac{\sigma_e^2}{\sigma_d^2} = \lambda$,

$$\begin{aligned} L(\beta_a, \beta_f, \sigma_d^2, \sigma_e^2, X_i) = -nln(2\pi) - \frac{1}{2}nln\sigma_d^2 - \frac{1}{2}nln\lambda\sigma_d^2 - \sum_{i-1}^{n} \frac{(x_i - X_i)^2}{2\sigma_d^2} \\ - \frac{1}{2\lambda\sigma_d^2} \sum_{i=1}^{n} (y_i - \beta_a - \beta_f X_i)^2 \end{aligned} \tag{6}$$

then the maximum likelihood estimators of parameters $\beta_a, \beta_f, \sigma_d^2, and X_i$ respectively and equate the result to zero:

$$\frac{\delta L}{\delta \beta_a} = - \frac{1}{2\lambda\hat{\sigma}_d^2} \sum_{i=1}^{n} 2(y - \beta_a - \beta_f * X_i)(-1) = 0 \tag{7}$$

$$\frac{\delta L}{\delta \beta_f} = - \frac{1}{2\lambda\hat{\sigma}_d^2} \sum_{i=1}^{n} 2(y - \beta_a - \beta_f * X_i)(-X_i) = 0 \tag{8}$$

$$\frac{\delta L}{\delta X_i} =-\frac{1}{2\hat{\sigma}_d^2}\sum_{i=1}^{n}2(x_i - X_i)(-1)-\frac{1}{2\lambda\hat{\sigma}_d^2}\sum_{1}^{n}2(y - \beta_a-\beta_f * X_i)(-\beta_f) = 0 \quad (9)$$

$$\frac{\delta L}{\delta\sigma_f} = -\frac{n}{\hat{\sigma}_d} - \frac{n}{\hat{\sigma}_d} + \sum_{i-1}^{n}\frac{(x_i - X_i)^2}{\hat{\sigma}_d}+\frac{1}{\lambda\sigma_d}\sum_{1}^{n}(y_i - \beta_a-\beta_f X_i)^2 = 0 \quad (10)$$

After simplification of the equations the maximum likelihood estimators of parameters $\beta_a, \beta_f, \sigma_d^2, X_i$ are as follows:

$$\beta_a = \overline{y} - \beta_f\overline{x} \quad (11)$$

$$\beta_f = \frac{(S_{yy} - \delta S_{xx}) + (((S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2)^{\frac{1}{2}}}{2S_{xy}} \quad (12)$$

$$\sigma_d^2 = \frac{1}{n - 2(\sum(x_i - X_i)^2 + \frac{1}{\lambda}\sum(y_i - \beta_a - \beta_f X_i)^2)} \quad (13)$$

$$X_i = \frac{\delta x_i + \beta_f(Y_i - \beta_a)}{\lambda + \beta_f} \quad (14)$$

where $\overline{y} = \frac{\sum y_i}{n}, \overline{x} = \frac{\sum x_i}{n}, S_{yy} = \sum(y_i - \overline{y})^2, S_{xx} = \sum(x_i - \overline{x})^2, S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$ and Coefficient of determination of ULFR R_f^2 for δ=1

$$R_f^2 = \frac{SS_r}{S_{yy}} \quad (15)$$

$$SS_r = \frac{\beta_f(S_{yy} - S_{xx}) + 2\beta_f S_{xy}}{1 + \beta_f^2} \quad (16)$$

5.2 Holt's Method

Exponential smoothing was suggested in the late 1950s, and has driven few successful predicting approaches [14]. Predicting models has been created using exponential smoothing methods which are weighted averages of former remarks, with the measurement decaying exponentially because the observations get older. Alternatively, Holt's two-parameter model is a popular smoothing model for forecasting data with trend and without seasonality. The model also known as linear exponential smoothing model. To produce an ultimate forecast, Holt's model consists of three distinct equations that work all together. A basic smoothing equation, level equation, is that openly modifies the last smoothed value for the preceding trend. In the second equation, trend equation, is articulated as the difference between the last two smoothed values. To end the third, forecast equation, is used to measure the final forecast. Holt's model applies two smoothing parameters. Two parameters are defined respectively overall smoothing and the trend smoothing. The Holt's method is also named trend-enhanced exponential smoothing or else double exponential smoothing [37]. Therefore, the following equations are as follows:

$$\textbf{Forecast equation: } F_{t,k} = Y_t + K Z_t \tag{17}$$

$$\textbf{Level equation: } Y_t = \alpha X_t + (1 - \alpha)(Y_{t-1} + Z_{t-1}) \tag{18}$$

$$\textbf{Trend equation: } Z_t = \beta(Y_t - Y_{t-1}) + (1 - \beta) Z_{t-1} \tag{19}$$

where Y_t represents an estimation of the level at time t, Z_t symbolizes measurement of the trend which means slope at time t, ($0 \leq \alpha \leq 1$) is the smoothing parameter(SP) for the level and β ($0 \leq \beta \leq 1$) is the SP for the trend. The variable X_t is defined as the period t base level from the current period. Additionally, estimation of the period t base level based on previous data is noted as $(Y_{t-1} + Z_{t-1})$. To measure Z_t, a weighted average of the resulting two measures are taken:

(i) An estimate of trend from the current period given by the increase in the smoothed trend from period (t-1) to period t.
(ii) The notation Z_{t-1}, which is the previous estimate of the trend at time (t-1)

To start Holt's method, a primary estimation (call it Y_0) of the level and an initial estimation (call it Z_0) of the trend are needed. Here, Z_0 equals to the average increase in the time series during the previous year and Y_0 equals to last observation.

6 Result and Discussion

In this section, we present the findings of the spread of COVID-19 followed by the discussion on forecasting results.

6.1 Findings from Spreading Analysis

The result of Fisher's Exact test is reported in Table 3. Based on the results as shown in Table 3, we have analysed the association between demographic variables and the spread of COVID-19 in the case of Bangladesh. Although existing studies such as [21] and [23] provided there is a significant association between divisions and infected groups of COVID-19, in our findings in Table 3, it appears that there is no significant association between divisions and infected groups of COVID-19 (Fisher's Exact test = 18.521, p-value = $0.063 > 0.05$) at 5% level of significance.

Additionally, Table 3 present the results of our investigation that there is no significant association between literacy rate classes and affected groups of COVID-19 (Fisher's Exact test = 0.676, p-value = $0.776 > 0.05$) at 5% level of significance. However, it shows that there is significant association between population density groups and affected groups of COVID-19 (Fisher's Exact test = 14.686, p-value = $0.027 < 0.05$) at 5% level of significance. This finding supports the findings of [1], [28] and [32]. Therefore, to measure the strength of the association between population density groups and infected groups of COVID-19; we use the Cramer's V results

Table 3 Fisher's exact test

Fisher's Exact Test	Test Value	p-value	Cramer's V	p-value
Divisions and infected groups of COVID-19	18.521	0.063	×	×
Literacy rate classes and infected groups of COVID-19	0.676	0.776	×	×
Population density groups and infected groups of COVID-19	14.686*	0.027	0.374*	0.027

Notes: * indicates significant at 5% level of significance in Table 3. From the result of Cramer's V, it can be concluded that the strength of the association between population density groups and affected groups of COVID-19 is very strong (Cramer's V = 0.374) and the strength is significant (p-value = 0.027 $<$ 0.05).

The results of ANOVA test and Turkey test are available in Tables 4 and 5 respectively. From the Table 4, we found that the p-value $< \alpha = 0.05$ for the second hypothesis. Therefore, reject H_{02} at 5% level of significance and conclude that there is sufficient evidence to show a significant difference in the mean affected number of COVID-19 cases across the five different population density groups in Bangladesh. For all other remaining hypotheses, the p-value $> \alpha = 0.05$, therefore do not reject the null hypotheses at 5% level of significance. Thus, it can be concluded that the mean infected number of COVID-19 cases is the same for the eight divisions and two literacy rate classes in Bangladesh. There is insufficient evidence to show a significant interaction between divisions and population density; divisions and literacy rate; population density and literacy rate. In addition, there is no sufficient evidence to show a significant combine effect of divisions, population density and literacy rate.

we found that the p-value $< \alpha = 0.05$ for the pairs 4, 7, 9 and 10. Therefore, reject H_0 at 5% level of significance and conclude that the mean infected number of COVID-19 cases is different across the low and high population density groups, semi-low and high population density groups, medium and high population density groups, semi-high and high population density groups in Bangladesh. For all other remaining pairs, the p-value $> \alpha = 0.05$, therefore, do not reject H_0 at 5% level of significance and conclude that the mean infected number of COVID-19 cases is not different across the remaining pairs of population density groups in Bangladesh.

6.2 Discussion of Forecasting Results

In this section, we will discuss the effectiveness of our proposed model as well as show the forecasting trend of COVID-19 in Bangladesh which is determined based on the demographic data and COVID-19 infected cases from 5th April to 6th June 2020. There are six parameters A = ($\alpha = 0.2$, $\beta = 0.8$), B = ($\alpha = 0.8$, $\beta = 0.2$), C = ($\alpha = 0.5$, $\beta = 0.5$), D = ($\alpha = 0.05$, $\beta = 0.45$), E = ($\alpha = 0.45$, $\beta = 0.05$) and F*= (α

Table 4 ANOVA test

Source	F	p-value
Divisions	.015	1.000
Population Density	7.962*	.000
Literacy Rate	.077	.783
Divisions * Population Density	.015	1.000
Divisions * Literacy Rate	.005	1.000
Population Density * Literacy Rate	.064	.938
Divisions * Population Density * Literacy Rate	.001	.977

Notes: * indicates significant at 5% level of significance

Table 5 Post hoc test — Tukey test

Population density	Population density	Mean Difference (I-J)	Std. Error	Sig.
Low	Semi-low	−126.90	891.797	1.000
≤ 686 people/sq km	Medium	−800.86	1220.361	.964
	Semi-high	−617.36	2018.892	.998
	High	−13751.36*	2018.892	.000
Semi-low	Low	126.90	891.797	1.000
687–1211 people/sq km	Medium	−673.96	1015.112	.963
	Semi-high	−490.46	1901.867	.999
	High	−13624.46*	1901.867	.000
Medium	Low	800.86	1220.361	.964
1212–1935 people/sq km	Semi-low	673.96	1015.112	.963
	Semi-high	183.50	2076.313	1.000
	High	−12950.50*	2076.313	.000
Semi-high	Low	617.36	2018.892	.998
1936–4308 people/sq km	Semi-low	490.46	1901.867	.999
	Semi-high	−183.50	2076.313	1.000
	High	−13134.00*	2626.351	.000
High	Low	13751.36*	2018.892	.000
≥ 4309 people/sq km	Semi-low	13624.46*	1901.867	.000
	Medium	12950.50*	2076.313	.000
	Semi-high	13134.00*	2626.351	.000

Notes: * indicates significant at 5% level of significance

= 0.76, β = 0.24) in our experiments where F* is our proposed smoothing constant parameter. Therefore, the Holt's method is employed to forecast over the post-sample period from June 2020 to December 2021.

In our study, we choose Dhaka as the forecasting division because this division is among the highest in the total number of confirmed COVID-19 cases, and highly populated in Bangladesh. In Table 6, we provide a detailed information about the different smoothing constant (SC) value, and the forecasted value of the number of COVID-19 infected people in Dhaka at the end of 2021. The table is also containing mean absolute percentage error (MAPE) value for each SC value. The forecasted values for different SC values show that the number of COVID-19 infected people will be around 0.5 million in Dhaka at end-of-year 2021. Table 6 also presents mean absolute percentage error, original cumulative number and the forecasted cumulative number of COVID-19 infected people for our proposed smoothing constant parameter F*. Prediction value is closer for the SC value of F*. The MAPE value shows 13.6% for F* whereas for other different parameters the MAPE values are arround 24%. Therefore, the smoothing constant values for F* ($\alpha = 0.76$, $\beta = 0.24$), which is determined our ULFR provide the best results for forecasting.

Table 6 Forecasting number of infected people for different parametric value

Smoothing constant	Date	Original Cumulative No.	Forecasted Value	MAPE (%)	Date	Forecasted value
A=($\alpha = 0.2$, $\beta = 0.8$)	6/6/2020	28273	29820	23.29	31/12/2021	683131
B=($\alpha = 0.2$, $\beta = 0.8$)		28273	29145	23.29		498575
C=($\alpha = 0.5$, $\beta = 0.5$)		28273	29392	23.29		511045
D=($\alpha = 0.05$, $\beta = 0.45$)		28273	29120	24.98		559815
E=($\alpha = 0.45$, $\beta = 0.05$)		28273	28871	23.29		394233
F*=($\alpha = 0.76$, $\beta = 0.24$)		28273	28799	13.60		517093

Note: * indicates our proposed parameter

In Fig. 2, we compared the daily short-term forecasts of cumulative case counts. Table 6 provides the forecasting values 683,131 and 394,233 for the SC A, E respectively and the predicted values are 498,575, 511,045 and 559,815 for the smoothing constant B, C, and D, respectively. For the SC F^*, forecasting value at the end of December 2021 is 517,093. We also observe from Table 6 that F^* is closer to the

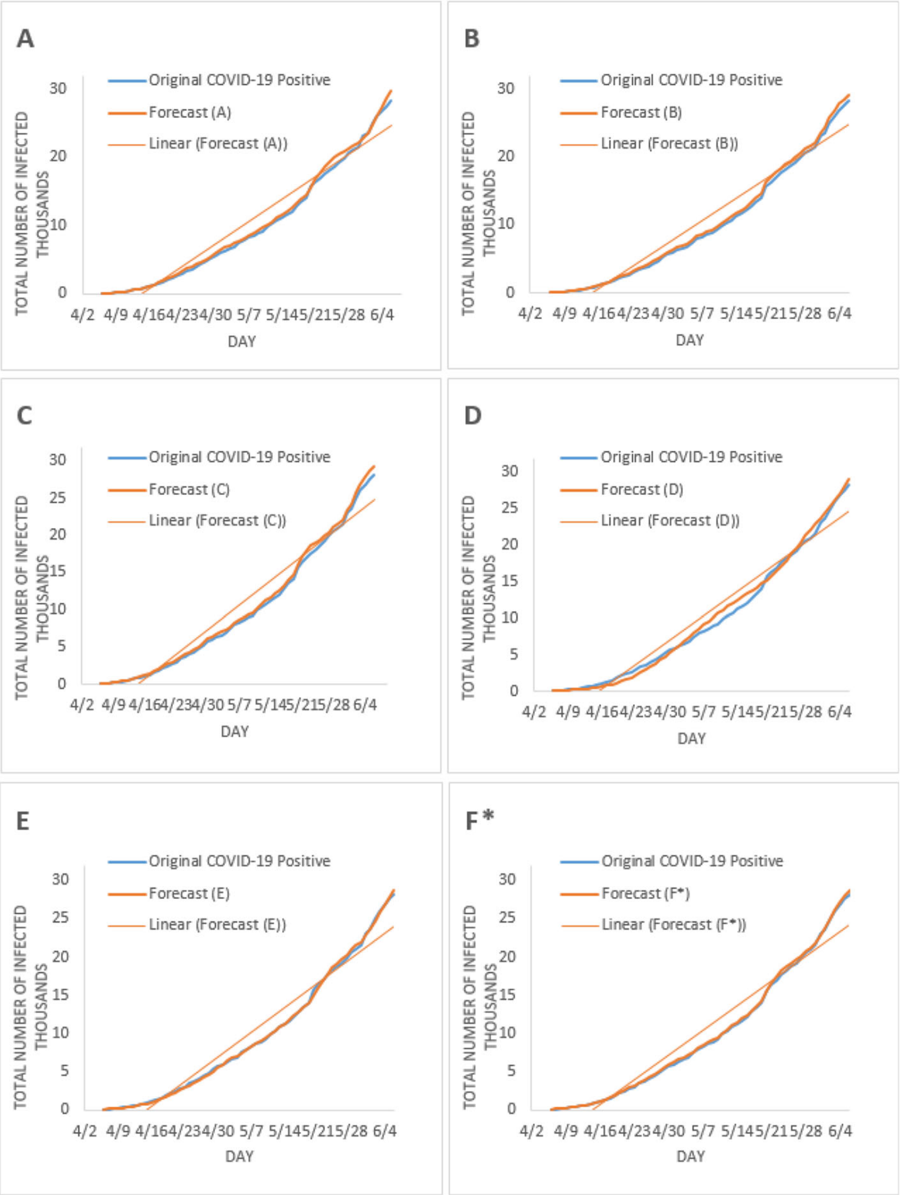

Fig. 2 Day wise forecasting number and original number of infected people

forecast values of B, C, and D but the short-term accuracy level is higher than all other forecasting constant values. According to Rohit et al. [31], COVID-19 is now increasing globally at a rate of 3% to 5% daily. Therefore, the result in this paper showed that the total number of infected cases will be around half a million the Dhaka division at the end of 2021.

Overall, this study will help people to prepare and plan for better regular life activities as the study included consciousness and understanding of various variables that can accelerate or decline the rate of COVID-19. While the study needs more data to make better explicit predictions, our model could help to forecast Confirmed COVID-19 cases if the spread of the virus does not change dramatically (means beyond explanation). Therefore, implementation of advanced machine learning, deep learning algorithms and visualisation techniques could help to forecast and visualise confirmed COVID-19 cases accurately [15–18]. However, to the best of our knowledge, the proposed model is highly effective at the time of writing this paper.

7 Conclusion and Future Work

This research investigated the association between COVID-19 and demographic variables in Bangladesh. It made several contributions to the literature. First, this study uses the Fisher's Exact test to investigate whether there is an association between the infected groups of COVID-19 and demographical variables such as divisions, literacy rate classes, and population density in Bangladesh. Second, it also uses the ANOVA test to examine whether there is any significant difference in the mean infected number of COVID-19 cases across the divisions, literacy rate, and population density in Bangladesh. Third, using Holt's method, this research forecasts the number of infected cases in the epidemic peak region, Dhaka division by the end of the year 2021. Our result shows that there is a significant association between population density groups and infected groups of COVID-19 in Bangladesh as well as the strength of the association is very strong and it is statistically significant. ANOVA test indicates a statistically significant difference in the mean infected number of COVID-19 cases across the five different population density groups in Bangladesh. Finally, the post hoc test, Tukey test finds that the high population density group shows a significant difference in the mean infected number of COVID-19 cases.

In summary, our most recent forecasts, based on the last two months data (5th April to 6th June 2020), remained relatively stable. The proposed models predict that the epidemic has not reached its peak in Dhaka division, yet it would do so on December 2021. This likely shows the impact of the population density on spreading the virus. Educated people's awareness does not impact on reducing the spread of the virus in its peak level in Bangladesh. The forecasts presented are based on the assumption that current mitigation efforts will continue. However, during our research, we encountered several limitations such as availability of required dataset, implementation of some other more accurate machine learning algorithms such as logistic regression and deep learning. The result will be more accurate with a huge dataset where all the COVID-19 patients information is confirmed. In future work, we will explore a machine learning technique to investigate the association and to predict the number of infected cases in the epidemic peak region in Bangladesh.

References

1. Ahmadi M, Sharifi A, Dorosti S, Ghoushchi SJ, Ghanbari N (2020) Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci Total Environ:138705
2. Ahmed O, Ahmed MZ, Alim SMAHM, Khan MAU, Jobe MC (2020) Covid-19 outbreak in Bangladesh and associated psychological problems: An online survey. Death Stud:1–10
3. Chen Y, Chu CW, Chen MI, Cook AR (2018) The utility of lasso-based models for real time forecasts of endemic infectious diseases: a cross country comparison. J Biomed Inf 81:16–30
4. Chien L-C, Chen L-W (2020) Meteorological impacts on the incidence of covid-19 in the us. Stoch Env Res Risk A 34(10):1675–1680
5. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Piontti AP, Mu K, Rossi L, Sun K et al (2020) The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. Science 368(6489):395–400
6. Cuevas A, Febrero M, Fraiman R (2004) An anova test for functional data. Comput Stat Data Anal 47(1):111–122
7. Das M, Das A, Sarkar R, Saha S, Mandal A (2020) Examining the impact of lockdown (due to covid-19) on ambient aerosols (pm 2.5): a study on indo-gangetic plain (igp) cities, india. Stoch Env Res Risk A:1–17
8. Fair KM, Zachreson C, Prokopenko M (2019) Creating a surrogate commuter network from australian bureau of statistics census data. Sci Data 6(1):1–14
9. Fanelli D, Piazza F (2020) Analysis and forecast of covid-19 spreading in china, Italy and france. Chaos, Solitons Fractals 134:109761
10. Fokas A, Cuevas-Maraver J, Kevrekidis P (2020) A quantitative framework for exploring exit strategies from the covid-19 lockdown. arXiv:2005.13698
11. Fuller W (1987) Error measurement models
12. Grasselli G, Pesenti A, Cecconi M (2020) Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response. Jama 323(16):1545–1546
13. Hasan NA, Siddik MS (2020) Possible role of meteorological variables in covid-19 spread: A case study from a subtropical monsoon country, bangladesh
14. Holt C (2004) Forecasting seasonals and trends by exponentially weighted averages. onr memorandum 52/1957. Carnegie Institute of Technology
15. Islam MR, Akter S, Ratan MR, Kamal ARM, Xu G (2021a) Deep visual analytics (dva): applications, challenges and future directions. Hum-Centric Intell Syst 1(1-2):3–17
16. Islam MR, Liu S, Biddle R, Razzak I, Wang X, Tilocca P, Xu G (2021b) Discovering dynamic adverse behavior of policyholders in the life insurance industry. Technol Forecast Soc Chang 163:120486
17. Islam MR, Liu S, Wang X, Xu G (2020) Deep learning for misinformation detection on online social networks: a survey and new perspectives. Soc Netw Anal Min 10(1):1–20
18. Islam MR, Miah SJ, Kamal ARM, Burmeister O et al (2019) A design construct of developing approaches to measure mental health conditions. Aust J Inf syst:23
19. Khan MHR, Hossain A (2020) Covid-19 outbreak situations in bangladesh: An empirical analysis. medRxiv
20. Kucharski AJ, Kwok KO, Wei VW, Cowling BJ, Read JM, Lessler J, Cummings DA, Riley S (2014) The contribution of social behaviour to the transmission of influenza a in a human population. PLoS Pathog 10(6):e1004206
21. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD et al (2020) Early dynamics of transmission and control of covid-19: a mathematical modelling study. The lancet infectious diseases
22. Mahmoudi MR, Baleanu D, Mansor Z, Tuan BA, Pho K.-H. (2020) Fuzzy clustering method to compare the spread rate of covid-19 in the high risks countries. Chaos, Solitons Fractals:110230
23. Mastrandrea R, Fournet J, Barrat A (2015) Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. PloS one 10(9):e0136497

24. McHugh ML (2013) The chi-square test of independence. Biochem Med: Biochem Med 23(2):143–149
25. Orubu ESF, Zaman MH, Rahman MT, Wirtz VJ (2020) Veterinary antimicrobial resistance containment in bangladesh: Evaluating the national action plan and scoping the evidence on implementation. J Glob Antimicrobial Resist 21:105–115
26. Paul A, Chatterjee S, Bairagi N (2020) Prediction on covid-19 epidemic for different countries: Focusing on south asia under various precautionary measures. medRxiv
27. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus covid-19. PloS one 15(3):e0231236
28. Rader B, Astley CM, Sy KTL, Sewalk K, Hswen Y, Brownstein JS, Kraemer MU (2020) Increased travel times to united states sars-cov-2 testing sites: a spatial modeling study. medRxiv
29. Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49(6):1280–1283
30. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman J, Yan P, Chowell G (2020) Real-time forecasts of the covid-19 epidemic in China from february 5th to february 24th, 2020. Infect Disease Modell 5:256–263
31. Salgotra R, Gandomi M, Gandomi AH (2020) Time series analysis and forecast of the covid-19 pandemic in India using genetic programming. Chaos, Solitons & Fractals:109945
32. Sannigrahi S, Pilla F, Basu B, Basu AS (2020) The overall mortality caused by covid-19 in the european region is highly associated with demographic composition: A spatial regression-based approach. arXiv preprint arXiv:2005.04029
33. Shen M, Peng Z, Xiao Y, Zhang L (2020) Modelling the epidemic trend of the 2019 novel coronavirus outbreak in china. BioRxiv
34. Shinde GR, Kalamkar AB, Mahalle PN, Dey N, Chaki J, Hassanien AE (2020) Forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. SN Comput Sci 1(4):1–15
35. Starnes DS, Yates D, Moore DS (2010) The practice of statistics. Macmillan
36. Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for covid-19 pandemic in india. Stoch Env Res Risk A:1
37. Swamidass PM (2000) Encyclopedia of production and manufacturing management. Springer Science & Business Media
38. Wang J, Tang K, Feng K, Lv W (2020) High temperature and high humidity reduce the transmission of covid-19. Available SSRN 3551767
39. Wright DJ (1986) Forecasting data published at irregular time intervals using an extension of holt's method. Manag Sci 32(4):499–510
40. You C, Deng Y, Hu W, Sun J, Lin Q, Zhou F, Pang CH, Zhang Y, Chen Z, Zhou X-H (2020) Estimation of the time-varying reproduction number of covid-19 outbreak in china. Int J Hygiene Environ Health:113555
41. Yun C, Bakar SARA (2010) Multidimensional unreplicated linear functional relationship model with single slope and its coefficient of determination

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Omar Sharif[1] · Md Rafiqul Islam[2] · Md Zobaer Hasan[3] · Muhammad Ashad Kabir[4] · Md Emran Hasan[5] · Salman A. AlQahtani[6] · Guandong Xu[2]

Omar Sharif
omarsharif.ged@diu.edu.bd

Md Zobaer Hasan
MdZobaer.Hasan@monash.edu

Muhammad Ashad Kabir
akabir@csu.edu.au

Md Emran Hasan
writetoemran@gmail.com

Salman A. AlQahtani
salmanq@ksa.edu.sa

Guandong Xu
Guandong.Xu@uts.edu.au

[1] Daffodil International University, Dhaka, Bangladesh

[2] Advanced Analytics Institute (AAi), University of Technology Sydney (UTS), Ultimo, Australia

[3] School of Science, Monash University Malaysia, Subang Jaya, Selangor D. E., Malaysia

[4] School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW, Australia

[5] City University, Dhaka, Bangladesh

[6] College of Computer and Information Sciences, King Saud University, Riyadh,
Kingdom of Saudi Arabia