Regressive Class Modelling for Predicting Trajectories of COVID-19 Fatalities Using Statistical and Machine Learning Models

Rafiqul I. Chowdhury¹ · M. Tariqul Hasan² · Gary Sneddon³

Received: 10 January 2022 / Revised: 17 March 2022 / Accepted: 20 March 2022
© The Author(s), under exclusive licence to Malaysian Mathematical Sciences Society and Penerbit Universiti Sains Malaysia 2022

Abstract
The COVID-19 (SARS-CoV-2 virus) pandemic has led to a substantial loss of human life worldwide by providing an unparalleled challenge to the public health system. The economic, psychological, and social disarray generated by the COVID-19 pandemic is devastating. Public health experts and epidemiologists worldwide are struggling to formulate policies on how to control this pandemic as there is no effective vaccine or treatment available which provide long-term immunity against different variants of COVID-19 and to eradicate this virus completely. As the new cases and fatalities are recorded daily or weekly, the responses are likely to be repeated or longitudinally correlated. Thus, studying the impact of available covariates and new cases on deaths from COVID-19 repeatedly would provide significant insights into this pandemic’s dynamics. For a better understanding of the dynamics of spread, in this paper, we study the impact of various risk factors on the new cases and deaths over time. To do that, we propose a marginal-conditional based joint modelling approach to predict trajectories, which is crucial to the health policy planners for taking necessary measures. The conditional model is a natural choice to study the underlying property of dependence in

Communicated by Shahariar Huda.

¹ M. Tariqul Hasan thasan@unb.ca
Rafiqul I. Chowdhury rachowdhury@upei.ca

² Department of Mathematics and Statistics, University of New Brunswick, Fredericton, NB E3B 5A3, Canada

³ Department of Mathematics and Statistics, Mount Saint Vincent University, Halifax, NS B3M 2J6, Canada

Published online: 13 April 2022
consecutive new cases and deaths. Using this model, one can examine the relationship between outcomes and predictors, and it is possible to calculate risks of the sequence of events repeatedly. The advantage of repeated measures is that one can see how individual responses change over time. The predictive accuracy of the proposed model is also compared with various machine learning techniques. The machine learning algorithms used in this paper are extended to accommodate repeated responses. The performance of the proposed model is illustrated using COVID-19 data collected from the Texas Health and Human Services.

Keywords  SARS-CoV-2 virus · Repeated measures · Model accuracy · Deep learning techniques · Joint modelling

Mathematics Subject Classification  62H20 · 62F03 · 62F10 · 62N03

1 Introduction

Countries around the world are overwhelmed by the outbreak of the COVID-19 pandemic, as the SARS-CoV-2 virus (COVID-19) has been transmitted across the globe at an alarming rate. As of December 31, 2021, according to World health organization (WHO) [1], a total confirmed new cases and deaths are over 281 millions and 5.4 millions, respectively. Due to its astonishing nature of transmission, public health experts and epidemiologists are struggling to formulate policies on how to control and end this pandemic. Health policy planners from all the countries are trying to flatten the curves for both the new cases and deaths of COVID-19. However, public health and government officials are witnessing a recent new spike of cases in countries where the spread of the COVID-19 was once assumed to be under control. Thus, the initial flattened curves of COVID-19 do not mean the crisis is over. Health officials introduced various restrictions on people’s mobility and socialization to reduce the spread and hence death. Consequently, the socio-economic, psychological, and other impacts of COVID-19 on people have already started to become visible continuously [1, 2]. However, as there has been more than one wave of the virus already, there could be another increase or spike in cases after the waves of infections stabilized. Besides, this pandemic curve can have multiple severe surges or spikes in new cases before it stabilizes with more effective doses of vaccine. Lifting of existing restrictions to control this pandemic too quickly could send countries around the world "back into confinement" once a consecutive wave hits.

To identify the COVID-19 transmission patterns, spread and development, various statistical modelling techniques were employed by health researchers around the world since the beginning of 2020 [3–9]. For example, to predict COVID-19-related new cases and deaths, autoregressive integrated moving average (ARIMA)-based modelling techniques were incorporated by researchers in the literature ([3–7]). Shastri et al [8] proposed recurrent neural network (RNN)-based deep learning techniques to forecast COVID-19 new cases and death. To predict the dynamic deportment of COVID-19, Da Silva et. al. [9] have exploited artificial intelligence (AI)-based modelling techniques such as Bayesian regression neural network, cubist regression,
k-nearest neighbours, quantile random forest, and support vector regression. To predict the transmission of COVID-19, machine learning algorithm has also been employed in the literature [10].

Our research is motivated by COVID-19 data collected from the Texas Health and Human Services website (https://dshs.texas.gov/coronavirus/additionaldata.aspx). In this paper, we have considered the daily new cases and deaths for 254 counties in Texas between March 4, 2020, and July 25, 2020, along with the population size for each county. As some counties have no daily deaths for several days, we summed weekly deaths and new cases for our analysis. Thus, for modelling, we have used any death or no death during a week for a total of 20 consecutive weeks. As COVID-19 is causing excess fatalities all over the world, predicting fatality trajectories has gained compelling interest among public health researchers. In this paper, we are interested in predicting the trajectory risks of weekly death and the impact of available risk factors on the trajectory. As the repeated weekly deaths in a specific county are likely to be correlated, we propose a joint modelling-based regressive class approach to estimate the trajectory risks [11, 12].

Studying the impact of available covariates and new cases on deaths from COVID-19 longitudinally would provide significant insights into this pandemic’s dynamics. To better understand the dynamics of spread, we need to study the impact of risk factors and new cases on deaths over time, first conditionally and then jointly, which is crucial to the health policy planners for taking necessary measures. The conditional model is a natural choice to study the underlying property of dependence in consecutive weeks [13]. Using this model, one can examine the relationship between outcomes and predictors, and it is possible to calculate risks of the sequence of events longitudinally [14]. Also, we must understand how the transitions between weeks occur and how the covariates influence these transitions for better prediction. The advantage of repeated measures is that one can see how individual responses change over time [15]. Lindsey and Lamber [15] suggested that this must generally be conditioned on the previous history of a subject. Lee and Nelder [16] concluded that the conditional models are of fundamental interest, and marginal predictions can be made from conditional models. Sometimes the transition probabilities may depend on different models, each model representing a transition from one stage to another. Lindsey and Lamber [15] examined some important theoretical aspects concerning the use of marginal models (e.g. GEE) and demonstrated various limitations. Another alternative is the subject-specific models considering the random effects by allowing random effect terms in the linear predictor [17]. However, the proposed method provides a more general and flexible set up for addressing the risk prediction of repeated categorical outcomes emerging from longitudinal studies. The regressive models at the subsequent follow-ups provide the estimates of the parameters of the conditional models.

We organize the rest of the paper as follows. In Sect. 2, we present proposed joint modelling approach for the repeated measurements to predict fatality trajectories along with various machine learning techniques. Estimation techniques of the model parameters are presented in Sect. 3. The results of data analysis are presented in Sect. 4. Finally, in Sect. 5, some concluding remarks are presented.
2 Predicting Fatality Trajectories for COVID-19

In this section, we present a proposed marginal-conditional based joint modelling approach for predicting trajectories of repeated COVID-19 deaths. In addition, we also incorporate three machine learning algorithms neural network (NN), support vector machine (SVM) and random forest (RF), and compared their performance with the proposed method. The chosen machine learning algorithms have widespread use in various fields. However, these machine learning algorithms are developed for the classification of categorical outcomes measured cross-sectionally or data collected from a single time point. These models also allow the prediction of the probability of the outcome categories. But the problem at our hands is longitudinal, and we are interested in predicting trajectory risk. In other words, we want to predict the joint probability of a sequence of events that occurs longitudinally. Besides, we need to assess the covariate impact on the series of events. We extended these machine learning algorithms using the proposed marginal conditional approach to obtain the marginal and conditional probability and, ultimately, the joint probability to predict the trajectory risks. In other words, we extend these machine learning models for repeated measures data to predict the trajectory risks based on the covariates. First we present proposed joint modelling approach to predict fatalities trajectories for COVID-19 in Sect. 2.1. We then present various machine learning algorithms in Sect. 2.2.

2.1 Joint Modelling Approach

Let $Y_{i1}, Y_{i2}, ..., Y_{ij}$ represent the occurrence of COVID-19-related death in the $i$-th county of Texas, USA at $j$-th week ($i = 1, 2, ..., n$ and $j = 1, 2, ..., J_i$). Note that the data set we used in this paper has $n = 254$ and $J_i = J = 20$. We also assume $Y_{ij} = s$ follows a binomial distribution where ($s = 0, 1$) with the category 0 denotes no COVID-19 related death and 1 denotes COVID-19 related death during a week for the $i$th county in Texas, USA. To be specific, in Fig. 1, we present the trajectory path using three consecutive weeks for the $i$th county.
The marginal trajectory probability of the response at the first week $Y_{i1}$ for the $i$th county can be calculated using the following formula:

$$P_{is.y1}(z_{i1}) = P(Y_{i1} = s \mid z_{i1}) = \frac{e^{(z'_{ij} \beta_{y1})s}}{1 + e^{(z'_{ij} \beta_{y1})}}, \quad s, = 0, 1. \quad (1)$$

where $\beta_{y1} = [\beta_0, \beta_1, ..., \beta_p]'$ is a $1 \times (p + 1)$ regression coefficients corresponding to the covariate vector $z = [1, z_1, ..., z_p]' = [1, x_1, ..., x_p]'$. Similarly, the conditional trajectory probability of the response at the second week $Y_{i2}$ given the first week $Y_{i1}$ can be calculated using the first-order regressive logistic models as:

$$P_{is.y2|y1}(z_{i2}) = P(Y_{i2} = s \mid Y_{i1} = y_{i1}, z_{ij}) = \frac{e^{(z'_{ij} \beta_{y2,y1})s}}{1 + e^{(z'_{ij} \beta_{y2,y1})}}, \quad s, y_{i1} = 0, 1, \quad (2)$$

where $\beta_{y2,y1} = [\beta_{y2,0}, \beta_{y2,1}, ..., \beta_{y2, p}, \beta_{y2,y1}]'$ is $1 \times (p + 2)$ vector of regression coefficients corresponding to the covariate vector $z = [1, z_1, ..., z_p, z_{p+1}]' = [1, x_1, ..., x_p, y_{1}]'$. Thus, the general form of the first-order regressive logistic models for $j$ consecutive outcomes $Y_{i1}, \ldots, Y_{ij}$ can be shown to be:

$$P_{is.yj|yj-1}(z_{ij}) = P(Y_{ij} = s \mid Y_{i(j-1)} = y_{i(j-1)}, z_{ij}) = \frac{e^{(z'_{ij} \beta_{yj,yj-1})s}}{1 + e^{(z'_{ij} \beta_{yj,yj-1})}}, \quad s, y_{ij-1} = 0, 1, \quad (3)$$

where $\beta_{yj,yj-1} = [\beta_{yj,0}, \beta_{yj,1}, ..., \beta_{yj, p}, \beta_{yj,yj-1}]'$ is a $1 \times (p + 2)$ vector of regression coefficients corresponding to the covariate vector $z = [1, z_1, ..., z_p, z_{p+1}]' = [1, x_1, ..., x_p, y_{j-1}]'$.

Consequently, the joint probability mass function of $Y_{i1}, Y_{i2}, \ldots, Y_{ij}$ with covariate vector $X = x$ can be expressed as:

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{ij} = y_{ij} \mid X = x) = P(Y_{i1} = y_{i1} \mid X = x) \times P(Y_{i2} = y_{i2} \mid Y_{i1} = y_{i1}; X = x) \times \cdots \times P(Y_{ij} = y_{ij} \mid Y_{i(j-1)} = y_{i(j-1)}; X = x) \quad (4)$$

$$= P_{is.y1}(x) \times P_{is.y2|y1}(x) \times \cdots \times P_{is.yj|yj-1}(x), \quad s, y_{i1}, ..., y_{ij-1} = 0, 1,$$

where $X' = [1, x_1, ..., x_p]$ is vector of covariates. It should be noted that $X = x$ can be time dependent.
Now using the fitted marginal and conditional models, we can estimate the trajectory risks as follows:

$$\hat{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iJ} = y_{iJ} | X = x) = \hat{P}(Y_{i1} = y_{i1} | X = x) \times \hat{P}(Y_{i2} = y_{i2} | Y_{i1} = y_{i1}; X = x) \times \ldots \times \hat{P}(Y_{iJ} = y_{iJ} | Y_{i(j-1)} = y_{i(j-1)}; X = x) = \hat{P}_{is}y_{i1}(x) \times \hat{P}_{is,y_{i2}|y_{i1}}(x) \times \ldots \times \hat{P}_{is,y_{iJ}|y_{i(J-1)}}(x),$$

$$s, y_{i1}, ..., y_{ij-1} = 0, 1,$$

In the next subsection, we discuss various machine learning techniques to predict fatalities of COVID-19.

### 2.2 Machine Learning Approaches

In this section, we present three machine learning algorithms: neural network (NN), support vector machine (SVM) and random forest (RF), which will be used to predict trajectories. Although there are other machine learning approaches available, we considered these three commonly used techniques for illustration and comparisons. It may be noted that these machine learning algorithms are developed for cross-sectional data or data from a single time point. However, using the proposed approach, we extended these algorithms for trajectory risk prediction for longitudinal events.

#### 2.2.1 Neural Networks (NN)

Neural networks are often used as a building block for deep learning due to the excellent predictive performance. The NN algorithm is tuned by constructing many layers and neurons, then trimming off the unnecessary neurons and layers using regularization through cross-validation. A neural network does not assess associations between predictors and the response as its objective is to predict outcome class. Also, multicollinearity does not create many problems in the neural network. More details of the theory of a neural network can be found in Ripley [18].

#### 2.2.2 Support Vector Machine (SVM)

A popular classifier, support vector machine (SVM), was originally introduced by Boser et. al. [19]. It has been updated by Cortes and Vapnik [20]. The main idea is to separate the space of the feature variables using hyperplanes so that the response classes become as distinct as possible. Also, SVM is not affected by the presence of multicollinearity. This model often provides excellent predictive performance and becomes intuitive, visualizing the support vectors in the lower dimension. This algorithm is computationally demanding, and much attention is required to tune the model parameters.
2.2.3 Random Forests (RF)

A random forest is an aggregated collection of classifiers of classification trees (an ensemble). This prediction model is compelling and provides one of the top prediction performances. The strength of the ensemble depends on the depths of the constituent classification trees and diversity between the trees. The RF needs to tuned using the number of trees to grow in a forest. The random selection of independent variables keeps the correlation between predictors to the minimal. When the objective is to make better predictions, multicollinearity does not affect the results. However, the algorithm may appear computationally demanding for high-dimensional data.

3 Estimation of Model Parameters

In this section, we develop estimation techniques for estimating model parameters to predict trajectories using the proposed approach. To do that, we obtain the log-likelihood function of the joint mass function in (4) which can be expressed as:

\[
\begin{align*}
\ell &= \sum_{i=1}^{n} \sum_{j=1}^{J} \left[ \ln P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{ij} = y_{ij} | x) + \ln P(Y_{i2} = y_{i2} | y_{i1}; x) + \cdots + \ln P(Y_{ij} = y_{ij} | y_{i(j-1)}; x) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J} \left[ \ln P_{iS,y_{i1}}(x_{ij}) + \ln P_{iS,y_{i2}}(y_{i1}) + \cdots + \ln P_{iS,Y_{ij}|y_{i(j-1)}}(x) \right] \\
&= n \sum_{i=1}^{n} \sum_{j=1}^{J} \left[ \ln P_{iS,y_{ij}}(x_{ij}) \right] \\
&\quad + \ln \left[ \prod_{i=1}^{n} P_{iS,Y_{ij}|y_{i(j-1)}}(x) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J} \left[ \ln P_{iS,y_{ij}}(x_{ij}) \right]
\end{align*}
\]  

(6)

From Eq. (6), it is clear that the joint mass function boils down to univariate mass function for any first-order regressive model. In other words, all the first-order regressive models are equivalent to that of the marginal model and follow the same estimation procedure. Then, the likelihood function of the first-order regressive model for sample of \( n \) independent observations can be written as:

\[
\ln \left[ P_{iS,y_{ij}|y_{i(j-1)}}(z) \right] = \sum_{i=1}^{n} \left[ Y_{is,y_{j}} z_{ij}^{\prime} \beta_{y_{j},y_{j-1}} - \log \left[ 1 + e^{(z_{ij}^{\prime} \beta_{y_{j-1},y_{j-1}})} \right] \right]
\]

where \( z = [1, z_1, z_2, \ldots, z_z, y_{(j-1)}]' = [1, x_1, x_2, \ldots, x_p, y_{(j-1)}]' \). Differentiating the log-likelihood with respect to the parameters and equating the derivatives to zero, we obtain the equations whose solutions give the following maximum likelihood estimates for \((p + 2)\) parameters.

\[
\frac{\partial \ln \left[ P_{iS,y_{ij}|y_{i(j-1)}}(z) \right]}{\partial \beta_{jq}} = \sum_{i=1}^{n} \left[ Y_{is,y_{j}} - P_{iS,y_{ij}|y_{i(j-1)}}(z) \right] z_{iq}, \quad q = 0, 1, \ldots, p, (p + 1).
\]
Thus, the observed information matrix can be obtained using the second derivatives as follows:

\[
\frac{\partial^2 \ln \left[ P_{is,yij|y_{i(j-1)}}(z) \right]}{\partial \beta_{jq} \partial \beta_{jq}'} = \sum_{i=1}^{n} \left[ P_{is,yij|y_{i(j-1)}}(z) (1 - P_{is,yij|y_{i(j-1)}}(z)) \right] z_q' z_{q'},
\]

and

\[
\frac{\partial^2 \ln \left[ P_{is,yij|y_{i(j-1)}}(z) \right]}{\partial \beta_{jq} \partial \beta_{jq}'} = \sum_{i=1}^{n} \left[ P_{is,yij|y_{i(j-1)}}(z) P_{is,yij|y_{i(j-1)}}(z) \right] z_q' z_{q'},
\]

where \( q, q' = 0, 1, ..., p, (p + 1) \). The observed information matrix \( I(\beta) \) is the \((p + 2) \times (p + 2)\) matrix where elements are the negative of the second derivatives. It can be shown that the asymptotic covariance matrix is \( \left[ I(\beta) \right]^{-1} \). Consequently, we can use a Newton–Raphson iterative logarithm to obtain the estimated regressive logistic coefficients. To do that, we first arrange the data into a new structure. Then, all these models can be estimated in a usual manner with an appropriate data structure and using Python, R, or other software capable of fitting logistic regression. Then, we can exploit the parallel programming to analyse all the subsets of data using multiple cores in a single computer or using several CPU in a distributed system.

### 3.1 Significance of the Joint Model

The significance of the joint model based on Eq. (4) can be tested using a likelihood ratio test between the joint constant only model (reduced) and joint full model (full) as follows:

\[
-2 \left[ \ln L_{Reduced}(\hat{\beta}_{Reduced}) - \ln L_{Full}(\hat{\beta}_{Full}) \right] \tag{7}
\]

which is distributed asymptotically as \( \chi^2 \) with \([(p + 1) + ((p + 2) \times (j - 1))] - j \) degrees of freedom. It is noteworthy to point out that \( \hat{\beta}_{Reduced} \) includes all the regression parameters from the constant only joint model and \( \hat{\beta}_{Full} \) includes all the parameters from the full joint model.

### 3.2 Predictive Models and Joint Probabilities

The predicted joint probability that a subject with covariate vector \( (X^* = x^*) \) for a trajectory as shown in Fig. 1 can be predicted using Eq. (5) and using the predicted marginal first-order conditional probabilities from the fitted marginal and all first-order regressive logistic models. Based on Eq. (5), the predicted joint probabilities for
\[ Y_{i1} = y_{i1} \text{ and } Y_{i2} = y_{i2} \text{ are:} \]

\[
\hat{P}_{y_{i1},y_{i2}}(x) = \hat{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} | x) \\
= \hat{P}(Y_{i1} = y_{i1} | X = x) \times \hat{P}(Y_{i2} = y_{i2} | Y_{i1} = y_{i1}; X = x) \\
= \hat{P}_{1x,y_{i1}}(x) \times \hat{P}_{1x,y_{i2}|y_{i1}}(x).
\]

Now, the predicted marginal probabilities \( \hat{P}(Y_{i1} = 1 | X = x) \) and \( \hat{P}(Y_{i1} = 0 | X = x) \) can be estimated from the fitted marginal model. The first-order conditional probabilities \( \hat{P}(Y_{i2} = s | Y_{i1} = y_{i1}; X = x), s, y_{1} = 0, 1, \) can be estimated from the fitted first-order regressive model using covariate vector \( Z' = [x^*, y_{i1}] \) and \( X^* = x^* \).

Then, we can plug-in the predicted marginal and conditional probabilities in Eq. (5) to obtain the joint probabilities for events.

## 4 Analysis of Texas COVID-19 Data

Texas COVID-19 data used in this paper were downloaded from the Texas Health and Human Services website (https://dshs.texas.gov/coronavirus//additionaldata.aspx). It provides daily new cases and deaths for 254 counties in Texas. We used data from March 4, 2020, to July 25, 2020. Also, the population size for each county is considered as covariate, which ranges from 92 to 4978845 for different counties. As some counties do not have any daily fatality for many days, we summed weekly deaths and new cases for our analysis. For modelling purposes, we have converted death as binary variable with any death as 1 or no death as 0, as the goal of public health experts and scientists are to minimize the death as 0. In our analysis, we have considered death as repeated binary outcome variable which repeats over the 20 weeks and are defined as \( Y_{i1}, \cdots, Y_{i20} \), respectively. We are interested in predicting the trajectory risks of death over the weeks and the impact of available risk factors on the trajectory. Due to the small sample size, i.e. only 254 counties in Texas, we divide the data into training (90%) and test (10%) sets based on the first-week samples to determine over(under)fitting and to assess the generalization ability of the models used. We used R and parallel programming package to fit all the models simultaneously. Table 1 displays the distribution of weekly death.

The prediction accuracy of the proposed joint model along with the three machine learning techniques is presented in Table 2. In general, the weekly accuracy of all three models such as proposed regressive model, neural network (NN), random forest (RF) is relatively very high compared to support vector machine (SVM) technique. The accuracy of SVM is low for most of the weeks. The same is true for both the training and test data. Also, we observed overfitting a few models. Regarding the model performance, comparing the accuracy of all these three models, it is found that the random forest model showed the highest accuracy, followed by the proposed regressive model and neural network, respectively. The fitted coefficients for marginal and proposed regressive models are shown in Table 3. The population size is positive and significantly associated with the outcomes for all the marginal and first-order regressive models with few exceptions. A similar pattern is observed between daily
### Table 1  Distribution of deaths for consecutive weeks

| Fatality | Weeks | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------|-------|---|---|---|---|---|---|---|---|---|----|
| No deaths |       | 221 | 119 | 99 | 106 | 94 | 96 | 84 | 41 | 26 | 20 |
| Any deaths |     | 33  | 135 | 155 | 148 | 160 | 158 | 170 | 213 | 228 | 234 |
| Total |     | 254  | 254  | 254  | 254  | 254  | 254  | 254  | 254  | 254  | 254  |

| Fatality | Weeks | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------|-------|----|----|----|----|----|----|----|----|----|----|
| No deaths |       | 249 | 228 | 207 | 209 | 211 | 214 | 205 | 213 | 191 | 175 |
| Any deaths |     | 5  | 26  | 47  | 45  | 43  | 40  | 49  | 41  | 63  | 79  |
| Total |     | 254  | 254  | 254  | 254  | 254  | 254  | 254  | 254  | 254  | 254  |

### Table 2  Training and test data accuracy for different models

| Weeks | Models | Proposed model | NN | SVM | RF |
|-------|--------|----------------|----|-----|----|
|       |        | Train | Test | Train | Test | Train | Test | Train | Test |
| 1     |        | 0.925 | 0.964 | 0.938 | 0.964 | 0.925 | 0.964 | 0.983 | 1.000 |
| 2     |        | 0.841 | 0.964 | 0.845 | 0.964 | 0.181 | 0.143 | 0.904 | 0.933 |
| 3     |        | 0.801 | 0.750 | 0.841 | 0.821 | 0.195 | 0.357 | 0.904 | 1.000 |
| 4     |        | 0.805 | 0.857 | 0.801 | 0.857 | 0.199 | 0.143 | 0.921 | 0.867 |
| 5     |        | 0.832 | 0.786 | 0.832 | 0.786 | 0.177 | 0.143 | 0.891 | 0.733 |
| 6     |        | 0.827 | 0.821 | 0.832 | 0.821 | 0.208 | 0.179 | 0.891 | 0.800 |
| 7     |        | 0.796 | 0.857 | 0.819 | 0.786 | 0.204 | 0.143 | 0.904 | 0.867 |
| 8     |        | 0.889 | 0.929 | 0.889 | 0.929 | 0.128 | 0.179 | 0.950 | 0.933 |
| 9     |        | 0.903 | 0.964 | 0.907 | 0.964 | 0.102 | 0.107 | 0.975 | 1.000 |
| 10    |        | 0.916 | 1.000 | 0.934 | 0.036 | 0.088 | 1.000 | 0.975 | 1.000 |
| 11    |        | 0.987 | 1.000 | 0.982 | 0.964 | 0.982 | 0.964 | 0.992 | 1.000 |
| 12    |        | 0.925 | 0.893 | 0.898 | 0.893 | 0.925 | 0.857 | 0.962 | 1.000 |
| 13    |        | 0.885 | 0.964 | 0.885 | 0.964 | 0.881 | 0.964 | 0.925 | 1.000 |
| 14    |        | 0.889 | 0.929 | 0.903 | 0.929 | 0.894 | 0.964 | 0.933 | 0.933 |
| 15    |        | 0.916 | 0.786 | 0.920 | 0.786 | 0.920 | 0.786 | 0.921 | 0.933 |
| 16    |        | 0.889 | 0.857 | 0.850 | 0.786 | 0.894 | 0.857 | 0.933 | 0.933 |
| 17    |        | 0.863 | 0.821 | 0.867 | 0.821 | 0.867 | 0.786 | 0.867 | 1.000 |
| 18    |        | 0.898 | 0.857 | 0.889 | 0.857 | 0.903 | 0.857 | 0.925 | 0.933 |
| 19    |        | 0.845 | 0.821 | 0.850 | 0.786 | 0.150 | 0.179 | 0.904 | 1.000 |
| 20    |        | 0.854 | 0.821 | 0.845 | 0.821 | 0.841 | 0.750 | 0.887 | 0.867 |
new cases and death. Daily new counts of cases for all the first-order regressive models are positive and significantly associated with the outcomes except for some. While we assessed the effect of covariates using proposed statistical learning models as shown in Table 3, we could not do so using machine learning models as those have interpretation problems. But we used those as one of our objectives is to predict fatality.

The predicted trajectory of conditional probabilities for Harris county is displayed in Fig. 2. The population of this county is 4978845. We presented the predicted trajectories using all four models. Both the proposed regressive and random forest models predicted the highest risks of deaths over time and these two lines overlapped. The dashed line is from the proposed model and the dashed-dotted line from the random forest models. The neural network (dotted line) closely follows these trajectories. The trajectory (the solid bottom line) predicted using the support vector machine showed the worst performance. This county has the highest population size and observed death in all the 20 consecutive weeks. However, there were no new cases during week 11 for this county. The predicted trajectories using the proposed regressive model and random forest are similar and predicted the highest risk, which is expected. However, the support vector machine algorithm predicted much lower risks and downward trends up to the 10th week. This may be due to no new cases during the 11th week. The corresponding trajectories of joint probabilities are shown in Fig. 3. After the pre-processing, the data for two counties (Anderson and Archer) used for model fitting are shown in Table 4.

In Fig. 4, we present the trajectory of conditional probabilities for Andrewes county using all the models. The population size of this county is much smaller (22269) compared to Harris county. Also, this county did not observe death during the majority of the weeks. It is clear from the figure that the probability of death is lower than 0.50 or close to zero when no new cases are observed during a week. This flattening trend of lowering death risk is clearly visible from the trajectories of joint probabilities shown in Fig. 5. In general, the predicted paths from all the models are relatively close except for support vector machine.

5 Conclusions

This paper proposed a marginal-conditional modelling approach to predict trajectory risks of fatalities from COVID-19 longitudinally using daily new cases and other available covariates from 254 counties from Texas in the USA. We assess the impact of weekly new cases and the county population size on the trajectories. Also, we extended some commonly used machine learning algorithms that use for the classification of categorical outcomes from cross-sectional data for the trajectory risk predictions from repeated measures data. It is evident from our study that controlling the occurrence of new cases reduces the risk of death. Also, the trajectory risk of death is much higher for the densely populated county. Relevant authorities of all countries are trying hard to reduce the new cases of COVID-19 towards zero. Only by ending the pandemic everywhere, we can achieve this goal.

Governments around the world are implementing various strict restrictions, including travel bands, social distancing, and unnecessary movements, besides travel...
| Covariates | Weeks | $\hat{\beta}$ | SE | p-value | Weeks | $\hat{\beta}$ | SE | p-value |
|------------|-------|--------------|----|---------|-------|--------------|----|---------|
| Constant   | Week 1| -3.14        | 0.34 | 0.00    | Week 11| -20.61       | 3961.56 | 1.00    |
| New cases  |       | -1.11        | 3.11 | 0.72    |        |              |    |         |
| Population |       | 0.00         | 0.00 | 0.00    |        | 0.00         | 0.00 | 0.00    |
| Deaths     |       |              |     |         |        | 15.43        | 3961.56 | 1.00    |
| Constant   | Week 2| -2.36        | 0.34 | 0.00    | Week 12| -2.99        | 0.32 | 0.00    |
| New cases  |       | 1.36         | 0.51 | 0.01    |        | 0.77         | 1.15 | 0.51    |
| Population |       | 0.00         | 0.00 | 0.00    |        | 0.00         | 0.00 | 0.00    |
| Deaths     |       | 0.10         | 1.13 | 0.93    |        | 0.21         | 2.84 | 0.94    |
| Constant   | Week 3| -1.62        | 0.29 | 0.00    | Week 13| -2.57        | 0.28 | 0.00    |
| New cases  |       | 1.15         | 0.40 | 0.00    |        | 1.77         | 0.55 | 0.00    |
| Population |       | 0.00         | 0.00 | 0.00    |        | 0.00         | 0.00 | 0.00    |
| Deaths     |       | 0.80         | 0.43 | 0.07    |        | 0.45         | 0.74 | 0.54    |
| Constant   | Week 4| -2.05        | 0.33 | 0.00    | Week 14| -2.84        | 0.31 | 0.00    |
| New cases  |       | 1.52         | 0.40 | 0.00    |        | 1.09         | 0.56 | 0.05    |
| Population |       | 0.00         | 0.00 | 0.00    |        | 0.00         | 0.00 | 0.00    |
| Deaths     |       | 0.29         | 0.41 | 0.49    |        | 0.92         | 0.55 | 0.10    |
| Constant   | Week 5| -2.22        | 0.36 | 0.00    | Week 15| -3.12        | 0.35 | 0.00    |
| New cases  |       | 1.54         | 0.44 | 0.00    |        | 1.63         | 0.69 | 0.02    |
| Population |       | 0.00         | 0.00 | 0.00    |        | 0.00         | 0.00 | 0.08    |
| Deaths     |       | 1.19         | 0.43 | 0.01    |        | 1.82         | 0.61 | 0.00    |
| Constant   | Week 6| -1.93        | 0.33 | 0.00    | Week 16| -2.93        | 0.32 | 0.00    |
| New cases  |       | 0.72         | 0.40 | 0.07    |        | 2.03         | 0.54 | 0.00    |
| Population |       | 0.00         | 0.00 | 0.00    |        | 0.00         | 0.00 | 0.11    |
| Deaths     |       | 1.06         | 0.41 | 0.01    |        | 1.11         | 0.60 | 0.06    |
| Constant   | Week 7| -1.21        | 0.28 | 0.00    | Week 17| -2.58        | 0.28 | 0.00    |
| New cases  |       | 1.31         | 0.39 | 0.00    |        | 1.00         | 0.54 | 0.07    |
| Population |       | 0.00         | 0.00 | 0.08    |        | 0.00         | 0.00 | 0.01    |
| Deaths     |       | 1.28         | 0.39 | 0.00    |        | 1.71         | 0.52 | 0.00    |
| Constant   | Week 8| -1.15        | 0.39 | 0.00    | Week 18| -2.77        | 0.30 | 0.00    |
| New cases  |       | 2.04         | 0.55 | 0.00    |        | 0.73         | 0.58 | 0.21    |
| Population |       | 0.00         | 0.00 | 0.01    |        | 0.00         | 0.00 | 0.00    |
| Deaths     |       | 0.57         | 0.56 | 0.31    |        | 0.78         | 0.55 | 0.16    |
| Constant   | Week 9| -0.68        | 0.44 | 0.13    | Week 19| -2.34        | 0.27 | 0.00    |
| New cases  |       | 1.69         | 0.61 | 0.01    |        | 0.47         | 0.63 | 0.46    |
| Population |       | 0.00         | 0.00 | 0.15    |        | 0.00         | 0.00 | 0.00    |
| Covariates | Weeks | $\hat{\beta}$ | SE  | p-value | Weeks | $\hat{\beta}$ | SE  | p-value |
|------------|-------|---------------|-----|---------|-------|---------------|-----|---------|
| Deaths    |       | 1.40          | 0.62| 0.02    | 0.83  | 0.55          | 0.13|         |
| Constant  | Week 10 | -0.61         | 0.55| 0.27    | -2.45 | 0.29          | 0.00|         |
| New cases |       | 1.70          | 0.66| 0.01    | 1.03  | 0.44          | 0.02|         |
| Population|       | 0.00          | 0.00| 0.42    | 0.00  | 0.00          | 0.00|         |
| Deaths    |       | 1.76          | 0.63| 0.00    | 1.25  | 0.46          | 0.01|         |

Fig. 2  Trajectory of conditional probabilities for Harris county using four models

Fig. 3  Trajectory of joint probabilities for Harris county using four models
restriction and quarantine of both suspected individuals and subjects who have had close contacts with suspected cases. The aim is that reducing new infections towards zero will reduce the spread of this disease hence death. While people around the world overwhelmingly support pandemic-related restrictions, but many communities became impatient for these restrictions and want those removed. A positive association of population size with death from our studies may reiterate the need for those restrictions for highly populated places to reduce the new cases and hence an increasing number of fatalities. We believe that analysis with the availability of more detailed data along with related risk factors may provide an in-depth understanding regarding the dynamics.
Table 4  Pre-processed data for model fitting from Anderson and Archer counties

| County | Nd | Nc | Population | County | Nd | Nc | Population | Week |
|--------|----|----|------------|--------|----|----|------------|------|
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 1    |
| Anderson | 1  | 0  | 58199      | Archer | 0  | 0  | 1948       | 2    |
| Anderson | 1  | 1  | 58199      | Archer | 1  | 0  | 1948       | 3    |
| Anderson | 1  | 1  | 58199      | Archer | 1  | 0  | 1948       | 4    |
| Anderson | 1  | 1  | 58199      | Archer | 0  | 0  | 1948       | 5    |
| Anderson | 1  | 1  | 58199      | Archer | 1  | 0  | 1948       | 6    |
| Anderson | 1  | 1  | 58199      | Archer | 0  | 0  | 1948       | 7    |
| Anderson | 1  | 1  | 58199      | Archer | 0  | 1  | 1948       | 8    |
| Anderson | 1  | 1  | 58199      | Archer | 0  | 0  | 1948       | 9    |
| Anderson | 1  | 1  | 58199      | Archer | 0  | 0  | 1948       | 10   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 11   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 12   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 13   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 14   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 15   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 16   |
| Anderson | 0  | 0  | 58199      | Archer | 0  | 0  | 1948       | 17   |
| Anderson | 0  | 1  | 58199      | Archer | 0  | 0  | 1948       | 18   |
| Anderson | 1  | 0  | 58199      | Archer | 0  | 0  | 1948       | 19   |
| Anderson | 0  | 1  | 58199      | Archer | 0  | 0  | 1948       | 20   |

Nd: weekly death from COVID 19: 0 = No death and 1 = 1 or more deaths
Nc: Number of weekly new cases

**Acknowledgements**  This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

**Declarations**

**Conflict of interest**  The authors declare that they do not have any conflict of interest to declare.

**References**

1. WHO Coronavirus Disease (COVID-19) Dashboard, Map Data, 2020. Available from: https://covid19.who.int/
2. Carpenter, T.E., O’Brien, J.M., Hagerman, A.D., et al.: Epidemic and economic impacts of delayed detection of foot-and-mouth disease: a case study of a simulated outbreak in California. J. Vet. Diagn. Invest. 23, 26−33 (2010)
3. Sahai, A.K., Rath, N., Sood, V., Singh, M.P.: ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes Metab. Syndr.: Clin. Res. Rev. 14(5), 1419−1427 (2020). https://doi.org/10.1016/j.dsx.2020.07.042
4. Dal Molin Ribeiro, M.H., da Silva, R.G., Mariani, V.C., dos Santos Coelho, L.: Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil Chaos. Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.109853

5. Vena, A., Giacobbe, D. R., Di Biagio, A., Mikulsk, M., Taramasso, L., De Maria, A., et al.: Clinical characteristics, management and in-hospital mortality of patients with coronavirus disease 2019 in Genoa, Italy, Clinical Microbiology and Infection, Available online 15 August 2020, doi: https://doi.org/10.1016/j.cmi.2020.07.049

6. Ceylan, Z.: Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci. Total Environ. 729, 10 (2020). https://doi.org/10.1016/j.scitotenv.2020.138817

7. Yang, Q., Wang, J., Ma, H., Wang, X.: Research on COVID-19 based on ARIMA model-Taking Hubei, China as an example to see the epidemic in Italy. J. Infect. Public Health 13(10), 1415–1418 (2020). https://doi.org/10.1016/j.jiph.2020.06.019

8. Shastri, S., Singh, K., Kumar, S., Kour, P., Mansotra, V.: Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study, Chaos. Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.110227

9. da Silva, R.G., Dal Molin Ribeiro, M.H., Mariani, V.C., dos Santos Coelho, L.: Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables, Chaos. Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.110027

10. Tuli, S., Shikhar, T., Rakesh, T., Rakesh, S.S.: Gill Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet Things (2020). https://doi.org/10.1016/j.iot.2020.100222

11. Islam, M.A., Chowdhury, R.I.: Prediction of disease status: a regressive model approach for repeated measures. Stat. Methodol. 7, 520–540 (2010). https://doi.org/10.1016/j.stamet.2010.03.001

12. Chowdhury, R.I., Islam, M.A.: Regressive models for risk prediction for repeated multinomial outcomes: an illustration using Health and Retirement Study (HRS) data. Biometric. J. 62, 898–915 (2020). https://doi.org/10.1002/bimj.201800101

13. Islam, M.A., Chowdhury, R.I., Huda, S.: Markov Models with Covariate Dependence for Repeated Measures. Nova Science, New York (2009)

14. Islam, M.A., Chowdhury, R.I., Singh, K.P.: A Markov model for analyzing polytomous outcome data. Pak. J. Stat. Op. Res. 8, 593–603 (2012)

15. Lindsey, J.K., Lamber, P.: On the appropriateness of marginal models for repeated measurements in clinical trials. Stat. Med. 17, 447–469 (1998)

16. Lee, Y., Nelder, J.A.: Conditional and marginal models: another view. Stat. Sci. 19, 219–238 (2004)

17. Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88, 9–25 (1993)

18. Ripley, Brian D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)

19. Boser, B.E., Guyon, I.M. and Vapnik, V.N.: A training algorithm for optimal margin classifiers. 5th Annual ACM Workshop on COLT. Pittsburgh, PA, ACM Press, pp. 144-152 (1992)

20. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learn. 20, 273–297 (1995)

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