Eur. Phys. J. Spec. Top. (2022) 231:3419–3426
https://doi.org/10.1140/epjs/s11734-022-00430-y

THE EUROPEAN
PHYSICAL JOURNAL
SPECIAL TOPICS

Regular Article

Estimating a continuously varying offset between multivariate time series with application to COVID-19 in the United States

Nick James[1] and Max Menzies[2,a]

[1] School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia
[2] Beijing Institute of Mathematical Sciences and Applications, Tsinghua University, Beijing 101408, China

Received 28 October 2021 / Accepted 18 December 2021 / Published online 11 January 2022

Abstract This paper introduces new methods to track the offset between two multivariate time series on a continuous basis. We then apply this framework to COVID-19 counts on a state-by-state basis in the United States to determine the progression from cases to deaths as a function of time. Across multiple approaches, we reveal an "up-down-up" pattern in the estimated offset between reported cases and deaths as the pandemic progresses. This analysis could be used to predict imminent increased load on a healthcare system and aid the allocation of additional resources in advance.

1 Introduction

Understanding the trajectories of and relationships between COVID-19 case and death counts assists governments in anticipating and responding to the impact of the pandemic. In the United States (US) and elsewhere, high case counts have generally been closely followed by high hospital admissions, the use of costly equipment such as ICU beds and ventilators [1], and deaths. The strain on the healthcare system may be considerable and can even threaten the health of patients who are not afflicted by COVID-19 [2].

Unfortunately, the dynamics of the COVID-19 pandemic have been consistently difficult to describe and predict. Numerous factors may influence the virus' spread, including the emergence of new variants, changes in government policy and restrictions, community adherence to health recommendations, exhaustion with mitigation measures [3], community frustration, differing risk appetites by different population groups, and changing testing policies [4–7]. Thus, the actual and recorded counts of COVID-19 cases have exhibited complex dynamics since the arrival of the pandemic. One of the most significant attributes to be aware of is the delay between the onset of cases and their progression to deaths. This may predict the peak of hospitalisations and provide advance warning of increased loads on the healthcare system.

This paper serves this purpose by providing an in-depth mathematical study on the estimated average offset between reported cases and deaths, investigating this as a function of time. Several clinical trials and mathematical studies have aimed to do this in isolated incidences, but this paper is the first we are aware of to develop a nonlinear dynamical framework to calculate a continuously changing time-varying offset. In Sect. 2, we describe several approaches applicable to any two multivariate time series, and we report our results on the US in Sect. 3. We supply a more in-depth discussion in Sect. 4.

This paper builds on a long literature of *multivariate time series analysis* and a rich literature of nonlinear dynamics applied to the COVID-19 pandemic. Existing methods of time series analysis include parametric models [8] such as exponential [9] or power-law models [10] and nonparametric methods such as distance analysis [11], distance correlation [12–14] and network models [15]. Mathematical approaches to the COVID-19 pandemic are almost too numerous to cover. First, many papers based on existing mathematical models, such as the susceptible-infected-recovered (SIR) model and the (effective) reproductive ratio R_t [16], have been proposed and systematically collated by researchers [17,18]. Next, nonlinear dynamics researchers have proposed several sophisticated extensions to the classical predictive SIR model, including finding analytical solutions [19,20], modifications with additional variables [21–26], incorporation of Hamiltonian dynamics [27] or network models [28], and a closer analysis of uncertainty in the SIR equations [29]. Other mathematical approaches to prediction and analysis include power-law models [30–32], forecasting models [33], fractal curves [34], Bayesian methods [35], regression models and feature selection [36,37], Markov chain Monte Carlo models [38], distance analysis [39,40], network

[a] e-mail: max.menzies@alumni.harvard.edu (corresponding author)

models [41–43], analyses of the dynamics of transmission and contact [44,45], clustering [46,47] and many others [48–53]. Finally, numerous articles have been devoted specifically to the dynamics of COVID-19 in the United States [54], including incorporating spatial components of the virus' spread [55–57]. Our paper builds on this rich literature by developing a new mathematical method and a more extensive analysis of the progression of COVID-19 cases to deaths in the US than previously performed.

2 Methodology

Our data spans 26 February 2020 to 25 May 2021 across $n = 51$ regions (50 US states and the District of Columbia), a period of $T = 454$ days. We begin here to avoid periods of sparse reporting early in the pandemic. We end here due to changes in the CDC's reporting of case data, particularly between vaccinated and unvaccinated individuals, which will be detailed in Sect. 4. We order the states alphabetically and index them $i = 1, \ldots, n$. Let $x_i(t), y_i(t)$ be the multivariate time series of new daily COVID-19 cases and deaths, respectively, in each of the n regions, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. We introduce several new methods of analysis to find a continuously varying offset between the multivariate time series $x_i(t)$ and $y_i(t)$. All four methods involve 7-day averaging; this is performed due to the consistent weekly patterns of COVID-19 reporting, with lower reporting on the weekends. Thus, let $\hat{x}_i(t)$ be the rolling 7-day case average, defined by

$$\hat{x}_i(t) = \frac{1}{7} \sum_{s=t-6}^{t} x_i(s), \quad t = 7, \ldots, T, \tag{1}$$

and analogously let $\hat{y}_i(t)$ be the rolling 7-day death average. The following four methods, described in the proceeding subsections, are contributions to the literature.

2.1 Probability vector method

First, we estimate a continuously varying offset between multivariate time series $x_i(t)$ and $y_i(t)$ via a comparison of probability vectors of total counts. Let $p^X(t) \in \mathbb{R}^n$ be the probability vector of 7-day rolling averaged cases in each state, observed over an interval $[t-6, t]$. That is,

$$\begin{aligned} p_i^X(t) &= \frac{\hat{x}_i(t)}{\sum_{j=1}^{n} \hat{x}_j(t)} \\ &= \frac{\sum_{s=t-6}^{t} x_i(s)}{\sum_{s=t-6}^{t} \sum_{j=1}^{n} x_j(s)}, \quad i = 1, \ldots, n, \quad t = 7, \ldots, T. \end{aligned} \tag{2}$$

Equivalently, Eq. (2) shows that $p^X(t)$ is the probability vector of new cases in each state, observed across an interval $t - 6 \leq s \leq t$, divided by the total number of US cases across this period. Let $p^Y(t)$ be the analogous vector for deaths. As these probability vectors are suitably normalised, it is possible to compare them directly. Given two vectors $p, q \in \mathbb{R}^n$, let their L^1 distance be defined as $\|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$.

Next, we define a *search interval length* of $S = 50$. With this, let the offset between the multivariate time series be defined by the following function:

$$f_1 : [7, T - S] \cap \mathbb{Z} \to [0, S] \cap \mathbb{Z}; \tag{3}$$

$$t \mapsto \operatorname{argmin}_s \{\|p^X(t) - p^Y(t + s)\|_1 : s = 0, 1, \ldots, S\}. \tag{4}$$

That is, for any probability vector of (averaged) cases at time t, $f_1(t)$ is defined as the time at most $S = 50$ days in the future with the closest probability vector of (averaged) deaths. We remark that the domain of f_1 is restricted to $[7, T - S] \cap \mathbb{Z}$ to allow an entire search interval of $S = 50$ days for each t. Were this not included, the function would be trivially bounded and decrease to zero as t approached T.

2.2 Affinity matrix method

In this section, we estimate a continuously varying offset between multivariate time series $x_i(t)$ and $y_i(t)$ by comparing affinity matrices of counts between states. Let $D^X(t) \in \mathbb{R}^{n \times n}$ be the distance matrix between 7-day rolling averaged cases in each state, observed over an interval $[t-6, t]$. That is,

$$D_{ij}^X(t) = |\hat{x}_i(t) - \hat{x}_j(t)|, \quad i, j = 1, \ldots, n, \quad t = 7, \ldots, T. \tag{5}$$

Let $D^Y(t)$ be the analogous matrix for deaths. Given such a distance matrix D, we associate a $n \times n$ affinity matrix A by

$$A_{ij} = 1 - \frac{D_{ij}}{\max D}. \tag{6}$$

This is suitably normalised with all elements in $[0, 1]$ to allow direct comparison between different affinity matrices. Let $A^X(t)$ and $A^Y(t)$ be the affinity matrices corresponding to the distance matrices $D^X(t)$ and $D^Y(t)$, respectively. Given two matrices $A, B \in \mathbb{R}^{n \times n}$, let their L^1 distance be defined as $\|A - B\|_1 = \sum_{i,j=1}^{n} |A_{ij} - B_{ij}|$.

Again we use a *search interval length* of $S = 50$. With this, let the offset between the multivariate time series be defined by the following function:

$$f_2 : [7, T - S] \cap \mathbb{Z} \to [0, S] \cap \mathbb{Z}; \tag{7}$$

$$t \mapsto \operatorname{argmin}_s \{\|A^X(t) - A^Y(t + s)\|_1 : s = 0, 1, \ldots, S\}. \tag{8}$$

That is, for any affinity matrix between states' (averaged) cases at time t, $f_2(t)$ is defined as the time at most $S = 50$ days in the future with the closest affinity matrix between (averaged) deaths. Again, the domain of f_2 is restricted to $[7, T - S] \cap \mathbb{Z}$ to allow a complete search interval of $S = 50$ days. Were this not included, the function would be trivially bounded and decrease to zero as t approached T.

2.3 Inner product method

This section estimates a continuously varying offset between multivariate time series $x_i(t)$ and $y_i(t)$ via normalised inner products between individual states' time series. As before, we make use of the 7-day rolling averaged counts $\hat{x}_i(t)$ and $\hat{y}_i(t)$, but this time we restrict to one state at a time for our calculations. For the proceeding exposition, let $\hat{x}(t)$ and $\hat{y}(t)$ be the 7-day averaged counts of cases and deaths for a single candidate state.

Suppose $a \leq t \leq b$ and $c \leq t \leq d$ are two intervals within $[7, T]$ of equal length $L = b - a = d - c$. Let the normalised inner product between $\hat{x}(t)_{a \leq t \leq b}$ and $\hat{y}(t)_{c \leq t \leq d}$ be defined and notated as follows:

$$\langle \hat{x}(a:b), \hat{y}(c:d) \rangle_n = \frac{\sum_{t=0}^{L} \hat{x}(a+t)\hat{y}(c+t)}{\left(\sum_{t=a}^{b} \hat{x}(t)^2\right)^{\frac{1}{2}} \left(\sum_{t=c}^{d} \hat{y}(t)^2\right)^{\frac{1}{2}}}. \tag{9}$$

This normalised inner product is derived from the standard Euclidean inner product on $\mathbb{R}^{L+1}$. Indeed, for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{L+1}$, let $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{L+1} u_i v_i$. Then $\langle ., . \rangle$ is symmetric, bilinear and positive-definite; $\langle \mathbf{u}, \mathbf{u} \rangle = \sum_{i=1}^{L+1} u_i^2$ recovers the Euclidean norm on $\mathbb{R}^{L+1}$. We can re-express Eq. (9) as follows:

$$\langle \hat{x}(a:b), \hat{y}(c:d) \rangle_n = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{(\langle \mathbf{u}, \mathbf{u} \rangle)^{\frac{1}{2}} (\langle \mathbf{v}, \mathbf{v} \rangle)^{\frac{1}{2}}}, \tag{10}$$

where $\mathbf{u} = \hat{x}(t)_{a \leq t \leq b}$ and $\mathbf{v} = \hat{y}(t)_{c \leq t \leq d}$. That is, Eq. (9) presents a normalised analogue of the standard Euclidean inner product on $\mathbb{R}^{L+1}$.

We have chosen these normalised inner products to have maximal value 1 if and only if there is a proportionality relation $\hat{y}(t) = k\hat{x}(t + \tau)$ for all $t = c, \ldots, d$ for some constant $k > 0$ and offset τ. Indeed, we are seeking the offset in time where deaths are most closely proportional to cases. They are more suitable than other metrics, such as correlation or distance correlation [12]. Correlation or distance correlation would each return maximal value 1 if $y = kx + b$ for an additional constant b, which is unsuitable.

Next, we use a rolling window of length $L = 150$ days in which to compute a varying maximised offset. This longer window is chosen here in order to capture undulations in the time series, which are necessary for the inner product comparison to work well. Indeed, the inner product is maximised when local maxima and minima in cases are aligned with future local maxima and minima in deaths. Within each window, we again use a search interval length of $S = 50$. Then, let the offset between the univariate time series $\hat{x}_i(t)$ and $\hat{y}_i(t)$ for each state i be defined by the following function:

$$g_i : [7, T - L] \cap \mathbb{Z} \to [0, S] \cap \mathbb{Z}; \tag{11}$$

$$t \mapsto \mathrm{argmax}_\tau \{\langle \hat{x}_i(t : t + L - \tau), \hat{y}_i(t + \tau : t + L) \rangle_n : \tau = 0, 1, \ldots, S\}. \tag{12}$$

Effectively, this function considers the interval $[t, t + L]$ as fixed and selects an appropriate offset only by considering case and death counts within the interval. For that purpose, we must consistently truncate the case time series at the end, and the death time series at the beginning, hence the computation of the normalised inner product between $\hat{x}(s)_{t \leq s \leq t+L-\tau}$ and $\hat{y}(s)_{t+\tau \leq s \leq t+L}$.

Finally, the overall offset between the multivariate time series $\hat{x}_i(t)$ and $\hat{y}_i(t)$, $i = 1, \ldots, n$ is simply defined as

$$g : [7, T - L] \cap \mathbb{Z} \to \mathbb{R}; \tag{13}$$

$$g(t) = \frac{1}{n} \sum_{i=1}^{n} g_i(t). \tag{14}$$

Due to the averaging process, this is not necessarily integer-valued.

2.4 Vector comparison method

In this final methodological section, we estimate not only a continuously varying offset between multivariate time series of cases and deaths, but also a time-varying mortality rate. We proceed by directly comparing vectors of cases and deaths and attempting to minimise appropriate linear combinations thereof. We present multiple variations within this framework based on different "loss" functions—these record differences between vectors of cases and deaths, up to linear rescaling. As in Sect. 2.3, we are seeking an offset in time where deaths are most closely proportional to cases. Unlike Sect. 2.3, we use all states concurrently.

First, we define an L^1 1-day loss function as follows:

$$\mathcal{L}_1^1(t, \tau, \lambda) = \sum_{i=1}^{n} |\hat{x}_i(t) - \lambda \hat{y}_i(t + \tau)|. \tag{15}$$

We remark that λ plays the role of the inverse of the mortality rate between cases and deaths and is chosen for increased interpretability when plotting our results. Equivalently, we expect one death out of every λ cases (for an optimal λ).

There are three parameters we can vary in our loss function. First, we can use sums of squares (an L^2 difference) rather than the above L^1 difference. Second, rather than fixing a single day t, we could compute a loss function over a longer period of length P. For example, we define an L^1 P-day loss function as follows:

$$\mathcal{L}^1_P(t,\tau,\lambda) = \sum_{i=1}^{n}\sum_{j=0}^{P-1} |\hat{x}_i(t+j) - \lambda\hat{y}_i(t+j+\tau)|. \tag{16}$$

Third, we could modify the loss functions with a division term. For example, we define an L^1 1-day divided loss function as follows:

$$\mathcal{L}^1_{1,div}(t,\tau,\lambda) = \sum_{i=1}^{n} \frac{|\hat{x}_i(t) - \lambda\hat{y}_i(t+\tau)|}{|\hat{x}_i(t)|}. \tag{17}$$

We can also combine these modifications, for example using sums of squares in (16) and (17).

If we use L^2 differences, we have an analytically determined value of λ that minimises the function for any candidate τ. For example, consider the L^2 1-day loss function,

$$\mathcal{L}^2_1(t,\tau,\lambda) = \sum_{i=1}^{n} |\hat{x}_i(t) - \lambda\hat{y}_i(t+\tau)|^2. \tag{18}$$

The partial derivative with respect to λ is

$$2\sum_{i=1}^{n} \hat{y}_i(t+\tau)(\lambda\hat{y}_i(t+\tau) - \hat{x}_i(t)). \tag{19}$$

By minimising a quadratic, there exists a distinguished value

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \hat{y}_i(t+\tau)\hat{x}_i(t)}{\sum_{i=1}^{n} \hat{y}_i(t+\tau)^2} \tag{20}$$

that minimises $\mathcal{L}^2_1(t,\tau,\lambda)$ for fixed t and τ, and similarly for other L^2 loss functions.

With the framework of loss functions as defined above, we can now define the continuous time-varying offset and associated inverse mortality. Again, we use a search interval length of $S = 50$ days. For any candidate loss function $\mathcal{L}$, we define the following function:

$$\mathbf{h}_{\mathcal{L}} : [7, T-S] \cap \mathbb{Z} \to [0,S] \cap \mathbb{Z} \times \mathbb{R}^+ \subset \mathbb{R}^2 \tag{21}$$
$$t \mapsto \text{argmin}_{\tau,\lambda}\{\mathcal{L}(t,\tau,\lambda) : \tau = 0,\ldots,S, 1 \leq \lambda \leq 200\}. \tag{22}$$

We remark that $\mathbf{h}_{\mathcal{L}}$ effectively has two outputs. We write $\mathbf{h}_{\mathcal{L}}(t) = (\tau_{\mathcal{L}}(t), \lambda_{\mathcal{L}}(t)) \in \mathbb{R}^2$. Then, $\tau_{\mathcal{L}}(t)$ gives the time-varying offset between the multivariate time series, while $\lambda_{\mathcal{L}}(t)$ gives the continuously varying inverse mortality rate. For the L^2 loss functions, $\lambda_{\mathcal{L}}(t)$ can be determined analytically through Eq. (20). For the L^1 loss functions, the optimisation can be performed via a grid search. For the inverse mortality rate λ, we search over a closed bounded interval $[1, 200]$, corresponding to a search of mortality rate between 0.5 and 100%.

3 Results

Figures 1 and 2 show the determined time-varying offset for the probability vector and affinity matrix method, respectively. In these plots, the value of the function at a date index of 2020-03, for example, records the optimal offset τ between cases at 1 March 2020 and deaths

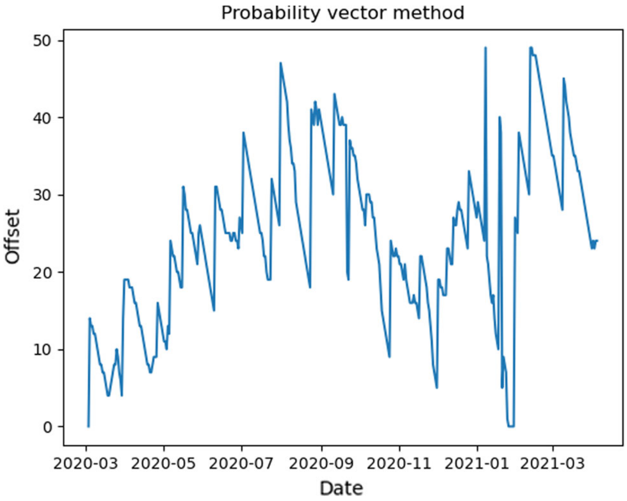


Fig. 1 Continuous time-varying offset $f_1(t)$ determined by the probability vector method, detailed in Sect. 2.1. A pattern of increase, decrease and then increase is observed. In order to accommodate the $S = 50$-day search window, the indexed dates end 50 days from the end of our analysis period

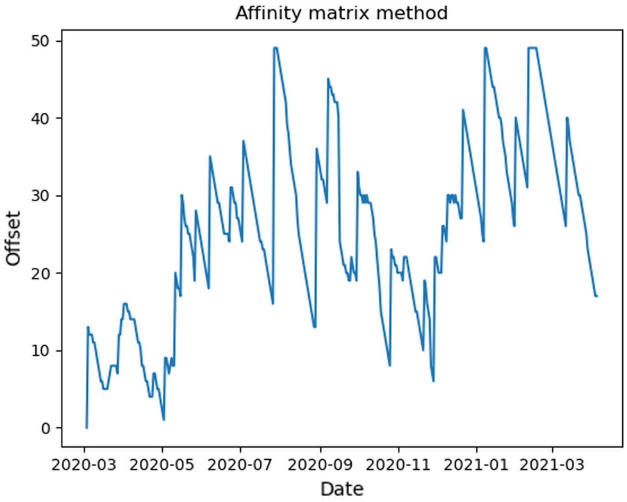


Fig. 2 Continuous time-varying offset $f_2(t)$ determined by the affinity matrix method, detailed in Sect. 2.2. A pattern of increase, decrease and then increase is observed. In order to accommodate the $S = 50$-day search window, the indexed dates end 50 days from the end of our analysis period

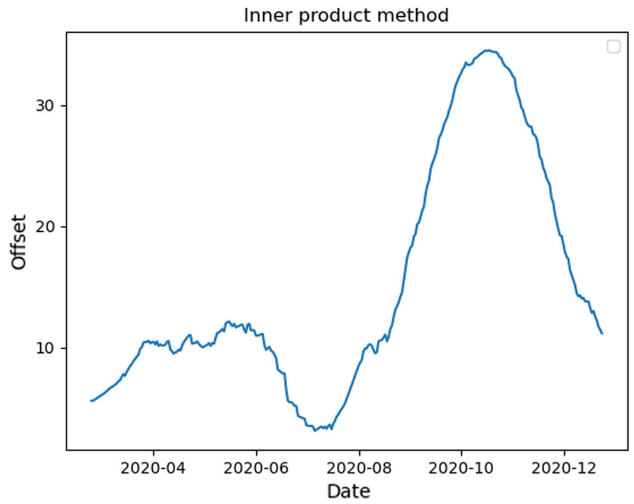
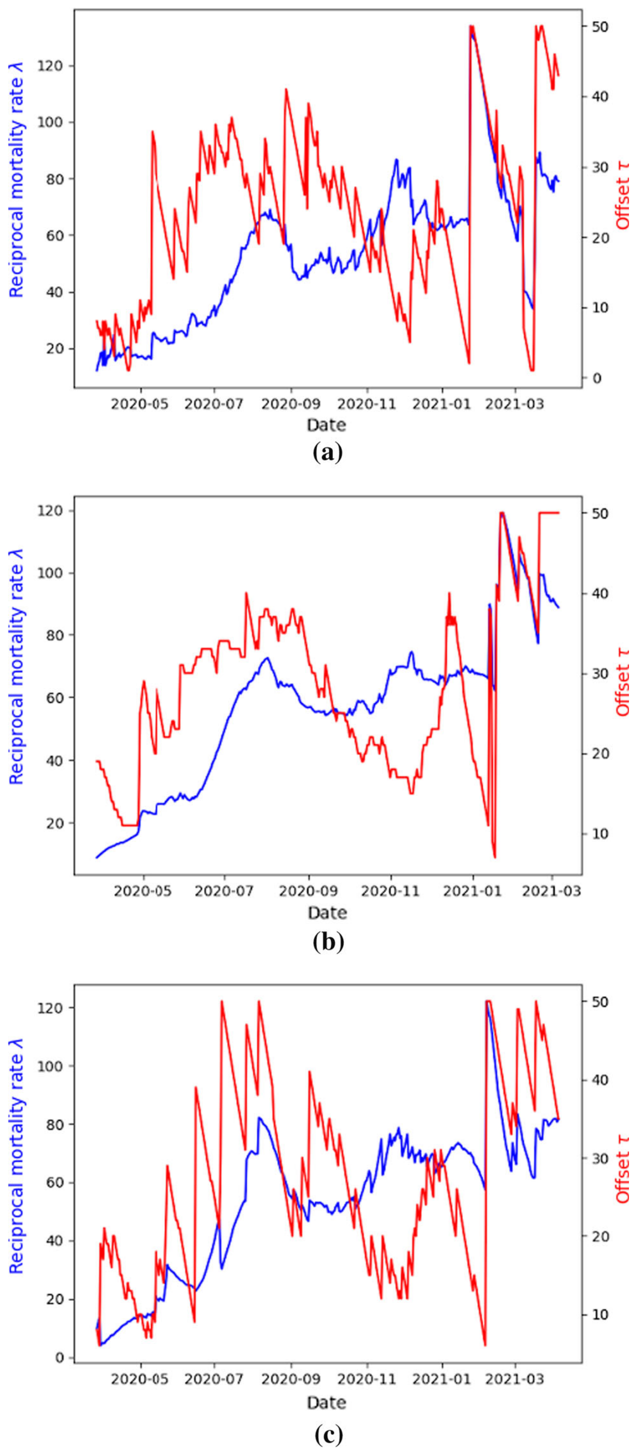


Fig. 3 Continuous time-varying offset $g(t)$ determined by the normalised inner product method, detailed in Sect. 2.3. A pattern of increase, decrease and then increase is observed. This function is substantially smoother than Figs. 1 and 2 due to the averaging in its definition (14). In order to accommodate the $L = 150$-day rolling computation window, the indexed dates end 150 days from the end of our analysis period

τ days later. Considerable similarity in these results is observed, which is to be expected, as both methods work similarly, by finding a future day in deaths with similar internal structure among states as a given day in cases. An "up-down-up" pattern is visible. Initially, the calculated offset during March 2020 is about 10 days. The offset rises to approximately 30 around September 2020 and then declines once more to 10–20 towards the end of 2020. Subsequently, an increase is observed to around 40, albeit with some irregularity during February 2021.

Figure 3 shows the offset for the normalised inner product method. This function is substantially smoother than the other offset functions in this paper due to the averaging in its definition (14). Again, an "up-down-up" pattern is observed but with consistently smaller values than the previous two methods. The determined offset rises from approximately 5 in March 2020 to 10, back down almost to zero, and up to a peak of over 30. We remark that the inner product method examines data $L = 150$ days in advance, while the other methods search data only $S = 50$ days in advance, so the determined offsets in Fig. 3 lead ahead of all the other figures. Aside from this, the inner product method is quite different to the other methods presented in this manuscript. Indeed, Sect. 2.3 shows how an offset is computed individually for each state, while every other method uses the entire multivariate data in conjunction.

In Fig. 4, we present three plots for three different loss functions within our vector comparison framework of Sect. 2.4. We make sure to trial some variation of all three available parameters in our loss function. In Figs. 4a–c, we use an L^1 1-day divided loss function,



(a)

(b)

(c)

Fig. 4 Several alternative time-varying offset functions computed within our framework of direct vector comparison, detailed in Sect. 2.4. To show a range of options, we present examples with every possible variation of parameters. In **a**, we use an L^1 1-day divided loss function. In **b**, we use an L^1 30-day loss function (without division). In **c**, we use an L^2 1-day loss function (without division), involving an analytically determined λ (as in (20)). A pattern of increase, decrease and then increase is observed in the offset τ, and a rather consistent increase in the reciprocal mortality rate λ

an L^1 30-day undivided loss function, and an L^2 1-day undivided loss function, respectively. In all three figures, we display both the time-varying offset τ and inverse mortality rate λ. These figures are quite consistent with Figs. 1 and 2 in the offset. Initially, the offset is consistently about 10 days, rising to 30, declining to 10–20, and dramatically rising to 40–50. Like Figs. 1 and 2, some irregularity is observed during January-February 2021. All three figures show a general increase in λ, signifying a consistent decrease in the mortality of COVID-19, at least with respect to observed cases and deaths. However, a brief period in reduction in λ is observed in the fall of 2020.

We remark that we could easily apply an averaging or smoothing procedure to the probability vector, affinity matrix or vector comparison methods (Figs. 1, 2, 4) to generate smoother curves like Fig. 3, but have chosen to display the initial raw result. In addition, smoothing could be applied for use in a predictive setting.

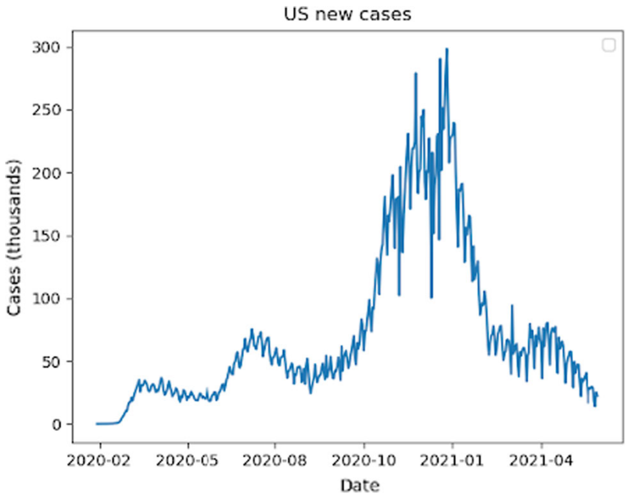


Fig. 5 New daily cases for the entire United States

4 Discussion

All four methods and six plots displayed thus far exhibit an “up-down-up” pattern in the estimated time-varying offset between case and death time series. Early on, a small offset is observed—this has several explanations. First, US states were slow to implement effective and wide-scale testing regimes [58], so cases were likely substantially underreported. Secondly, treatments were limited, thus infected patients may have passed away from infection within a quicker time frame. Third, due to the novelty of the virus, many vulnerable individuals such as the elderly may have contracted the disease early on and passed away relatively quickly. Later on, it is likely that vulnerable individuals took greater precautions than the rest of the population.

Subsequently, the offset increases until July–September, depending on the precise method. (The offset estimated by the inner product method (Fig. 3) peaks $\sim$ 3 months earlier, likely due to examining data $L = 150$ days in advance rather than $S = 50$ days in advance.) This could be attributed to improved treatment [59–62], non-pharmaceutical interventions, including social distancing, business closures, and better management of nursing homes, and more widespread testing.

Curiously, all methods observe a subsequent decrease in the estimated progression between cases and deaths. This decrease begins around August–September 2020 for the probability vector, affinity matrix and vector comparison methods (Figs. 1, 2, 4 respectively), and proportionately earlier for the inner product method. This period heralds a consistent worsening in the status of the pandemic throughout the US. As seen in Fig. 5, cases consistently rise from early September to the end of 2020. In addition, many states relax and do not reimpose lockdown measures during this time [63], and the colder climate yields worse outcomes both in terms of spread and illness [64]. The change in offset is not necessarily only due to individual progressions from infection to death, but involves mediating factors like stresses on hospital capacity. For example, perhaps initial waves of patients can be treated with ventilators, but these may quickly run out, causing more deaths from later cases.

The status of the pandemic changes drastically following the beginning of 2021. First, cases precipitously fall (Fig. 5), perhaps following the increased gathering of people over Thanksgiving and Christmas. Second, the rollout of vaccines produced at the end of 2020 [65,66] targeted vulnerable populations first and had a beneficial effect on the mortality of COVID-19 among the elderly. The drastic change in the status of the pandemic during this time could be the cause of the irregularity observed in several figures. The dramatically higher determined offset at the end of the time window, at least for Figs. 1, 2 and 4, is a welcome testament to the effectiveness of vaccines and the still improving treatment for unvaccinated individuals.

One strength in our paper is the fact that four different methods, including different loss functions within the vector comparison framework, yield relatively similar results. The loss functions in Sect. 2.4 allow variation of three parameters, which are all trialled at least once in the subfigures of Fig. 4. In particular, the choice of whether to implement divided loss functions, such as in (17), notably changes the properties of the loss function. In an undivided loss function such as (15), larger states with larger absolute values of $\hat{x}_i(t)$ and $\hat{y}_i(t)$ are likely to disproportionately influence the selection of τ and λ. In a divided loss function, this is no longer the case. It is a strength of the framework of Sect. 2.4 that this normalisation produces little difference in results.

Several limitations exist in this paper and even in future work. First, the framework of Sect. 2.4 implicitly assumes (and aims to find) a constant (inverse) mortality rate λ among all the states. While the US mostly has a similar standard of living and healthcare system quality from state to state, this is not uniformly the case, and different states differ substantially in population density and socioeconomic demographics. However, we believe that the remarks above, wherein two drastically different methods that prioritise larger states and all states, respectively, give similar results, show that perhaps this limitation is not too grave.

Second, it is notable that Fig. 3, defined by the inner product method of Sect. 2.3, is the one figure most different to the others. That is, it appears to be the odd one out relative to Figs. 1, 2, 4a–c. This difference is to be partially expected, as the inner product method works quite differently to the other methods. Namely, it examines $L = 150$ days in advance rather than $S = 50$ days, and computes an offset for each individual state rather than the multivariate time series as a whole. One may consider the outlier of Fig. 3 to be both a strength and limitation of the manuscript. It is potentially a limitation as it does not match the other figures exactly, but it may be a strength as it suggests that computing a separate offset for each state and simply averaging them is too naive a procedure. We remark that the inner product method is the simplest and the most closely related to (quite naive) existing methods of computing offsets between time series, such as cross-correlation [67]. In Sect. 2.3, we explained why our chosen normalised inner product is more suitable in this context than an offset correlation (essentially equivalent to the method of cross-correlations). That is, the outlier status of Fig. 3 may have a notable upshot: that a full consideration of the multivariate structure of the time series is necessary, and not simply an individual consideration of each state at a time.

Third, extending our analysis into the future may be difficult due to complexities in the epidemiology of COVID-19 and the availability of data. Specifically, the rollout of vaccines has created very different progressions from cases to deaths in the vaccinated vs unvaccinated populations. In addition, the Centers for Disease Control and Prevention (CDC) has changed its reporting of cases among the vaccinated population, only tracking "breakthrough cases" that result in hospitalisation or death [68]. Future work could use the analysis presented in this paper, but more precise data needs to be collected and made available on an ongoing basis. More broadly, we encourage future work to carefully separate out the mathematical epidemiology of COVID-19 between vaccinated and unvaccinated populations, studying phenomena not limited to the offset between cases and deaths, and further exploring the positive impact of COVID-19 vaccines on the community. For example, future work could separate out the progression from COVID-19 infection to either death or recovery among the vaccinated and unvaccinated populations, including a consideration of "long Covid" [69]. There may be numerous non-trivial benefits to be discovered with careful analysis. At the same time, as near-entire vaccination of the US population seems unlikely (that is, the US is unlikely to reach herd immunity), measures to contain and reduce the impact of the virus on the healthcare system remain highly relevant for the reduction of casualties and economic and other social consequences [70, 71].

5 Conclusion

Overall, we have proposed four methods to determine a continuously varying offset between two multivariate time series and applied this to the state-by-state counts of COVID-19 cases and deaths in the United States. Our final method is a framework of loss functions in which we have trialled the variation of several parameters. Our methods exhibit considerable robustness with broadly similar results obtained, including under relatively substantial changes such as normalising by case counts in Sect. 2.4 to de-prioritise larger states. Our findings reveal new insights into the time-varying progression from cases to deaths in the US and discuss how this reflects the changing status of the pandemic. We show that the estimated offset between cases and deaths rises between the first and second waves of COVID-19 in the US, falls towards the end of 2020, and dramatically rises in 2021. Minor modifications such as smoothing, combined with updated and reliable data on cases among the vaccinated and unvaccinated populations could provide a valuable predictive tool regarding future periods of high load on the healthcare system. Our analysis could also be applied to other multivariate time series outside epidemiology.

Acknowledgements The authors would like to thank Howard Bondell for helpful discussions.

Data availability This manuscript has associated data in a data repository. [Authors' comment: The data we use can be found at Ref [72], https://www.github.com/nytimes/covid-19-data.]

References

1. D.M. Cutler, L.H. Summers, JAMA **324**, 1495 (2020)
2. L. Rosenbaum, N. Engl. J. Med. **382**, 2368 (2020)
3. J.A.T. da Silva, Curr. Res. Behav. Sci. **2**, 100014 (2021)
4. G. Pullano et al., Nature (2020)
5. M.D. Bari, D. Balzi, G. Carreras, G. Onder, Front. Med. **7**, 402 (2020)
6. J. Miller, C. Copley, B.H. Meijer, *Countries turn to rapid antigen tests to contain second wave of COVID-19, Reuters, October 15, 2020*
7. *Prise en charge par les médicins de ville des patients atteints de COVID-19 en phase de déconfinement*, Ministère des Solidarités et de la Santé, May 13, 2020
8. H.W. Hethcote, SIAM Rev. **42**, 599 (2000)
9. G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, Phys. Life Rev. **18**, 66 (2016)
10. A. Vazquez, Phys. Rev. Lett. **96**, 038702 (2006)
11. R. Moeckel, B. Murray, Phys. D **102**, 187 (1997)
12. G.J. Székely, M.L. Rizzo, N.K. Bakirov, Ann. Stat. **35**, 2769 (2007)
13. C.F. Mendes, M.W. Beims, Phys. A **512**, 721 (2018)
14. C.F.O. Mendes, R.M. da Silva, M.W. Beims, Phys. Rev. E **99**, 06220 (2019)

15. K. Shang, B. Yang, J.M. Moore, Q. Ji, M. Small, Chaos Interdiscip. J. Nonlinear Sci. **30**, 041101 (2020)
16. G. Bonifazi, L. Lista, D. Menasce, M. Mezzetto, D. Pedrini, R. Spighi, A. Zoccoli, Eur. Phys. J. Plus **136**, 1–4 (2021)
17. L. Wynants et al., BMJ m1328 (2020)
18. E. Estrada, Phys. Rep. **869**, 1 (2020)
19. N.S. Barlow, S.J. Weinstein, Phys. D **408**, 132540 (2020)
20. S.J. Weinstein, M.S. Holland, K.E. Rogers, N.S. Barlow, Phys. D **411**, 132633 (2020)
21. K.Y. Ng, M.M. Gui, Phys. D **411**, 132599 (2020)
22. C. Vyasarayani, A. Chatterjee, Phys. D **414**, 132701 (2020)
23. M. Cadoni, G. Gaeta, Phys. D **411**, 132626 (2020)
24. A.G. Neves, G. Guerrero, Phys. D **413**, 132693 (2020)
25. A. Comunian, R. Gaburro, M. Giudici, Phys. D **413**, 132674 (2020)
26. A. Abidemi, Z.M. Zainuddin, N.A.B. Aziz, Eur. Phys. J. Plus **136**, 1–35 (2021)
27. A. Ballesteros, A. Blasco, I. Gutierrez-Sagredo, Phys. D **413**, 132656 (2020)
28. S. Liu, M.Y. Li, Phys. D **422**, 132903 (2021)
29. N.M. Gatto, H. Schellhorn, Math. Biosci. **333**, 108539 (2021)
30. C. Manchein, E.L. Brugnago, R.M. da Silva, C.F.O. Mendes, N.W. Beims, Chaos Interdiscip. J. Nonlinear Sci. **30**, 041102 (2020)
31. B. Blasius, Chaos Interdiscip. J. Nonlinear Sci. **30**, 093123 (2020)
32. B.K. Beare, A.A. Toda, Phys. D **412**, 132649 (2020)
33. M. Perc, N.G. Miksić, M. Slavinec, A. Stožer, Front. Phys. **8**, 127 (2020)
34. A. Gowrisankar, L. Rondoni, S. Banerjee, Eur. Phys. J. Plus **135**, 1–9 (2020)
35. D. Manevski, N.R. Gorenjec, N. Kejžar, R. Blagus, Math. Biosci. **329**, 108466 (2020)
36. B. Gross et al., EPL (Europhys. Lett.) **131**, 58003 (2020)
37. A. Maiti et al., Sustain. Cities Soc. **68**, 102784 (2021)
38. R. Paul, A.A. Arif, O. Adeyemi, S. Ghosh, D. Han, J. Rural Health **36**, 591 (2020)
39. N. James, M. Menzies, Phys. D **425**, 132968 (2021)
40. N. James, M. Menzies, L. Azizi, J. Chan, Phys. D **412**, 132636 (2020)
41. A. Karaivanov, PLoS One **15**, e0240878 (2020)
42. J. Ge, D. He, Z. Lin, H. Zhu, Z. Zhuang, Math. Biosci. **330**, 108484 (2020)
43. L. Xue, S. Jing, J.C. Miller, W. Sun, H. Li, J.G. Estrada-Franco, J.M. Hyman, H. Zhu, Math. Biosci. **326**, 108391 (2020)
44. F. Saldaña, H. Flores-Arguedas, J.A. Camacho-Gutiérrez, Math. Biosci. Eng. **17**, 4165 (2020)
45. A. Danchin, G. Turinici, Math. Biosci. **331**, 108499 (2021)
46. J.A.T. Machado, A.M. Lopes, Nonlinear Dyn. (2020)
47. N. James, M. Menzies, P. Radchenko, Chaos Interdiscip. J. Nonlinear Sci. **31**, 031105 (2021)
48. C.N. Ngonghala, E.A. Iboi, A.B. Gumel, Math. Biosci. **329**, 108452 (2020)
49. J. Cavataio, S. Schnell, Math. Biosci. **333**, 108545 (2021)
50. N. James, M. Menzies, Chaos Interdiscip. J. Nonlinear Sci. **31**, 083116 (2021)
51. L.Ó. Náraigh, Á. Byrne, Math. Biosci. **330**, 108496 (2020)
52. D.H. Glass, Math. Biosci. **330**, 108472 (2020)
53. N. James, M. Menzies, Phys. D **417**, 132809 (2021)
54. N. James, M. Menzies, Chaos Interdiscip. J. Nonlinear Sci. **30**, 091102 (2020)
55. Y. Zhou et al., Harvard Data Sci. Rev. (2020)
56. Y. Wang, Y. Liu, J. Struthers, M. Lian, Clin. Infect. Dis. **72**, 643 (2020)
57. N. James, M. Menzies, H. Bondell, EPL (Europhys. Lett.) (2021)
58. M.D. Shear, A. Goodnough, S. Kaplan, S. Fink, K. Thomas, N. Weiland, *The lost month: How a failure to test blinded the U.S. to Covid-19*, The New York Times, March 28, 2020. https://www.nytimes.com/2020/03/28/us/testing-coronavirus-pandemic.html
59. M. Wang et al., Cell Res. **30**, 269 (2020)
60. E.M. Bloch, Blood **136**, 654 (2020)
61. X. Xu et al., Proc. Natl. Acad. Sci. **117**, 10970 (2020)
62. B. Cao et al., N. Engl. J. Med. **382**, 1787 (2020)
63. M. Iati et al., *All 50 U.S. states have taken steps toward reopening in time for Memorial Day weekend, The Washington Post, May 20, 2020*
64. S. Mallapaty, Nature **586**, 653 (2020)
65. F.P. Polack et al., N. Engl. J. Med. **383**, 2603 (2020)
66. E.E. Walsh et al., N. Engl. J. Med. **383**, 2439 (2020)
67. A. Papoulis, *The Fourier Integral and its Applications* (McGraw-Hill, New York, 1962)
68. K.J. Wu, *What breakthrough infections can tell us*, The Atlantic, May 28, 2021. https://www.theatlantic.com/science/archive/2021/05/tracking-breakthrough-infections/619027/
69. E. Mahase, BMJ m2815 (2020)
70. V. Priesemann et al., Lancet **397**, 92 (2021)
71. S. Momtazmanesh et al., Am. J. Trop. Med. Hyg. **102**, 1181 (2020)
72. Coronavirus (Covid-19) data in the United States (2021), The New York Times. https://www.github.com/nytimes/covid-19-data. Accessed 24 July 2021