Count-valued time series models for COVID-19 daily death dynamics

William R Palmer | Richard A Davis | Tian Zheng

Department of Statistics, Columbia University,
New York, New York, USA

Correspondence
William R Palmer, Department of Statistics,
Columbia University, 1255 Amsterdam Ave.,
New York, NY, 10027, USA.
Email: wrp2110@columbia.edu

We propose a generalized non-linear state-space model for count-valued time series of COVID-19 fatalities. To capture the dynamic changes in daily COVID-19 death counts, we specify a latent state process that involves second-order differencing and an AR(1)-ARCH(1) model. These modelling choices are motivated by the application and validated by model assessment. We consider and fit a progression of Bayesian hierarchical models under this general framework. Using COVID-19 daily death counts from New York City's five boroughs, we evaluate and compare the considered models through predictive model assessment. Our findings justify the elements included in the proposed model. The proposed model is further applied to time series of COVID-19 deaths from the four most populous counties in Texas. These model fits illuminate dynamics associated with multiple dynamic phases and show the applicability of the framework to localities beyond New York City.

KEYWORDS
Bayesian estimation, count-valued time series, COVID-19 modelling, state-space models

1 INTRODUCTION

Since early 2020, the COVID-19 pandemic has posed a sustained threat to public health across the globe. With more than 100 million confirmed cases and 2.15 million confirmed deaths across 192 territories (Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, 2020), the infectious respiratory disease has spread, sickened, and killed on an unprecedented scale.

The staggering global totals arise from local, transient outbreaks, in which disease spreading contacts, new cases, and subsequent deaths are concentrated in space and time. Long range human mobility creates contacts across the globe, and a few such links can carry an epidemic between distant cities. Local contacts, however, drive the outbreaks of COVID-19, creating waves of cases and then, later on, waves of hospitalizations and deaths.

In late February and early March, following the original outbreak in China and amid the escalating situation in Europe, COVID-19 quietly took hold in New York City (NYC). At the time of the city's first confirmed case on March 1, it is estimated that there were over 10,000 infections (Carey & Glanz, 2020). Control measures were slow to be put in place, with bars, restaurants, and schools staying open for the next 2 weeks, and no public face covering requirement until April 15. NYC quickly became the epicentre of the pandemic, with hospitals stretched, intensive care units (ICUs) filling up, and sickened New Yorkers dying at an alarming rate.

In this paper we assess how a dynamic model can track the trends in observed COVID-19 mortality data. We propose a non-linear state-space time series model that is fitted to daily COVID-19 deaths in NYC, and we evaluate forecasts of held-out, look-ahead counts. Moreover, we fit and assess a succession of models within the same framework, where the progression from simpler to more complex is driven by the behaviour we wish to capture—volatility in the region of high curvature and the subsequent settling following the observed peak. We use second-order differencing to model these dynamics in order to de-trend the observed series and capture the smoothness in the daily deaths curvature.

Mathematical epidemiological models can generally be characterized as either mechanistic or 'phenomenological' (Chowell et al., 2016). Existing models of COVID-19 deaths fall in both categories (Altieri et al., 2020; Cooper et al., 2020; Harvey & Kattuman, 2020; He et al., 2020). Our
model falls into the second category, using observed historical data to make predictions, without explicitly accounting for any epidemiological mechanisms.

Good data alongside good modelling is essential for successful COVID-19 forecasting (Jewell et al., 2020; Chin et al., 2020). We use deaths reported by the NYC Department of Health New York City Department of Health and Mental Hygiene (2020). These deaths are confirmed by the City's Office of the Chief Medical Examiner and the Health Department's Bureau of Vital Statistics, a measure that can improve uniformity and reliability (Pappas, 2020). The counts include decedents who were NYC residents and had a positive laboratory test for the coronavirus. Figure 1 shows these confirmed counts in the five NYC boroughs in the spring of 2020.

This paper describes the statistical methods we employ and presents the results of applying these methods to observed COVID-19 death counts. Section 2.1 introduces the modelling framework, and Section 2.2 specifies the general form and progression of our considered models. Section 2.3 details the tools for predictive model assessment. In Section 3.1 we compare the progression of model fits to NYC data. Sections 3.2 and 3.3 further examine fits from the proposed model. Section 3.2 shows posterior estimates around the observed April peaks for all five NYC boroughs, whereas Section 3.3 shows fits to longer series of observed deaths from the four most populous Texas counties.

## 2 METHODS

We take a non-linear state-space model approach to modelling daily death counts. This approach fits nicely into a hierarchical modelling framework for which Bayesian estimation methods are accessible. Specifically, we use the Monte Carlo methods (Carpenter et al., 2017; Stan Development Team, 2020) for fitting our model, obtaining posterior samples of the model parameters. Out-of-sample look-ahead predictions are then found by drawing from posterior predictive distributions.

### 2.1 Model framework

Let \( y_t \) be the observed daily count on Day \( t \). We assume that \( y_t \) given \( \lambda_t \) follows a parametric distribution for count data with mean \( \lambda_t \). The dynamics of the time-varying mean \( \lambda_t \) is specified through a latent state process \( s_t \), which updates according to a density that depends on the previous state \( s_{t-1} \).

Let \( p(y_t|\lambda_t, \beta) \) and \( p(s_t|s_{t-1}, \beta) \) denote the conditional observational density and the state update density, respectively, where \( \beta \) includes model parameters that do not vary with time. To complete the Bayesian specification, we assign prior distributions \( p(\beta) \) and \( p(s_1, \ldots, s_{t_0}) \) to the conditioning parameters \( \beta \) and the initial states \( s_1, \ldots, s_{t_0} \), respectively. For second-order differencing we must initialize \( t_0 = 3 \) states. The model specification is

\[
\begin{align*}
  y_t & \sim p(y_t|\lambda_t, \beta), \quad t = 1, \ldots, T \\
  \lambda_t & \mapsto s_t, \quad t = 1, \ldots, T \\
  s_t & \sim p(s_t|s_{t-1}, \beta), \quad t = t_0 + 1, \ldots, T \\
  (s_1, \ldots, s_{t_0}) & \sim p(s_1, \ldots, s_{t_0}) \\
  \beta & \sim p(\beta),
\end{align*}
\]

where \( \lambda_t \mapsto s_t \) means that \( \lambda_t \) is encoded in the latent state \( s_t \). Given \( T \) observations \( y_{1:T} = (y_1, \ldots, y_T) \), we can fit this generative model using Monte Carlo methods and obtain estimates of the posterior distribution \( p(s_{1:T}, \beta|y_{1:T}) \).

We make \( k \)-day ahead forecasts based on the posterior predictive distribution

\[
p(y_{T+1:k}|y_{1:T}) = \int p(y_{T+1:k}|s_{T+1:k}, \beta) \prod_{t=T}^{T+k} p(s_{t+1}|s_t, \beta) p(s_1, \ldots, s_{t_0}) d\beta.
\]
It is straightforward to sample from this distribution using the posterior samples from $p(s_T, \beta | y_{1:T})$. In particular we can draw

$$
(s_T, \beta) \sim p(s_T, \beta | y_{1:T})
$$

$$
s_{T+\epsilon} \sim p(s_{T+\epsilon} | s_{T+\epsilon-1}, \beta) \quad \epsilon = 1, \ldots, k
$$

$$
y_{T+k} \sim p(y_{T+k} | s_{T+k}, \beta).
$$

A set of draws $y_{T+k}^{(1)}, \ldots, y_{T+k}^{(B)}$ from (3) approximates the predictive performance of the model. We assess model fit by considering held-out $k$-day ahead predictions relative to these posterior predictive draws.

### 2.2 Models considered

Within this generalized state-space, parameter-driven modelling framework (Davis et al., 1999), we consider and fit a sequence of models. These models vary in their observation equation $p(y_t | \lambda_t, \beta)$ and state update density $p(s_t | s_{t-1}, \beta)$. For the former, we begin with the most natural and simple choice, the Poisson distribution, and later compare with the negative binomial distribution. The state update density is specified according to an underlying process for the second-order difference of $\log \lambda_t$, which we denote by $\delta_t$. For $t > t_0$. The model is

$$
p(s_t | s_{t-1}, \beta) = p(\delta_t | \delta_{t-1}, \rho, a_0, a_1)
$$

$$
\delta_t := \nu^2 \log \lambda_t = (\log \lambda_t - \log \lambda_{t-1}) - (\log \lambda_{t-1} - \log \lambda_{t-2})
$$

$$
\delta_t = \rho \delta_{t-1} + \frac{\sqrt{a_0 + a_1 \delta_{t-1}^2}}{\nu} \epsilon_t \quad \epsilon_t \sim \text{Normal}(0, 1).
$$

Under the model, $\delta_t = \nu^2 \log \lambda_t$ is assumed to follow an AR(1)-ARCH(1) process. This parametric form of the state equation accommodates our beliefs about how second-order differences drive the dynamics of the observed daily counts. Note that because $\lambda_t$ must be positive, it is natural to work with the logarithm of the underlying mean, $\log \lambda_t$. In a wave of COVID-19 death counts, we expect the $\delta_t$’s to have high lag 1 autocorrelation and settle towards 0 following the regime of high curvature around the peak. Our state equation can capture this behaviour—if $\rho \in (0, 1)$ and $a_1 > 0$, we have $\delta_t$ negatively autocorrelated, and $\lim_{t \to \infty} \mathbb{E}[\delta_t | \delta_t] = 0$. Moreover, $\delta_t$ is more concentrated around 0 when $|\delta_{t-1}|$ is small.

The $\delta_t$ update depends on $\delta_{t-1}$ whenever at least one of $\rho$ or $a_1$ is nonzero. In this case, $t_0 = 3$, and for $t \geq 3$, we explicitly define $s_t$ as the vector of length three containing $\log \lambda_t$, the first order difference $\log \lambda_t - \log \lambda_{t-1}$, and the second-order difference $\delta_t$. Because of the twice-differencing, three dimensions are necessary to satisfy the second and third lines of (1). If $\rho$ and $a_1$ are both 0, then $\delta_t$ is white noise, $t_0 = 2$, and $s_t$ need only contain $\log \lambda_t$ and $\log \lambda_t - \log \lambda_{t-1}$.

The update step (4) for $\delta_t$ has the general form of a first order autoregressive process with ARCH(1) errors. Fixing the AR(1) and ARCH(1) terms ($\rho$ and $a_1$, respectively) gives rise to boundary cases and simpler models: $a_1 = 0$ gives iid noise; $\rho = 1$ gives a random walk. We consider a sequence of models that vary in their treatment of $\rho$ and $a_1$, and grow in complexity. Table 1 shows the progression of models. Note that when $a_1$ is set to 0, we reparameterize with $\sigma = \sqrt{a_0}$ for notational ease.

In our simplest model, $P(0,0)$, we fix $\rho$ and $a_1$ equal to 0, giving the white noise process for $\delta_t$ discussed above. Keeping $a_1 = 0$, we next consider a random walk for $\delta_t$, setting $\rho = 1$ in $P(1,0)$. In $P(0,0)$ we introduce a prior on $\rho$ with support $[0, 1]$. Conditionally, $\delta_t$ now follows a stationary AR(1) process. In $P(\rho, a_1)$, we set priors for both $\rho$ and $a_1$, so that conditionally $\delta_t$ follows a stationary AR(1)-ARCH(1) process. This is the most complicated state equation we consider. We take this one step further in the NB($\rho$, $a_1$) case by introducing a negative binomial observation distribution with scale parameter $\phi$, parameterized so that it has mean $\lambda_t$ and variance $\lambda_t + \lambda_t^2 / \phi$.

Each model in Table 1 has associated priors for the $\beta$ parameters and the initial states $s_1, \ldots, s_{t_0}$. In Appendix A1 we provide details of the full joint distribution for NB($\rho$, $a_1$). With the full joint specified, we can use MCMC methods to fit each model. In Appendix B1 we give the stan model code used to fit NB($\rho$, $a_1$).
2.3 Model assessment

We perform predictive model assessment in several ways based on posterior predictive draws \( y_{T+k} \sim p(y_{T+k}|y_1:T) \) sampled according to (3). These evaluations allow us to meaningfully compare the models in Table 1.

We make point and interval forecasts by taking quantiles of posterior predictive draws. We produce forecasts and compare to observed data. As a summary measure of predictive performance, we consider the absolute relative error of 7-day ahead cumulative death predictions

\[
\left| \frac{\text{Median} \left( \sum_{t=1}^{7} y_{T+k} | y_1:T \right) - \sum_{t=1}^{7} y_{T+k} }{\sum_{t=1}^{7} y_{T+k}} \right|.
\]

The other predictive model assessments we perform evaluate model fit using the full set of posterior predictive draws.

2.3.1 Probability integral transform

The probability integral transform (PIT) is a useful tool generally used for assessing models for continuous data; however, it can be extended to count-valued data. We use the nonrandomized PIT histogram for count data proposed in (Czado et al., 2009) for assessing the fits of our considered models.

We fit our model to observed counts \( y_1:T \) and make a \( k \)-day ahead prediction by sampling posterior predictive draws, \( y_{T+k}^{(1)}, \ldots, y_{T+k}^{(B)} \) as in (3). We seek a probability integral transform to assess this empirical predictive distribution against future observations \( y_{T+k} \sim F_{T+k} \). The empirical predictive distribution function \( \hat{F}_{B,T+k}(k) \) is given by

\[
\hat{F}_{B,T+k}(k) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{y_{T+k}^{(b)} \leq k\}}. \tag{6}
\]

If we successfully target the true predictive distribution, so that \( y_{T+k}^{(1)}, \ldots, y_{T+k}^{(B)} \) are independent and identically distributed \( y_{T+k} \sim F_{T+k} \), then \( \lim_{B \to \infty} \hat{F}_{B,T+k}(k) = F_{T+k}(k) \), and it follows that the PIT using \( \hat{F}_{B}(y_{T+k}) \) will be very close to Uniform on the unit interval for \( B \) large. For each model fit, we make \( B = 4000 \) post-warmup posterior draws.

2.3.2 Scoring rules

Scoring rules are another tool for predictive model assessment, providing a numerical score \( s(P, y) \) based on a predictive distribution \( P \) and an observation \( y \). We consider two scoring rules, the logarithm score and the ranked probability score. These scoring rules are strictly proper (Czado et al., 2009). We average the scores over sets of look-ahead predictive distributions to compare the considered models.

The logarithm score is defined as \( \log(P, y) = -\log p(y) \), where \( p(y) \) is the probability of observing \( y \) under the predictive distribution \( P \). We adapt this score to our framework for a \( k \)-day ahead forecast as \( \log(P_{B,T+k}, y_{T+k}) = -\log \hat{P}_{B,T+k}(y_{T+k}) \), where

\[
\hat{P}_{B,T+k}(k) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{y_{T+k}^{(b)} \leq k\}} + \frac{1}{B^2}
\]

is an approximation of the true posterior predictive probability mass function. This is similar to (6) above. We add the \( 1/B^2 \) term to avoid \( \log(P_{B,T+k}, y_{T+k}) = -\infty \) when \( y_{T+k}^{(b)} \neq y_{T+k} \) for all \( b \in \{1, \ldots, B\} \).

The ranked probability score is defined as \( \text{rps}(P, y) = \sum_{k=1}^{\infty} |P(k) - \mathbb{1}_{\{y \leq k\}}|^2 \), where \( P(\cdot) \) is the cumulative distribution function under \( P \). We take \( P(k) = \hat{F}_{B}(y_{T+k}) \), from (6). For both scoring rules, smaller values indicate a better fit.

3 RESULTS

We fit the models in Table 1 to multiple subsets of the NYC data. Each daily count time series corresponds to a single borough and begins with the day of the first COVID-19 death observed in that borough. Model assessment is based on a collection of these fits and their forecasts. Comparing results across the considered models, we determine \( \text{NB}(\rho, \alpha) \) to be generally best at capturing the observed dynamics in NYC’s daily COVID-19 death counts over various ranges of the data. We further fit the \( \text{NB}(\rho, \alpha) \) model to daily COVID-19 deaths data from the four most populous counties in Texas.

3.1 Model comparison

Figures 2 and 3 illustrate how the progression of state equation specifications (\( \delta_t \) updates) captures the dynamics of observed daily deaths around their peak. Here the models from Table 1 are fit to six ranges of Queens data. Figure 2 shows posterior estimates of the underlying time-varying mean \( \lambda_t \), along with observed deaths, whereas Figure 3 shows posterior estimates of \( \delta_t \).

Each subfigure corresponds to a different model fit. The rows align with Table 1 and each column corresponds to a different training series, where \( T \) is the number of training points. We indicate where the final training day falls relative to the observed peak. For the top left subfigure, \( T = 25 \) days of deaths are included in the training series, starting from the first confirmed COVID-19 death in Queens on March 11 and ending...
FIGURE 2  Daily COVID-19 deaths in Queens along with estimates of underlying means $\lambda_t$. Red shaded regions correspond to out-of-sample forecasts. Estimates vary by model and the length of training series ($T$).

FIGURE 3  Estimates of $\delta_t$, the second-order difference of the logarithm of the underlying mean. Red shaded regions correspond to out-of-sample forecasts. Estimates vary by model and the length of training series ($T$).

with the count from April 4, 2 days before the observed peak. We plot approximate 90% posterior intervals and approximate posterior medians based on Monte Carlo sampling. The red shaded regions show 3 weeks of out-of-sample, look-ahead forecasts.

The first row of Figures 2 and 3 show estimates based on $P(0,0)$. Under $P(0,0)$, $\delta_t$ is uncorrelated with $\delta_{t-1}$ and $\mathbb{E}[\delta_t | \delta_{t-1}] = 0$. These model features lead to poor forecasting around the observed peak, where second-order differences are negative with lag 1 autocorrelation. Although $P(0,0)$ fails to capture the dynamics close to the peak, this model forecasts well when the training series extends at least 18 days past the peak.

Under $P(1,0)$, we have that $\mathbb{E}[\delta_{T+k} | y_{1:T}] = \mathbb{E}[\delta_T | y_{1:T}]$ for $k \geq 1$. This feature is seen clearly in Figure 3, where the $P(1,0)$ posterior predictive distributions for $\delta_{T+k}$ remain centred around the posterior mean of $\delta_T$. In the first four columns, forecasts for $\delta_t$ stay negative, leading to a too quick decrease in the corresponding $\lambda_t$ forecasts.

Under $P(\rho,0)$ forecasts of $\delta_t$ increase gradually towards 0. This captures the dynamics of the observed counts much more closely than the previous two models. However, although predictive distributions for $\delta_{T+k}$ become centred around 0, there is still much uncertainty, and this seems to exaggerate the possibility of $\delta_t$ moving away from 0. Inclusion of the ARCH(1) term corrects for this. In the $P(\rho, \alpha_1)$ and $NB(\rho, \alpha_1)$ models, there is more uncertainty in the second and third columns where $\delta_T$ is negative, and less in the fifth and sixth where $\delta_T$ is close to 0. The final two rows in Figures 2 and 3, corresponding to models $P(\rho, \alpha_1)$ and $NB(\rho, \alpha_1)$, are very similar. This state update model is not able to properly predict the peak in the far left column, but performs well in the period after the peak.
Table 2 shows mean absolute relative errors (5) along with their standard errors. We average over the five NYC boroughs as well as the week-long range based on where the final training day falls relative to the observed peak. Each cell summarizes 5 × 7 = 35 values.

The P(1,0) model has the lowest absolute relative error when the final training day falls within 3 days of the observed peak. This follows because absolute errors are bounded for underestimates, and the forecasts from P(1,0) decrease quickly to 0. The NB(\(\rho, \alpha\)) model has the lowest error when the final training day falls 4 to 17 days after the observed peak, whereas the P(0,0) model performs best later on.

Figure 4 compares the nonrandomized PIT histograms (Section 2.3.1) for 1-day ahead predictions. Each histogram is based on 5 × 7 = 35 model fits, with the 28 training series from each borough covering the same four peak-relative weeks in Table 2. The histograms corresponding to the P(0,0), P(1,0), and P(\(\rho, 0\)) appear skewed. Although we see some overdispersion for NB(\(\rho, \alpha\)), the PITs for P(\(\rho, \alpha\)) and NB(\(\rho, \alpha\)) do not provide strong evidence that these models are incorrect.

Table 3 shows two proper scoring rules (Section 2.3.2) for the progression of models considered. These scores are averaged over the 21 days of forecasting for each model fit, the peak-relative week of the final training day, and the five boroughs. Each entry is an average of 21 × 7 × 5 = 735 values.

According to these measures for predictive model assessment, the models with both the AR(1) and ARCH(1) terms dominate when the final training day is 4 to 10 days after the peak, with NB(\(\rho, \alpha\)) performing the better of the two. The P(\(\rho, \alpha\)) and NB(\(\rho, \alpha\)) models have logarithm scores close to the lowest for the other peak-relative weeks, and NB(\(\rho, \alpha\)) has the lowest ranked probability score when the final training day is 18 to 24 days after the peak. We again see how P(0,0) performs well relative to the others in the later weeks, but very poorly when the final training day is close to the peak.

3.2 Fits around peak for all five boroughs

Figures 5 and 6 show further posterior estimates from NB(\(\rho, \alpha\)). We fit the model to counts from all five NYC boroughs, with final training days concentrated around the 2 weeks following the observed peaks. Again we plot approximate 90% posterior intervals and approximate posterior medians for \(\lambda_t\) and \(\delta_t\), which are based on posterior samples.
The trajectories of the posterior estimates of $\delta_t$ show a similar pattern across the five boroughs—an increase towards 0 with less posterior predictive uncertainty as more training points are included. For Brooklyn, Queens, and The Bronx there is a temporary flattening out or even decrease in $\delta_t$ around 2 weeks before the peak. For Brooklyn and The Bronx, the peak of the posterior $\lambda_t$ estimates falls before the observed peak, whereas for Queens there appears to be a sort of double observed peak, and the maximum posterior $\lambda_t$ is closer to the first.

In the far left hand column in Figure 5, we see a failure to recognize that the peak has been reached for Queens, Manhattan, and Staten Island. Here the model incorrectly forecasts dramatic continued growth in the death rates for these boroughs. As more training points are added, the posterior predictive forecasts show decreasing mean processes that are better aligned with the actual deaths and have less posterior uncertainty.

Across all the boroughs, the $\text{NB}(\rho, \alpha)$ model appears to perform reasonably well in capturing the dynamics of the observed daily counts once it has been fit with data that extend a few days past the observed peak. The fits improve as more data are included.

Figure 7 shows nonrandomized PIT histograms for $k$-day ahead predictions for the $\text{NB}(\rho, \alpha)$ model. The fits appear to break down around $k = 12$, with the predictive distributions becoming more and more overdispersed with an increasing $k$.

3.3 Fits to four Texas counties

In further exploration, we fit the $\text{NB}(\rho, \alpha)$ model to COVID-19 mortality data from the four most populous counties in Texas (Texas Department of State Health Services, 2020). These data come from the state, are subject to ongoing quality checks and updates, and include fatalities for which COVID-19 is listed as a direct cause of death on the death certificate. The four largest Texas counties are Harris County, which includes the city of Houston, Dallas and Tarrant counties, which include most of the Dallas–Fort Worth metroplex population, and Bexar County, which includes the city of San Antonio. We fit the $\text{NB}(\rho, \alpha)$ model to one long series for each county, beginning with its first recorded COVID-19 death, to the death count on December 20, 2020. The training data consist of the provisional counts available on January 12, 2021.
Posterior estimates of $\delta_t$ for 5 NYC boroughs – 90% interval and median

**FIGURE 6** Estimates of $\delta_t$ from the collection of NB($\rho$, $\alpha_1$) model fits. The red shaded regions show out-of-sample posterior predictive forecasting.

**FIGURE 7** PIT histograms for $k$-day ahead predictive distributions, for the NB($\rho$, $\alpha_1$) model.

Figure 8 shows the posterior estimates from these four fits. Over these longer training series, we observe multiple dynamic phases, including pronounced peaks in mid-to-late July and waves building at the end of the year. We can see how the dynamics of the estimated $\delta_t$'s on the right are reflected in the observed training counts and estimated $\lambda_t$'s on the left. The geographically adjacent Dallas and Tarrant counties show similar patterns of a more gradual, less devastating first wave, whereas Harris and Bexar Counties have dramatic and pronounced peaks. According to the model forecasts, the late fall/winter peaks may not yet be reached.
4 | CONCLUSION

We have presented a generalized state-space model for count-valued time series that seeks to capture the dynamics of local daily COVID-19 deaths. The model components, in particular the latent state process involving second-order differencing and an AR(1)-ARCH(1) model, are motivated by supposed behaviour of waves of epidemic counts. The results we see in applying our proposed model justify these components.

DATA AVAILABILITY STATEMENT

The data used in this paper are openly available at https://github.com/nychealth/coronavirus-data (New York City Department of Health and Mental Hygiene, 2020) and https://dshs.texas.gov/coronavirus/additionaldata (Texas Department of State Health Services, 2020).

ORCID

William R Palmer  
https://orcid.org/0000-0002-7160-0180

REFERENCES

Altieri, N., Barter, R. L., Duncan, J., Dwivedi, R., Kumbier, K., Li, X., Netzorg, R., Park, B., Singh, C., Tan, Y. S., Tang, T., Wang, Y., Zhang, C., & Yu, B. (2020). Curating a COVID-19 data repository and forecasting county-level death counts in the United States. Harvard Data Science Review. https://hdsr.mitpress.mit.edu/pub/p6isyf0g

Borkovec, M., & Klüppelberg, C. (2001). The tail of the stationary distribution of an autoregressive process with ARCH(1) errors. Annals of Applied Probability, 11(4), 1220–1241.

Carey, B., & Glanz, J. (2020). Hidden outbreaks spread through U.S. cities far earlier than Americans knew, estimates say. New York Times. https://www.nytimes.com/2020/04/23/us/coronavirus-early-outbreaks-cities.html

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, Articles, 76(1), 1–32.

Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2020). COVID-19 dashboard. Available from: https://coronavirus.jhu.edu/map.html [last accessed January 2021].

Chin, V., Samia, N. I., Marchant, R., Rosen, O., Ioannidis, J. P. A., Tanner, M. A., & Cripps, S. (2020). A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New York state. European Journal of Epidemiology, 35(8), 733–742.

Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth: A review. Physics of Life Reviews, 28, 66–97.

Cooper, I., Mondal, A., & Antonopoulos, C. G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons & Fractals, 139, 110057.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. Biometrics, 65(4), 1254–1261.
In Section 2 we describe the salient features of our modelling approach. Here we provide details of the full joint distribution of the $NB(\rho, \alpha_1)$ model.

For $t \geq 3$, we define the state $s_t$ as the three vectors containing $\log \lambda_t$ along with the first and second differences of $\log \lambda_t$. The first two states $s_t$ and $s_{t-1}$ need only contain initialized variables $\log \lambda_1$ and $\log \lambda_2$. The parameters $\beta$ are $(\phi, \rho, \alpha_0, \alpha_1)$, as in Table 1.

The negative binomial observation distribution gives

$$p(y_t|s_{t-1}; \phi, \rho, \alpha_0, \alpha_1) = \phi \left( \frac{\delta_t - \rho \delta_{t-1}}{\alpha_0 + \alpha_1 \delta_{t-1}} \right).$$

This satisfies conditional independence in the observation equation (Davis et al. 1999). For $t \geq 4$, it is clear by construction that $s_t$ is a function of $s_{t-1}$, $\beta$, and $\epsilon_t$, so that $s_t$ is conditionally independent of $s_{t-2}$ and $y_{t-1}$, given $s_{t-1}$ and $\beta$. For the state update density, we have

$$p(s_{t}|s_{t-1}; y_{t-1}; \beta) = p(\delta_t|\delta_{t-1}; \rho, \alpha_0, \alpha_1) = \phi \left( \frac{\delta_t - \rho \delta_{t-1}}{\alpha_0 + \alpha_1 \delta_{t-1}} \right),$$

where $\phi(\cdot)$ is the standard normal density. Thus, conditional independence in the state equation also holds. The complete joint distribution of the model is given by

$$p(y_{1:T}, s_{1:T}; \beta) = p(\beta) p(s_{1:2}, y_2) \prod_{t=4}^{T} p(\delta_t|\delta_{t-1}; \rho, \alpha_0, \alpha_1) \prod_{t=2}^{T} p(y_t|s_{t-1}; \phi).$$

What remains is to specify the prior distributions on the initial states and parameters $\beta$. For the initial states, we assume $p(s_{1:2}, y_2) = p(\log \lambda_1)p(\log \lambda_2|\lambda_1)p(\log \lambda_3|\lambda_1, \lambda_2)$, where

$$\log \lambda_1 \sim \text{Normal}(0, S^2)$$
$$\log \lambda_2|\lambda_1 \sim \text{Normal}(\log \lambda_1, 1^2)$$
$$\log \lambda_3|\lambda_1, \lambda_2 \sim \text{Normal}(\log \lambda_2 + \delta \log \lambda_2, .05^2).$$
The prior distributions for the parameters \( \beta \) are given by

\[
\begin{align*}
\phi & \sim \text{Half Normal}(0, 200^2) \\
\rho & = \frac{r_0}{2} \times \left( \sqrt{\tan^2 \theta + 4 - \tan \theta} \right) \\
a_1 & = \rho \times \tan \theta \\
\theta & \sim \text{Uniform} \left(0, \frac{\pi}{2}\right) \\
r_0 & \sim \text{Beta}(2, 1)
\end{align*}
\]

where \( \phi, \theta, r_0, \) and \( a_0 \) are all independent.

The priors for \( \phi \) and \( a_0 \) are restricted normal distributions, with only positive support. These are both strong, informative priors. The Half Normal distribution prevents the Negative Binomial scale parameter \( \phi \) from jumping to large values, which helps convergence in Stan. Note that as \( \phi \) increases, the conditional observation distribution approaches Poisson. Because we separately consider and fit the \( P(\rho, a_1) \) model, this restrictive prior does not limit our exploration.

For \( a_0 = \text{Var}(\delta_t|\delta_{t-1} = 0) \), setting the location and scale of the restricted normal to \( .001^2 \) places the prior mode away from 0 yet forces \( a_0 \) to be close to 0. This helps convergence in Stan. Moreover, the former effect helps prevent us getting stuck at \( \delta_t = 0 \), whereas the latter amounts to model regularization—restricting changes in the second-order difference \( \delta_t \) ensures smoothness of the mean process and prevents \( \lambda_t \) from following the observed counts too closely.

The joint prior \( p(\rho, a_1) \) has support on the region \( \mathcal{D} := \{(\rho, a_1) \in [0, 1]^2 : \rho^2 + a_1 < 1 \} \). It follows that draws of \( \rho \) and \( a_1 \) will satisfy sufficient stationarity conditions for autoregressive processes with ARCH(1) errors, as considered in Borkovec and Klüppelberg (2001). Although \( a_1 \) must be non-negative, we restrict \( \rho > 0 \) because a negative autoregressive term does not make sense in our model.

We construct \( p(\rho, a_1) \) so that it is close to uniform on \( \mathcal{D} \). A prior draw \( (\rho, a_1) \) is a transformation of a uniform draw on the unit quarter disc \( (\theta, r_0) \), given in polar coordinates. Note that \( p(\rho, a_1) \) and \( \mathcal{D} \) are given in the Cartesian coordinate system. To transform \( (\theta, r_0) \) to a point in \( \mathcal{D} \), we preserve the polar angle \( \theta \) while contracting the polar radius. In particular, we scale \( r_0 \) by \( r_{D}(\theta) \), where \( r_{D}(\theta) \) is the distance from the origin to the parabola \( x^2 + y = 1 \), along \( \theta \). Converting between coordinate systems, we have

\[
r_{D}(\theta) \sin \theta = 1 - r_{D}(\theta)^2 \cos \theta.
\]

This is a quadratic equation with one positive solution

\[
r_{D}(\theta) = \frac{-\sin(\theta) + \sqrt{1 + 3\cos^2 \theta}}{2\cos^2 \theta}.
\]

Finally, we convert the transformed draw \( (\theta, r_0 \times r_{D}(\theta)) \) to Cartesian coordinates to get

\[
\rho = \frac{r_0}{2} \times \left( \sqrt{\tan^2 \theta + 4 - \tan \theta} \right)
\]

and

\[
a_1 = \rho \times \tan \theta,
\]

as above.

**APPENDIX B: STAN MODEL CODE**

Here is the stan code for the NB(\( \rho, a_1 \)) model, with priors set up as described in Appendix A1.