Journal of Population Therapeutics & Clinical Pharmacology

Special Issue
DOI: 10.15586/jptcp.v27iSP1.721

An application of a mixture of exponential distributions for assessing hazard rates from COVID-19

Athanase Polymenis

Corresponding author: Athanase Polymenis, Department of Economics, University of Patras, University Campus at Rio, 26504 Rio-Patras, Greece. Email: athanase@upatras.gr

Submitted: 18 June 2020. Accepted: 3 August 2020. Published: 14 August 2020

ABSTRACT

In the present study, we are interested in modeling repose time periods (the length of the time intervals between successive deaths) caused by a new, widespread disease called covid-19. This is useful for predicting probabilities of new deaths that occur within pre-determined time intervals. In practical applications, the choice of the statistical model is crucial for obtaining accurate estimates of death hazard rates. Based on an earlier research, we propose to use a mixture of exponential distributions; this model is simple to implement when hazard rates obtained from the components of the mixture are easily calculated, and it is adequate for dealing with nonstationary time series as those appearing in the case of this disease. The model is then applied to the example of Italy, and it appears to be also useful for comparing hazard rates along time.

Keywords: *hazard rate; infectious disease; mixture of exponentials; nonstationary time series*

INTRODUCTION

A new coronavirus called SARS-CoV-2, which causes Covid-19, surfaced a few months ago in China. The new virus affects a patient's lungs, causing severe pneumonia, thereby increasing the risk of death, especially among the elderly or those suffering from other diseases. A major problem is that this infectious disease rapidly spreads in the population, and the high rates of infection lead to a large number of intensive care admissions. As a result, death rates are particularly high, causing a lot of fear and distress in the population, and this is even more so because presently there are no drugs or vaccine to treat this disease. At the same time, health systems in affected countries face many difficulties in tackling the high number of cases as well as in dealing with patients who need intensive care. In this scenario, in this work, we focus on forecasting death rates of patients with covid-19. The distribution of the time periods between deaths (repose periods) provides useful information about the probability that the next death will occur within some specific time interval, and thus it is very important for forecasting purposes. In general, Exponential, Binomial, and Weibull distributions are widely used in survival analyses for modeling occurrence rates. However, in a somewhat different context concerning volcanic hazard data, mixtures of exponentials[1] have been proposed in earlier literature as an alternative to the aforementioned methods; note that Exponential and Binomial distributions are not generally adequate because of the nonstationary nature of the time series involved. On the other hand, the authors have underlined the advantages of using mixtures of exponentials in comparison to a Weibull distribution. In the present study, we are also faced with nonstationary time series, as there are different regimes suggested by the data as will be shown in the sequel. In view of this, we also propose to use a mixture of exponential distributions for modeling death hazards over time due to covid-19. Results of this research could then be used for better organizing the health system (for instance, forecasting needs of hospitals with respect to numbers of ventilators, increasing intensive care units, etc.). We applied this model in the case of Italy, which is amongst the European countries most affected by the disease, with thousands of fatalities; however, the model can also be used in any heavily affected country. It is remarkable that there is a flourishing international literature on covid-19, and the case of Italy has been particularly investigated[2,3] (e.g., reference 2 describes how the Italian government takes measures in order to contain the health risk, whereas in reference 3, a model is proposed for predicting the course of the epidemic).

METHODS

The theoretical approach

Finite mixtures have been well documented in the existing literature. More specifically, a finite mixture of unknown exponential distributions with two components has a probability density function of the form[4]

$$f(t;\Lambda) = pf_1\,(t;\lambda_1) + (1-p)f_2\,(t;\lambda_2)$$

with $f_1(t;\lambda_1) = \lambda_1 e^{(-\lambda_1 t)}$ and $f_2(t;\lambda_2) = \lambda_2 e^{-\lambda_2 t}$; Λ is the vector parameter (p,λ_1,λ_2); the symbol $t(\geq 0)$ stands for time; p is the weight of the first component and $1-p$ the weight of the second component, with $p > 0$; f_1 and f_2 are the component densities, which are both exponentials with corresponding parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, also called rate parameters or hazard rates (for our case, these will be the rate of mortality per minute). Remark that this model can also be easily generalized to include more than two components. The cumulative distribution function of a two-component mixture of exponentials takes the form $F(t;\Lambda) = p(1-e^{-\lambda_1 t}) + (1-p)(1-e^{-\lambda_2 t})$. Then, the probability that at least a new death will occur in the next t min supposing that the last death has just occurred is a special case of equation (5) of Mendoza-Rosas and De la Cruz-Reyna,[1] obtained for $s = 0$, and it is equal to $F(t;\Lambda)$. In order to compute this probability, we need to estimate the parameters λ_1 and λ_2, and the weights p (and $1-p$) using the data in hand. In order to proceed, we use the idea of Mendoza-Rosas and De la Cruz-Reyna[1] for fitting a mixture model; that is, we consider the number of regimes to be the number of components of the model. Regimes can be represented graphically using the change of slope in the cumulative plot of the number of deaths.

Data collection design

Data concerning fatal cases were collected for the time period between 21 February 2020 (date of the first fatal case in Italy) and 30 April 2020, and between the 21 February 2020 and 8 May 2020 for comparison reasons using the World Health Organization database (https://bing.com/covid/local/italy).

Data classification

In the case of Italy, it appears from Figures 1 and 2, replicated from https://bing.com/covid/local/italy (see arrow for cumulative distribution function of deaths), that there are two regimes – representing the components of the mixture – which concern time periods from 21 February 2020 to 5 April 2020 (first category), corresponding to 45 days, and from 6 April 2020 onward (second category). The first category refers to increasing number of deaths on a daily basis, whereas the second one refers to number of deaths becoming stable on a daily basis and then slowly decreasing. This is because from the first week of April onward, the number of deaths on a daily basis appeared to stabilize and then slowly started to decrease. In view of this visual inspection,[1] a two-component mixture of exponentials seems suitable for modeling death rates, each component being assigned to a specific category. For comparison reasons, we consider two time periods corresponding to 21 February 2020–30 April 2020 (Figure 1) and 21 February 2020–8 May 2020 (Figure 2), where the latter period can be viewed as a complement to the former.

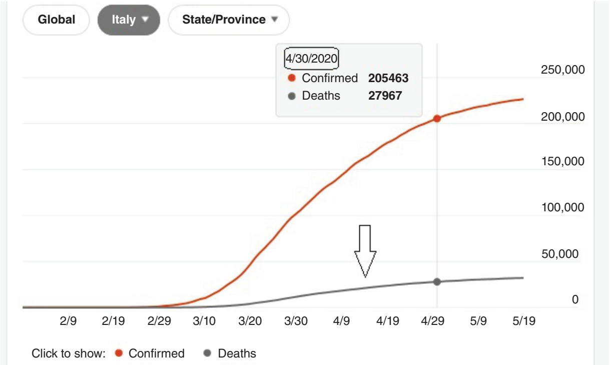

FIG 1. Cumulative distribution of number of deaths for the period between 21 February 2020 and 30 April 2020.

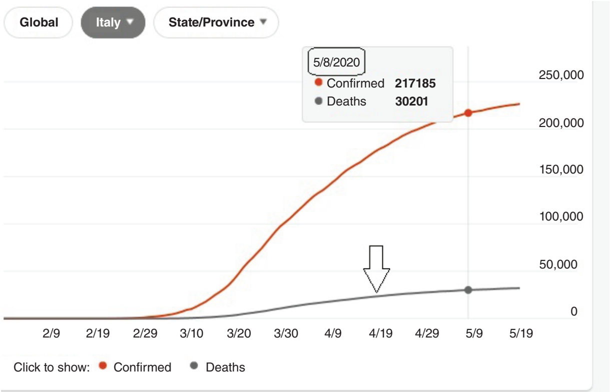

FIG 2. Cumulative distribution of the number of deaths for the time period between 21 February 2020 and 8 May 2020.

Statistical analysis

We now implement this model for the period from 21 February 2020 to 30 April 2020. The first category consists of 45 days (i.e. 45 × 1440 = 64,800 min), whereas the second one consists of 25 days (i.e. 25 × 1440 = 36,000 min). We thus have a two-component mixture model with 15,887 deaths for the category corresponding to 21 February 2020–5 April 2020 (see https://bing.com/covid/local/italy) and 12,080 deaths for the other category corresponding to 6 April 2020–30 April 2020 (calculated as total number of deaths up to 30 April 2020 = 27,967 – 15,887). The duration in minutes for each regime are $D_1 = 45 \times 1440 = 64{,}800$ (for 21 February 2020–5 April 2020) and $D_2 = 25 \times 1440 = 36{,}000$ (for 6 April 2020–30 April 2020).

The rates of deaths per minute λ_1 and λ_2 can be calculated as $\lambda_1 = \dfrac{15887}{64800} = 0.24517$ for the regime corresponding to 21 February 2020 to 5 April 2020 and $\lambda_2 = \dfrac{12080}{36000} = 0.33556$ for the regime corresponding to 6 April 2020 to 30 April 2020.

Note that the regime with a duration of 25 days (36,000 min) has a higher death rate (λ_2) than the regime with a duration of 45 days (64,800 min) with death rate (λ_1); this result is compatible with the remark by Mendoza-Rosas and De la Cruz-Reyna,[1] namely, that regimes of shorter duration tend to have higher occurrence rates. On the other hand, the corresponding weights are calculated as $p = \frac{D_2}{D_1 + D_2} = \frac{36000}{100800} = 0.357$ and $1 - p = \frac{D_1}{D_1 + D_2} = \frac{64800}{100800} = 0.643$. This result is a direct consequence of equation (6) of Mendoza-Rosas and De la Cruz-Reyna,[1] where $p = w_1 = \frac{(D_1 + D_2) - D_1}{((D_1 + D_2) - D_1) + ((D_1 + D_2) - D_2)}$ and $1 - p = w_2 = \frac{(D_1 + D_2) - D_2}{((D_1 + D_2) - D_1) + ((D_1 + D_2) - D_2)}$ are the normalized complements of the duration in years of each regime. Weighting factors and rates of deaths per minute for the period between 21 February 2020 and 30 April 2020 appear in Table 1. Using these results, probabilities of at least one death occurring within the next t min for period between 21 February 2020 and 30 April 2020 are immediately obtained via the cumulative distribution function F for pre-specified values of t. For instance, assume that a death has just occurred; then, the probability that at least a new death will occur within the next 10 min is $F(10;0.357,0.24517,0.33556) = 0.94625$.

We also repeated the same reasoning for the time period between 21 February 2020 and 8 May 2020, with the same first category as before and a second category pertaining to data collected between 6 April 2020 and 8 May 2020. In that case, D_1 = 64,800 min (as previously) and D_2 = 47,520 min. Noticing that the number of deaths concerning the second regime is now 14,314 (30,201–15,887), we obtain $\lambda_1 = \frac{15887}{64800} = 0.24517$ (as previously) and $\lambda_2 = \frac{14314}{47520} = 0.30122$ (Note that again a regime of shorter duration has a higher occurrence rate). On the other hand, we obtain $p = \frac{D_2}{D_1 + D_2} = \frac{47520}{112320} = 0.4231$ and $1 - p = \frac{D_1}{D_1 + D_2} = \frac{64800}{112320} = 0.5769$. Weighting factors and rates of deaths per minute for the period between 21 February 2020 and 8 May 2020 appear in Table 2. We can now compute probabilities of occurrence of at least one death in a predetermined time interval. For example, taking time t to be 10 min, and using our estimated parameters $w_1, \lambda_1, \lambda_2$, we find $F(10;0.4231,0.24517,0.30122) = 0.9352$.

Assessment of the model

The mixture model was assessed for the time period between 21 February 2020 and 8 May 2020 using a Kolmogorov–Smirnov goodness-of-fit test.[5] Like in Mendoza-Rosas and De la Cruz Reyna,[1] we also tested sensitivity in the quality of the fit by taking neighboring dates to 5 April 2020 (last day of the first regime selected by graphical inspection) as final dates for the first category and implementing again the Kolmogorov–Smirnov test.

RESULTS

Using the aforementioned analysis, we now present in Tables 1, 2, and 3, the estimated parameters and probabilities of deaths within predetermined time intervals.

The cumulative distribution function is computed in the same way for values of t = 2, 5, 10, 15, and 20 min, for both periods, and results are presented in Table 3.

DISCUSSION

Results presented in Table 3 are obtained using the hazard rates and the weighting factors, pertaining

TABLE 1. Observed Death Regimes and Calculated Parameters of the Mixture Model for the Period Between 21 February 2020 and 30 April 2020.

Regime	Time period	Number of deaths	Duration of regime (minutes)	Rate per minute λ	Weighting factor w
1	21 February 2020– 5 April 2020	15,887	64,800	0.24517	0.357
2	6 April 2020– 30 April 2020	12,080	36,000	0.33556	0.643

TABLE 2. Observed Death Regimes and Calculated Parameters of the Mixture Model for the Period Between 21 February 2020 and 8 May 2020.

Regime	Time period	Number of deaths	Duration of regime (minutes)	Rate per minute λ	Weighting factor w
1	21 February 2020– 5 April 2020	15,887	64,800	0.24517	0.4231
2	6 April 2020– 8 May 2020	14,314	47,520	0.30122	0.5769

to the mixtures of exponentials that appear in Tables 1 and 2. These results show that the survival function, that is, the probability of exceeding some pre-specified t (= 1 – value from cumulative distribution) is slightly larger for the period between 21 February 2020 and 8 May 2020, which reflects the fact that the number of deaths occurring within the period between 1 May 2020 and 8 May 2020 have decreased in comparison to previous dates.

TABLE 3. Probabilities of Occurrence of at Least One Death in the Next *t* min.

t	Cumulative Distribution 21 February 2020–30 April 2020	Cumulative Distribution 21 February 2020–8 May 2020
2	0.467	0.4251
5	0.775	0.7479
10	0.94625	0.9352
15	0.9883	0.983
20	0.99657	0.9955

It is interesting to note that if we had used a single exponential distribution (instead of a mixture), we would then have obtained very close results to those provided by our mixture model for both periods, namely, 21 February 2020–30 April 2020 and 21 February 2020–8 May 2020. Indeed, let us consider the cdf $F_0^*(t)$ of a single exponential distribution with parameter $\lambda = 0.27\left(=\frac{30201}{112320}\right)$, where λ is the overall hazard rate, for, say, the period between 21 February 2020 and 8 May 2020. We observe that the survival functions for the single exponential distribution at times 1, 2, and 3, corresponding to $1 - F_0^*(1) = 0.763$, $1 - F_0^*(2) = 0.583$, $1 - F_0^*(3) = 0.445$, are approximately equal to the survival functions $1 - F(1) = 0.758$, $1 - F(2) = 0.575$, $1 - F(3) = 0.437$, corresponding to the mixture of exponentials, and this result holds good for any value of t. This is a consequence of the fact that hazard rates λ_1 and λ_2 are close enough in this case and so the mixture model can be approximated by a single exponential distribution with hazard rate λ; therefore, in the case of Italy, one can use equivalently either a mixture of exponentials or a single exponential distribution for prediction purposes. However, the proposed model would certainly be useful for those cases where the rates are quite different, as it would also catch large changes in regimes. Finally, we conclude from the above that the probability of surviving more than *t* min obtained from the mixture distribution steeply declines as duration increases in the same way as mentioned in Indrayan and Holt[6] [see in particular Figure E.8(b)], which is a characteristic feature of an exponential distribution.

CONCLUSIONS

In the present paper, we have proposed a method, based on mixtures of exponentials, for forecasting death hazards in populations hit by covid-19. In an earlier application to volcanic hazards, this method was shown to outperform standard techniques used in survival analysis. Two main advantages of the method are that it captures the different underlying categories (or regimes) suggested by the data and that it is quite handy and simple to implement, as component-wise hazard rates are readily obtained from the data. This is a very important asset in the context of covid-19 because, in contrast with volcanic hazards, repose periods are very short and so it would be useful to apply the method as many times as required in order to follow the course of the disease and thus help officials cope with difficulties concerning the health system at any stage of the pandemic. As an example, an application to the case of Italy was performed, and hazard rates were provided for two time periods for comparison reasons. The mixture model provided encouraging results, as it showed that a decrease in death numbers results, as expected, in larger values of the survival function.

CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

FUNDING

This study had no financial support.

DATA AVAILABILITY STATEMENT

Data are free to download.

COMPLIANCE WITH ETHICAL STANDARDS

No ethical standards are needed for this study.

ACKNOWLEDGEMENTS

The author is grateful for useful comments from an anonymous referee which helped improve the present article.

REFERENCES

1. Mendoza-Rosas AT, De la Cruz-Reyna S. A mixture of exponentials distribution for a simple and precise assessment of the volcanic hazard. Nat Hazards Earth Syst Sci 2009;9:425–31. http://dx.doi.org/10.5194/nhess-9-425-2009.
2. Giangaspero M. Covid-19 epidemic in Italy: Lesson learning. J Fam Med Dis Prev 2020;6:119. http://dx.doi.org/10.23937/2469-5793/1510119.
3. Giordano G, Blanchini F, Bruno R, et al. Modelling the covid-19 epidemic and implementation of population-wide interventions in Italy. www.nature.com/articles/s41591-020-0883-7, 2020.
4. Titterington DM, Smith AFM, Makov UE. Statistical analysis of finite mixture distributions. New York: John Wiley & Sons, 1985:74–5.
5. Conover WJ. Practical nonparametric statistics. 3rd ed. New York: Wiley Series in Probability and Statistics, 1999:430–5.
6. Indrayan A, Holt MP. Concise encyclopedia of biostatistics for medical professionals, 1st ed. Boca Raton (Florida): CRC Press, Chapman Hall, 2016:227–8.