Modeling post-holiday surge in COVID-19 cases in Pennsylvania counties

Benny Ren1*, Wei-Ting Hwang1

1 Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

* bennyren@pennmedicine.upenn.edu

Abstract

COVID-19 arrived in the United States in early 2020, with cases quickly being reported in many states including Pennsylvania. Many statistical models have been proposed to understand the trends of the COVID-19 pandemic and factors associated with increasing cases. While Poisson regression is a natural choice to model case counts, this approach fails to account for correlation due to spatial locations. Being a contagious disease and often spreading through community infections, the number of COVID-19 cases are inevitably spatially correlated as locations neighboring counties with a high COVID-19 case count are more likely to have a high case count. In this analysis, we combine generalized estimating equations for Poisson regression, a popular method for analyzing correlated data, with a semivariogram, to model daily COVID-19 case counts in 67 Pennsylvania counties between March 20, 2020 to January 23, 2021 in order to study infection dynamics during the beginning of the pandemic. We use a semivariogram that describes the spatial correlation as a function of the distance between two counties as the working correlation. We further incorporate a zero-inflated model in our spatial GEE to accommodate excess zeros in reported cases due to logistical challenges associated with disease monitoring. By modeling time-varying holiday covariates, we estimated the effect of holiday timing on case count. Our analysis showed that the incidence rate ratio was significantly greater than one, 6-8 days after a holiday suggesting a surge in COVID-19 cases approximately one week after a holiday.

Introduction

COVID-19, a highly contagious respiratory disease, first appeared in China at the end of 2019 and quickly spread across the world [1]. Evidence suggests mask-wearing, and social distancing are effective strategies in containing COVID-19 [2,3]. During the beginning of the pandemic, local and state governments quickly moved to implement mask mandates, travel restrictions and community containment measures (e.g., shelter in place) to mitigate the spread of the disease [4-6]. However, many Americans still choose to travel and congregate during the pandemic which is heightened during a federal holiday. Due to lack of adherence to public health guidance during the holidays, one should expect to see a surge in COVID-19 cases after a holiday. While many reports reaffirm this hypothesis, they are based on anecdotal evidence such as summary statistics of case counts from moving time windows. There are only a handful epidemiological studies that estimate the association between holiday timing and the number of reported COVID-19 cases [7,8]. For the first year of the pandemic, we
hypothesize that we should see a surge in COVID-19 cases within two weeks after a holiday given that the incubation period for COVID-19 extends up to 14 days, with a median time of 4-5 days from exposure to symptoms onset and adding up to an additional 3 days, either by the PCR test or the instant rapid antigen test, for a positive test to be reported [10][11]. We consider daily case counts between March 20, 2020 to January 23, 2021 to study early pandemic dynamics prior to widespread vaccine distribution, COVID-19 variants and at-home testing. In addition, rigorous disease surveillance and reporting procedures were in place during the beginning of the pandemic resulting in comprehensive infection data.

Poisson count regression with the population size as an offset is a popular approach to model count data and incidence rate. Based on the reporting guidelines of the COVID-19 datasets, there are certain dates such as holidays and weekends that could impact whether cases are being reported [12]. These reporting practices have resulted in excess or structural zeros in the case count, also known as zero-inflation [13]. We also need to consider spatial correlation among county-level COVID-19 case counts because vector-borne and transmissible diseases such as COVID-19, exhibit non-negligible spatial correlation as the movement of people can spread the virus to nearby counties [14]. Furthermore, processes that are confounded by spatially correlated variables are also suited for spatial models, a well-studied example is disease and pollution [15][17]. Thus, we expect to see similar case numbers or trends among neighboring counties [18].

Mixed models are powerful tools for spatial modeling due to its ability to handle a complex spatial correlation structure usually represented as a semivariogram or kriging process [10][20]. Mixed modeling problems have been addressed from a convex optimization and Bayesian computation perspective [14][21]. Correlation in spatial epidemiology can also be captured using conditional autoregressive models [22][24]. Inferential summaries from these models assume a correctly specified spatial correlation, otherwise a post-estimation robust covariance can be derived to address misspecified correlation. One such class of robust estimators are the heteroskedasticity-consistent or sandwich estimators which can easily be derived from marginal models and generalized estimating equations (GEEs). Marginal models are flexible alternatives to mixed models when population-level effects are of interest [25][26]. In addition, formulation of GEEs through the quasi-likelihood provide a convenient set of score equations with well established optimization procedures. As a result, a GEE formulation can be derived to incorporate spatial-temporal relationships, as well as, zero-inflation. For its generalizability and robust inference under misspecified spatial correlation, GEEs are an appealing but under utilized tool in spatial epidemiology.

We propose the use of a mixture of marginal models for zero-inflated over-dispersed Poisson regression to model the daily number of new cases in 67 Pennsylvania counties [27]. We combine the zero-inflated Poisson regression with the framework of a transition model to estimate present case counts as a function of previously reported cases [24][25]. We account for spatial correlation by treating semivariograms as the working correlation under a generalized estimation equation framework and designate each date as a ‘cluster’ under the traditional longitudinal nomenclature [26][28][30]. We propose an Expectation-Solution (ES) algorithm to fit the mixture of marginal models, which conveniently reduces to simpler problems of weighted GEEs and weighted semivariograms [31][35]. We include a combination of past case counts and time-lagged covariates, as well as county-specific and date-specific covariates to model daily rates of new COVID-19 cases following the parameterization of semivariogram working correlations discussed in [30]. While there are other methods to model spatial temporal data, we elect to use a GEE due to consistent inference under misspecified working correlation, simplicity of formulation and estimation under an imbalance clusters study design [14][36][37]. Imbalance clusters occurs in our data because counties report their
first case at different dates, staggered entry into the study, when populous counties report cases before rural counties.

The rest of this manuscript is organized as follows. In the Methods section, we outline the zero-inflated Poisson model and the proposed semivariogram model; we detail the zero-inflated GEE for the Poisson counts with excess zeros and describe the estimating procedures for the semivariogram through the ES algorithm and outline the robust sandwich standard error. In the Results section, we analyze the daily COVID-19 cases from 67 Pennsylvania counties.

Methods

Zero-inflated Poisson model

Let $Y_{i,t}$ denotes the count of COVID-19 cases at county $i$ and date $t$. Often times, a case count of zero is due to logistical issues in reporting, which is known as an excess zero and is unrelated to the Poisson model. A true zero is unrealistic in many situations. For excess zeros, we define the zero-inflation process with Poisson count data as

$$Y_{i,t} | Y_{i,t-1} \sim \begin{cases} 0 & \text{with probability } p_{i,t} \\ \text{Poisson}(\mu_{i,t} | Y_{i,t-1}) & \text{with probability } 1 - p_{i,t} \end{cases}$$

That is, if $Y_{i,t}$ is a zero Poisson random variable, it belongs to a Poisson distribution with probability $1 - p_{i,t}$, otherwise it is an excess zero with probability $p_{i,t}$ [13]. We denote the latent membership, excess zero indicator, for county $i$ at date $t$ as $Z_{i,t} | Z_{i,t-1} \sim \text{Bernoulli}(p_{i,t} | Z_{i,t-1})$, following a transition model as a function of previous outcomes that is unobservable and treated as missing data.

To model case counts as a rate, we define the Poisson regression model as

$$E[Y_{i,t} | Y_{i,t-1} = y_{i,t-1}] = \mu_{i,t} = \exp (x_{i,t}^\top \beta + \lambda^* \times \log (y_{i,t-1}/m_i) + \log (m_i))$$  \hspace{1cm} (1)

and (1) can be rewritten as

$$E\left[\frac{Y_{i,t}}{m_i} | Y_{i,t-1} = y_{i,t-1}\right] = \exp (x_{i,t}^\top \beta) \left(\frac{y_{i,t-1}}{m_i}\right)^{\lambda^*}$$

where $m_i$ is the county population offset term, representing the population at county $i$ but fixed for all dates $t$ and we denote $\mu_{i,t} | Y_{i,t-1}$ as $\mu_{i,t}$ and $p_{i,t} | Z_{i,t-1}$ as $p_{i,t}$ for brevity. We include a transition term of a lagged case incidence $Y_{i,t-1}$ to account for the temporal correlation of counts in the same county. Note that for $y_{i,t-1} = 0$ values must be corrected such that $\log(y_{i,t-1})$ are defined [24]. Our transition term addresses temporal trends and correlation allowing us to treat the residuals from different dates as independent in the working correlation of the marginal model. Here, $x_{i,t}$ are county and date-specific covariates detailed in our real data analysis.

We can also construct a logistic regression to model the latent membership $Z_{i,t}$ as

$$E[Z_{i,t} | Z_{i,t-1} = z_{i,t-1}] = p_{i,t} = \expit (u_{i,t}^\top \lambda + \lambda^* \times z_{i,t-1})$$  \hspace{1cm} (2)

We also incorporate a transition term, $Z_{i,t-1}$, to account for temporal trends and $u_{i,t}$ are county and date-specific covariates specifically related to the zero-inflation process.

Theoretical semivariogram model

Standard zero-inflated Poisson models do not account for the longitudinal nature of the data which is expressed as spatial correlation among counties. Though mixed GLMs can
be used to account for spatial correlation as Gaussian random effects, these frameworks involve difficult computation such as the Laplace approximation involving the covariance matrix. Spatial random effects are often modeled using semivariograms which can be viewed as a kriging model in the residual space [14]. Alternatively, we propose a GEE procedure which iteratively updates a working correlation using residuals allowing for a more straightforward estimation procedure.

Semivariograms model the correlation between two locations \(i\) and \(j\) as a decreasing function of distance. This phenomenon is commonly referred to as Tobler's first law of geography which states: “everything is related to everything else, but near things are more related than distant things.” In our study, distance is measured by meters and locations are latitude and longitude coordinates of county centroids. In our GEE procedure, residuals are expected to be correlated based on distance and are used to calculate the semivariogram working correlation. At a given date, county residuals \(r_{i,t}\) and \(r_{j,t}\), separated by distance \(d_{i,j}\), are assumed to follow a theoretical semivariogram

\[
\gamma(d_{i,j}, \tau) = \phi \left( 1 - \exp \left( -\frac{d_{i,j}}{\tau} \right) \right)
\]

where \(\phi\) is an over-dispersion parameter and \(\tau\) is a scale parameter and correlation does not depend on time. Thus, the working correlation matrix between counties \(i\) and \(j\) at date \(t\) is given as

\[
R_t(\tau) = \exp \left( -\frac{d_{i,j}}{\tau} \right)_{ij}
\]

and is solely a function of distance \(d_{i,j}\) with \(\tau\) determining the rate correlation decays as distance increases. We make an isotropic covariance assumption, meaning the semivariogram is a function of the locations only through the Euclidean distance between them.

While standard GEEs treat longitudinal data from individuals as independent clusters, our approach differs by treating dates as independent clusters, counties as our repeated measurements, spatial correlation as the working correlation as summarized in Table 1. After incorporating date-specific covariates and case counts from the previous day, we assume residuals from different dates to be uncorrelated in the working correlation. Fig 1 shows low autocorrelation of residuals, \(\text{Corr}(r_{i,t}, r_{i,t-k})\), within each county from a naive standard Poisson generalized linear model described by equation (1) using a lagged autoregressive predictor \((y_{i,t-1})\) and study covariates while assuming no zero-inflation. Our implementation of zero-inflated modeling assumes some observations belong to a structural zero model which further decreases autocorrelation in residuals, suggesting a single lag autoregressive predictor adequately addresses autocorrelation in the residuals. Furthermore, we expect the structural zeros to be related to holiday and weekend timings based on dataset guidelines and case counts to be cumulatively reported after a holiday. As a result, we anticipate holiday and weekend timings to be the major driver of temporal case trajectories in our study. Next, we detail the estimation approach using GEEs and estimation of scale parameter \(\tau\) based on procedures from the spatial statistics literature. In many practical settings there is an abundance of time-series data, but limited number of locations; our approach allows the asymptotic results to be driven by the number of dates in the analysis.
Table 1. Differences between standard GEE found in longitudinal studies and the proposed GEE

<table>
<thead>
<tr>
<th></th>
<th>Standard GEE</th>
<th>Proposed GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters</td>
<td>Individuals/Patients</td>
<td>Time points (Dates)</td>
</tr>
<tr>
<td>Repeated Measures</td>
<td>Repeated at different time points</td>
<td>Repeated at different locations</td>
</tr>
<tr>
<td>Working Correlation</td>
<td>Independent, Exchangeable, AR, etc.</td>
<td>Semivariogram model</td>
</tr>
</tbody>
</table>

Fig 1. Box plot of the residual autocorrelations, $Corr(r_{i,t}, r_{i,t-k})$, for lag $k$ ($k = 1, \ldots, 7$) from 67 Pennsylvania counties. The residuals $r_{i,t}$ were obtained from a Poisson generalized linear model with study covariates and a lagged autoregressive predictor ($y_{i,t-1}$) as outlined in the COVID-19 data analysis section. Light blue lines connect the autocorrelation values of each county. Solid black line connects the autocorrelation values $Corr(r_t, r_{t-k})$ calculated with residuals from all counties combined.

**Estimation**

**GEE for over-dispersed Poisson counts**

The GEE for the Poisson model is given as

$$Q(\beta) = \sum_{t=1}^{T} Q_t(\beta) = \sum_{t=1}^{T} D_t^T (\mu_t) V_t^{-1} (\tau_P, \phi) W_t (y_t - \mu_t)$$  \hspace{1cm} (3)$$

where

$$V_t (\tau_P, \phi) = \phi A_t^{1/2} R_t (\tau_P) A_t^{1/2}$$

$$D_t (\mu_t) = \frac{\partial}{\partial \lambda} \mu_t = \left[ x_{i,t}^T \mu_{i,t} \right]_i = A_t X_t$$

$$A_t = \text{diag} (\mu_{1,t}, \mu_{2,t}, \mu_{3,t}, \ldots)$$

$$\mu_{i,t} = \exp (x_{i,t}^T \beta + \beta^* \times \log (y_{i,t-1}^* / m_i) + \log (m_i))$$
with \( X_t = [x_{1,t}, x_{2,t}, x_{3,t}, \ldots]^T \) and \( \beta \) are the design matrix of covariates and vector of regression coefficients related to case counts. Here \( Y_{i,t-1}^* \) is the last count greater than zero for county \( i \) and \( y_t = [y_{1,t}, y_{2,t}, y_{3,t}, \ldots]^T \) is a vector representing new COVID-19 cases with corresponding expected values \( \mu_i = [\mu_{1,t}, \mu_{2,t}, \mu_{3,t}, \ldots]^T \). Scale parameter \( \tau_p \) is specific to the Poisson model. We incorporated Poisson model membership probabilities, \( w_{i,t} = 1 - p_{i,t} \), through matrix \( W_t = \text{diag}(w_{1,t}, w_{2,t}, w_{3,t}, \ldots) \), into the weighted GEE. When \( w_{i,t} \approx 0 \), then the data point is effectively removed from the estimation of Poisson regression parameters. The estimation of these weights is outlined in the Expectation-Solution algorithm in next section. We modify (1) to carry forward the, highly correlated, last non-zero count if \( Y_{i,t-1} = 0 \). Typically, when \( Y_{i,t-1} = 0 \) an additional parameter may be introduced in order to ensure log of \( y_{i,t-1}^*/m_i \) are defined \([24]\).

In accordance with GEEs, we use residuals to estimate the working correlation but our procedure involves using a parameterization described in the theoretical semivariogram. Over-dispersion is common in COVID-19 case counts due to high variability in reporting practices across counties but is conveniently accounted for with an additional parameter in the GEE. We calculate the dispersion parameter \( \phi \) and standard residuals \( r_{i,t} \) as

\[
\phi = \frac{\sum_{t=1}^{T} \sum_{i \in S(t)} w_{i,t} \left( \frac{y_{i,t} - \mu_{i,t}}{\sqrt{\phi_{i,t}}} \right)^2}{\sum_{t=1}^{T} \sum_{i \in S(t)} w_{i,t} - \text{rank}(X)}
\]

\[
r_{i,t} = \frac{y_{i,t} - \mu_{i,t}}{\sqrt{\phi_{i,t}}},
\]

where \text{rank}(X) is the number of linearly independent covariates and \( S(t) \) is the set of observed counties at time \( t \). Through the quasi-likelihood, GEEs incorporate a dispersion parameter, resulting in an over-dispersed Poisson regression.

**GEE for excess zeros**

To accommodate for latent membership \( Z_{i,t} \), we construct another GEE for estimating \( p_{i,t} \), the probability of an excess zero, using the logit link function. This GEE formulation is given as

\[
Q(\lambda) = \sum_{t=1}^{T} Q_t(\lambda) = \sum_{t=1}^{T} D_t^T (p_t) V_t^{-1}(\tau_Z) (z_t - p_t)
\]

where

\[
V_t(\tau_Z) = A_t^{1/2} R_t(\tau_Z) A_t^{1/2}
\]

\[
D_t(p_t) = \frac{\partial}{\partial \lambda} p_t = [u_{i,t}^T p_i (1 - p_{i,t})]_i = A_t U_t
\]

\[
A_t = \text{diag} (p_{1,t} (1 - p_{1,t}), p_{2,t} (1 - p_{2,t}), \ldots)
\]

\[
p_{i,t} = \expit (u_{i,t}^T \lambda + \lambda^* \times \mathbb{I}(y_{i,t-1} = 0))
\]

and \( \tau_Z \) is scale parameter for the excess zero model. Here \( U_t \) and \( \lambda \) are the covariates and regression coefficients for modeling the excess zero. Because we do not have the complete data, in order to account for serial correlation among excess zeros, we replace the transition term \( Z_{i,t-1} \) in [2] with an indicator of zero cases on the previous day, \( \mathbb{I}(y_{i,t-1} = 0) \) in order capture temporal trends. In practice, \( Z \) is often an imbalanced binary outcome where over parameterization of covariates can cause numerical
instability when fitting the model. Therefore, it has been suggested that it be fitted with a parsimonious set of covariates [27].

We calculate standard Pearson residuals as

\[
e_{i,t} = \frac{z_{i,t} - p_{i,t}}{\sqrt{p_{i,t} (1 - p_{i,t})}}
\]

which are used for estimating the working correlation using a semivariogram as described in the next section.

**Estimation of \( \tau \) through the empirical semivariogram**

After each Newton Raphson update of the regression coefficients \( \beta, \lambda \). We also update the working correlation by estimating the scale parameter \( \tau \), using the empirical semivariogram. In this section we outline procedures for modeling semivariograms and their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by

\[
\gamma_u(u) = \frac{1}{2} \sum_{i,j,w} w_{i,j} \left( \frac{z_{i,t} - z_{j,t}}{\sqrt{p_{i,t} (1 - p_{i,t})}} \right)^2,
\]

with a parsimonious set of covariates [27].

Recall, we use separate GEEs for the Poisson and excess zeros, to account for the instability when fitting the model. Therefore, it has been suggested that it be fitted with a parsimonious set of covariates [27].

We calculate standard Pearson residuals as

\[
e_{i,t} = \frac{z_{i,t} - p_{i,t}}{\sqrt{p_{i,t} (1 - p_{i,t})}}
\]

which are used for estimating the working correlation using a semivariogram as described in the next section.

**Estimation of \( \tau \) through the empirical semivariogram**

After each Newton Raphson update of the regression coefficients \( \beta, \lambda \). We also update the working correlation by estimating the scale parameter \( \tau \), using the empirical semivariogram. In this section we outline procedures for modeling semivariograms and their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.

Each bin of residuals will be used to calculate an empirical estimate for the midpoint of their computational implementation. First, we calculate the empirical semivariogram by grouping pairs of residuals into \( U \) bins, based on equally spaced intervals of distance.
presented median based approaches with 50% breakpoint \cite{39,41}. Being that estimating \( \tau \) is a sub-routine in the ES algorithm, we elect to use the approach from \cite{39} as our empirical semivariogram: \( \hat{\gamma}_{CH} (d_{u}) \) for robustness considerations. We denote the corresponding \( \tau \) estimates as \( G_{CH} (\mathbf{r}, \mathbf{W}) \), and \( G_{CH} (\mathbf{e}, \mathbf{I}) \) in our analysis. The empirical semivariogram is calculated in parallel and \( \tau \) is estimated numerically using weighted least squares and the \texttt{optim} package in \texttt{R} \cite{42}.

**Expectation-Solution algorithm**

Following \cite{27} we construct an Expectation-Solution algorithm for our zero-inflated regression models as a mixture of marginal models \cite{27,31}. The ES algorithm is a modification to the well-known Expectation-Maximization (EM) algorithm, where the M-step is replaced by the solution from the GEE \cite{43}. We further modify the working correlation to follow a spatial correlation structure which is estimated using a semivariogram \cite{30}.

**E-Step**

In the E-step, we update the unobserved \( \mathbf{Z} \). The conditional expectation is given as:

\[
\begin{align*}
\hat{z}_{i,t}^{(s+1)} &= \frac{\hat{p}_{i,t}^{(s)}}{\hat{p}_{i,t}^{(s)} + \left(1 - \hat{p}_{i,t}^{(s)}\right) f\left(y_{i,t} \mid \mu_{i,t}^{(s)}\right)} \hat{I}(y_{i,t} = 0) \\
\omega_{i,t}^{(s+1)} &= 1 - \hat{z}_{i,t}^{(s+1)}
\end{align*}
\]

where \( \omega_{i,t}^{(s+1)} \) are updated Poisson model membership weights. We denote the iteration number of the ES algorithm as \( s \).

**S-Step**

The S-step replaces the maximization step in the EM algorithm. In the S-step, we iteratively estimate Poisson model parameters \( \beta \), \( \tau_{p} \) and the excess zero model parameters \( \lambda \), \( \tau_{Z} \). The iterative updates for \( \beta \) using Newton Raphson are given as:

\[
\beta^{(k+1)} = \beta^{(k)} + \mathbf{H}^{-1} \left( \mu^{(k)}, \tau_{p}^{(k)}, \phi^{(k)} \right) \mathbf{Q} \left( \mu^{(k)}, \tau_{p}^{(k)}, \phi^{(k)}, \mathbf{W}^{(s+1)} \right)
\]

\[
\mathbf{H}^{-1} \left( \mu^{(k)}, \tau_{p}^{(k)}, \phi^{(k)} \right) = \left( \sum_{t=1}^{T} \mathbf{D}_{t} \left( \mu_{t}^{(k)} \right) \right)^{\top} \left[ \mathbf{V}_{t} \left( \tau_{p}^{(k)}, \phi^{(k)} \right) \right]^{-1} \left[ \mathbf{D}_{t} \left( \mu_{t}^{(k)} \right) \right]^{-1}
\]

\[
\mathbf{Q} \left( \mu^{(k)}, \tau_{p}^{(k)}, \phi^{(k)}, \mathbf{W}^{(s+1)} \right) = \sum_{t=1}^{T} \mathbf{D}_{t} \left( \mu_{t}^{(k)} \right)^{\top} \left[ \mathbf{V}_{t} \left( \tau_{p}^{(k)}, \phi^{(k)} \right) \right]^{-1} \mathbf{W}_{t}^{(s+1)} (y_{t} - \mu_{t}^{(k)})
\]

and \( \tau_{p}^{(k+1)} \) is updated with \cite{5} using \( \hat{\gamma}_{CH} (d_{u}) \). We denote the update as

\[
\tau_{p}^{(k+1)} = G_{CH} \left( \mathbf{r}^{(k+1)}, \mathbf{W}^{(s+1)} \right)
\]

where

\[
\phi^{(k+1)} = \sum_{t=1}^{T} \sum_{i \in S(t)} \omega_{i,t}^{(s+1)} \left( \frac{y_{i,t} - \mu_{i,t}^{(k+1)}}{\sqrt{\omega_{i,t}^{(s+1)}}} \right)^{2}
\]

\[
\left( \sum_{t=1}^{T} \sum_{i \in S(t)} \omega_{i,t}^{(s+1)} \right) - \text{rank}(\mathbf{X})
\]
The iteration number of the S-step are denoted with $k$. We repeat [8] and [9] until convergence to obtain $\beta^{(s+1)}$, a new iteration in the greater ES algorithm.

We apply the same procedure in the S-step for excess zero model parameters: $\lambda$ and $\tau_Z$

$$\lambda^{(k+1)} = \lambda^{(k)} + H^{-1} (p^{(k)}, r^{(k)}_Z) Q (p^{(k)}, r^{(k)}_Z, z^{(s+1)})$$

$$H^{-1} (p^{(k)}, r^{(k)}_Z) = \left( \sum_{t=1}^{T} D_t (p_t^{(k)})^T [V_t (r^{(k)}_Z)]^{-1} [D_t (p_t^{(k)})] \right)^{-1}$$

$$Q (p^{(k)}, r^{(k)}_Z, z^{(s+1)}) = \sum_{t=1}^{T} [D_t (p_t^{(k)})]^T [V_t (r^{(k)}_Z)]^{-1} (z_t^{(s+1)} - p_t^{(k)})$$

$$r^{(k+1)}_Z = G_{CH} (e^{(k+1)}, W = I)$$

where

$$e_{i,t}^{(k+1)} = \frac{z_{i,t}^{(s+1)} - p_{i,t}^{(k)}}{\sqrt{p_{i,t}^{(k)} (1 - p_{i,t}^{(k)})}}$$

and the weight matrices are replaced by the identity matrix. Equations [10] and [11] are repeated until convergence to obtain $\lambda^{(s+1)}$. For the ES algorithm, we iteratively update the E-step: $z^{(s)}$, $W^{(s)}$, and the S-step: $\beta^{(s)}$, $\lambda^{(s)}$ until convergence.

Inference on coefficients

Using the heteroskedasticity-consistent or robust sandwich estimator for standard errors,

$$\text{cov}(\beta) = \left( \sum_{t=1}^{T} D_t^T V_t^{-1} W_t D_t \right)^{-1} \left( \sum_{t=1}^{T} Q_t (\beta) Q_t^T (\beta) \right) \left( \sum_{t=1}^{T} D_t^T V_t^{-1} W_t D_t \right)^{-1}$$

we derive confidence intervals for $\beta$; we have consistent estimates even under misspecified working correlations which are spatial correlations in our case [31]. We proposed a reasonable spatial correlation model but retain reliable inference even in situations when the correlation model is incorrect. Analogously, we can also calculate the covariance for $\lambda$ with the equivalent formulation without the weight matrices, $W_t$, using the score equation [1]. We use the asymptotic distribution, $\hat{\beta} \sim N(\beta, \text{cov}(\beta))$ for inference, substituting the final estimates from the ES algorithm.

Results

Data

Poisson model covariates

Daily reported cases for 67 counties in Pennsylvania from March 20, 2020 to January 23, 2021 were obtained from the New York Times GitHub repository, https://github.com/nytimes/covid-19-data [12]. We restrict our study dates to be prior to widespread vaccine distribution, COVID-19 variants and at-home testing. The New York Times dataset includes confirmed cases from PCR tests as well as probable
cases. Probable cases are derived from a set of testing, symptoms and exposure criteria recommended by the Council of State and Territorial Epidemiologists and was also adopted by the Center of Disease Control on April 14, 2020 [44]. Data collection for a county starts after the first recorded case and we exclude the first 14 days of documented cases per county as reporting is often inconsistent during the start of data collection [45]. This also ensures carried forward imputation values: \( y_{i,t-1} > 0 \) and \( \log (y_{i,t-1}/m_i) \) are all defined.

Our primary goal is to understand the relationship between the timing of a holiday and daily new case counts. We consider federal holidays: New Years, Martin Luther King, George Washington Birthday, Good Friday, Memorial, Independence, Labor, Columbus, Veterans, Thanksgiving, Christmas and election day (November 3rd, 2020) [46]. Federal holidays are dates which a large proportion of the population are not working, enabling congregation and subsequent transmission of COVID-19. We create a vector of holiday binary \( \{1, 0\} \) indicators for whether a day in the last 2 weeks was a holiday in the Poisson regression model from today to the prior 14 days. For example, the count recorded on January 1st, the associated 15 holiday lag indicators are \((1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0) \) for dates: New Years, Dec 31, Dec 30, Dec 29, Dec 28, Dec 27, Dec 26, Christmas, . . . , Dec 18. We abbreviate the indicator for a holiday being \( k \) days in the past as, Holiday Lag \( k \): HL\(k\). Another example, for the count data on January 2nd, the indicators are \((0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0) \).

Therefore, the regression coefficients of our holiday indicators can inform whether there are increased cases (i.e., surge) with incidence rate ratios greater than 1, in relation to the presence and timing of a holiday within the prior 2 weeks. If the hypothesis of a post-holiday surge in COVID-19 cases is true, we expect to see several incidence rate ratios (IRRs) associated with HL\(k\) to be significantly greater than 1 after a holiday.

However, to account for the association with the day of the week, we include an indicator for the day of the week with Wednesday as the baseline. Our regression analysis also controls for other factors or covariates that may be associated with the incidence of COVID-19 cases as well as excess zeros in reporting. For covariates associated with case incidence, there’s growing evidence suggesting that warm and wet climate conditions seem to reduce spread of COVID-19 [47]. Precipitation and maximum temperature data from daily summaries was obtained from National Oceanic and Atmospheric Administration (NOAA) web service were included as additional covariates [48]. Maximum daily temperature and daily precipitation from weather stations across Pennsylvania were downloaded from the NOAA web service. Using latitude and longitude for each county centroid and weather station, county-specific weather data was interpolated using weather data from the same date. Interpolation was carried out using the \texttt{VIM R} package, median of \( K \) nearest neighbors (K-\textit{nn}), with \( K = 5 \) [49]. Precipitation was further binarized into a \( \{0, 1\} \) indicator based on whether or not it rained that day. Similar to the holiday covariates, we also construct 15 lag covariate indicators for precipitation in the prior 2 weeks. We also construct 15 continuous covariates of Fahrenheit temperature values for the prior 2 weeks. Following the convention of abbreviating Holiday Lag: HL\(k\), we abbreviate the covariates for temperature as TL\(k\) and precipitation as PL\(k\).

Other covariates included for the Poisson regression consisted of well-known demographic factors that affect health outcomes [50][51]. Demographic information from the U.S. census and 2016 American Community Survey were also included in the model as county-specific covariates. For each county, quintiles for population density, percentage of the population living in poverty, log median household income, log median house value, percentage of Black residents, percentage of Hispanic residents, percentage of the adult population with less than high school (HS) education, and percentage of owner-occupied housing were included as county-specific covariates. The data was
obtained using the same source outlined in [51].

Zero-inflation covariates

The presence of excess zeros can bias coefficient estimates in the negative direction if ignored. A non-negligible 21% of the reported case counts are zeros. Therefore, we propose a parsimonious model for \( Z \) by considering only a subset of covariates from the main Poisson regression model such that zero-inflation is accounted for based on well-known relationships associated with reporting zero cases. Based on the COVID-19 case reporting guidelines, covariates associated with excess zeros include holidays and weekends. For example, there tends to be a gap in reporting during weekends and holidays, suggesting case count reporting is more likely to occur on business days which does not reflect the reality of the pandemic and back logged cases tend to be lumped into the next business day’s case count.

Indicators for tomorrow, today and yesterday being a holiday are included model for \( Z \). As an example, for outcome data recorded on December 31st, the covariates will be \((1, 0, 0)\). An indicator for weekend is also included. In addition, certain county-specific covariates, percentage of population in poverty and percentage of owner-occupied housing are also included as covariates. Owner-occupied housing rate is an indicator of urban-rural dichotomy, with rural counties having a high rate of owner-occupied housing. Poverty rate and owner-occupied housing rate, as socioeconomic markers, are potential confounders for zero-inflation since the economy of a county may determine the resources available for municipal services such as case reporting. We also include \( I(Y_{i,t-1} = 0) \) an indicator for whether the previous day was a zero as a covariate to account for temporal trends.

COVID-19 data analysis

In addition to the Poisson model, we fitted a negative binomial hurdle model to capture the severity of a COVID-19 outbreak conditional that cases counts are greater than zero \([52,53]\). The hurdle model fits a negative binomial regression truncated to the positive case counts, \( \{Y_{i,t} : Y_{i,t} > 0\} \) using the same covariates as \([3]\) and models outcomes \( I(Y_{i,t} = 0) \) with a logistic regression with the same covariates as \([4]\). Robust sandwich estimators for the hurdle model were obtained using the \texttt{sandwich} R package \([54,55]\).

We fitted the following four models: 1.) Poisson generalized linear model without spatial correlation, without over-dispersion and without zero-inflation, we denoted as GLM, 2.) negative binomial hurdle model without spatial correlation, we denoted as Hurdle-NB, 3.) over-dispersed Poisson marginal model, with spatial correlation, but without zero-inflation, we denote this as GEE-OP, 4.) a zero-inflated over-dispersed Poisson marginal model as described by the Expectation-Solution algorithm, which we denote this as GEE-ZIOP.

The GEE-ZIOP results in Fig 2 and Table 2 shows the estimated incidence rate ratio (and 95% confidence interval) of 1.41 (1.07, 1.85); 1.37 (1.09, 1.72); 1.3 (1.08, 1.57); at 6, 7, 8 respective days after a holiday. At 6 to 8 days after a holiday we expect there to be a surge with 1.3 to 1.41 times as many cases as there would be without a holiday. Intervals for IRR estimates were plotted for all models in Fig 2 and showed that IRR estimates associated with the GEE-ZIOP model at the surge was greater than GLM, Hurdle-NB and GEE-OP models. All four models yield maximum IRR estimates at 6-8 days, suggesting that a surge in case counts occurs about 6-8 days after a holiday.

We did not find a similar time-varying effects of temperature or precipitation (results can be found in Table 5 & 6) and most of the IRRs associated with TL\( k \) and PL\( k \) covariates were not significant. From the results of the GEE models in Table 2, we
observed a significant increase in case incidence on Mondays, which aligns with the dataset guideline. Cases that were not reported during the weekend, possibly due to logistical issues, were reported on Monday instead. Additional county demographic effects can be found in Table 3. Notably, counties with a high proportion African-Americans were associated with increased COVID-19 incidence in all four models, with significant effects in the Hurdle-NB and GEE-ZIOP models. In addition, high median household income counties were associated with decreased COVID-19 incidence, in three of the four model, with significant effect in the GEE-OP and GEE-ZIOP models. Counties with low rate of high school education attainment were also associated with increased COVID-19 incidence in all four models, albeit not always at a significant level. Although many county demographic effects varied across the four models, a number of social and demographic effects from our analysis aligned with results from the literature [56, 57].

Table 2. Intercept and holiday coefficients with 95% confidence intervals based on robust standard errors.

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>Hurdle-NB</th>
<th>GEE-OP</th>
<th>GEE-ZIOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.9081 (-3.503, -0.313)</td>
<td>0.8293 (-4.861, 6.52)</td>
<td>-0.6568 (-2.576, 1.244)</td>
<td>-0.2922 (-2.286, 1.702)</td>
</tr>
<tr>
<td>HL0</td>
<td>-0.1852 (-0.382, 0.012)</td>
<td>0.0819 (-0.069, 0.233)</td>
<td>-0.5037 (-0.857, -0.15)</td>
<td>-0.1482 (-0.504, 0.208)</td>
</tr>
<tr>
<td>HL1</td>
<td>0.0460 (-0.069, 0.161)</td>
<td>0.065 (-0.116, 0.209)</td>
<td>0.1855 (-0.075, 0.446)</td>
<td>0.2716 (0, 0.543)</td>
</tr>
<tr>
<td>HL2</td>
<td>-0.0525 (-0.236, 0.13)</td>
<td>0.0746 (-0.118, 0.267)</td>
<td>-0.1261 (-0.408, 0.156)</td>
<td>-0.1084 (-0.446, 0.229)</td>
</tr>
<tr>
<td>HL3</td>
<td>-0.1139 (-0.255, 0.027)</td>
<td>0.1150 (-0.103, 0.333)</td>
<td>-0.3874 (-0.95, 0.175)</td>
<td>-0.2959 (-0.839, 0.247)</td>
</tr>
<tr>
<td>HL4</td>
<td>0.1706 (0.032, 0.309)</td>
<td>0.2342 (0.083, 0.385)</td>
<td>-0.0053 (-0.194, 0.183)</td>
<td>0.0464 (-0.168, 0.261)</td>
</tr>
<tr>
<td>HL5</td>
<td>0.0942 (-0.023, 0.212)</td>
<td>0.0904 (-0.036, 0.216)</td>
<td>0.1231 (-0.09, 0.336)</td>
<td>0.1867 (-0.046, 0.42)</td>
</tr>
<tr>
<td>HL6</td>
<td>0.2146 (0.029, 0.4)</td>
<td>0.2781 (0.101, 0.455)</td>
<td>0.1958 (-0.097, 0.489)</td>
<td>0.3424 (0.07, 0.615)</td>
</tr>
<tr>
<td>HL7</td>
<td>0.2167 (0.064, 0.369)</td>
<td>0.2262 (0.052, 0.4)</td>
<td>0.3260 (0.094, 0.558)</td>
<td>0.3135 (0.085, 0.542)</td>
</tr>
<tr>
<td>HL8</td>
<td>0.1704 (0.05, 0.291)</td>
<td>0.1118 (-0.045, 0.269)</td>
<td>0.2408 (0.075, 0.406)</td>
<td>0.2620 (0.074, 0.45)</td>
</tr>
<tr>
<td>HL9</td>
<td>0.1213 (-0.045, 0.287)</td>
<td>0.0941 (-0.121, 0.309)</td>
<td>0.1146 (-0.131, 0.36)</td>
<td>0.1511 (-0.169, 0.472)</td>
</tr>
<tr>
<td>HL10</td>
<td>0.0002 (-0.172, 0.172)</td>
<td>0.0080 (-0.171, 0.187)</td>
<td>0.1957 (-0.055, 0.446)</td>
<td>0.2389 (-0.017, 0.494)</td>
</tr>
<tr>
<td>HL11</td>
<td>0.1673 (0.045, 0.289)</td>
<td>0.1597 (0.026, 0.293)</td>
<td>0.1604 (-0.084, 0.405)</td>
<td>0.1780 (-0.096, 0.452)</td>
</tr>
<tr>
<td>HL12</td>
<td>0.0870 (-0.029, 0.203)</td>
<td>0.1322 (-0.003, 0.268)</td>
<td>0.1434 (-0.048, 0.334)</td>
<td>0.1774 (-0.016, 0.371)</td>
</tr>
<tr>
<td>HL13</td>
<td>-0.0634 (-0.228, 0.101)</td>
<td>-0.0282 (-0.172, 0.115)</td>
<td>0.0729 (-0.148, 0.294)</td>
<td>0.0834 (-0.161, 0.328)</td>
</tr>
<tr>
<td>HL14</td>
<td>0.0136 (-0.169, 0.196)</td>
<td>0.0841 (-0.073, 0.241)</td>
<td>0.0632 (-0.175, 0.301)</td>
<td>0.0540 (-0.18, 0.288)</td>
</tr>
</tbody>
</table>

GEE-OP: $\tau_P = 156552.2$, $\phi = 30.0776$  
GEE-ZIOP: $\tau_P = 114791.4$, $\phi = 30.84434$

GLM: Poisson generalized linear model, without over-dispersion, without zero-inflation, without spatial correlation  
Hurdle-NB: negative binomial Hurdle model, without spatial correlation  
GEE-OP: over-dispersed Poisson marginal model, without zero-inflation, with spatial correlation  
GEE-ZIOP: zero-inflated over-dispersed Poisson marginal model, with spatial correlation

Our results for modeling zero counts can be found in Table 4. The zero-inflated model differs from a hurdle model by treating zero counts probabilistically through a mixture model, while the hurdle model treats every zero count as a success in a Bernoulli process. For our data, we believe zeros may originate from the Poisson model or be superficially generated due to logistical issues in the reporting and a mixture model is more appropriate. However, a hurdle model is equivalent to a zero-inflated model when zero counts from the underlying Poisson model are rare and serve as a useful sensitivity analysis under different modeling assumptions. Counties with a high rate of owner occupied housing were significantly associated with an increased probability of reporting zero cases in both models. Owner occupied housing may reflect the urban-rural dichotomy of counties, with rural counties having a high rate of owner occupied housing and, by extension, conclude that rural counties have an increased probability of
Table 3. Day of week and county demographic coefficients and 95% confidence intervals based on robust standard errors.

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>Hurdle-NB</th>
<th>GEE-OP</th>
<th>GEE-ZIOP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta )</td>
<td>CI</td>
<td>( \beta )</td>
<td>CI</td>
</tr>
<tr>
<td>Thursday</td>
<td>0.1402 (0.033, 0.247)</td>
<td>0.1073 (-0.009, 0.224)</td>
<td>0.1787 (-0.031, 0.389)</td>
<td>0.1864 (-0.027, 0.4)</td>
</tr>
<tr>
<td>Friday</td>
<td>-0.0154 (-0.133, 0.102)</td>
<td>-0.0382 (-0.149, 0.073)</td>
<td>0.1256 (-0.06, 0.311)</td>
<td>0.1618 (-0.032, 0.356)</td>
</tr>
<tr>
<td>Saturday</td>
<td>-0.0789 (-0.176, 0.018)</td>
<td>-0.0135 (-0.116, 0.089)</td>
<td>-0.0976 (-0.274, 0.079)</td>
<td>0.1794 (-0.022, 0.381)</td>
</tr>
<tr>
<td>Sunday</td>
<td>-0.2833 (-0.407, -0.16)</td>
<td>-0.1567 (-0.283, -0.03)</td>
<td>-0.2987 (-0.498, -0.099)</td>
<td>0.0710 (-0.158, 0.3)</td>
</tr>
<tr>
<td>Monday</td>
<td>-0.0064 (-0.134, 0.121)</td>
<td>-0.1910 (-0.309, -0.072)</td>
<td>0.3323 (0.13, 0.535)</td>
<td>0.3746 (0.163, 0.586)</td>
</tr>
<tr>
<td>Tuesday</td>
<td>0.1072 (0.001, 0.213)</td>
<td>0.1213 (0.006, 0.237)</td>
<td>0.1406 (-0.031, 0.328)</td>
<td>0.1864 (-0.027, 0.4)</td>
</tr>
<tr>
<td>poverty rate</td>
<td>-0.2701 (-0.809, 0.269)</td>
<td>-0.7307 (-1.622, 0.161)</td>
<td>0.0286 (-0.645, 0.702)</td>
<td>0.0613 (-0.607, 0.729)</td>
</tr>
<tr>
<td>log median household value</td>
<td>-0.0414 (-0.151, 0.068)</td>
<td>0.2398 (-0.03, 0.509)</td>
<td>0.2781 (0.013, 0.543)</td>
<td>0.3458 (0.075, 0.616)</td>
</tr>
<tr>
<td>percent black</td>
<td>0.2393 (-0.086, 0.565)</td>
<td>1.3918 (0.539, 2.245)</td>
<td>0.2837 (-0.099, 0.666)</td>
<td>0.5388 (0.144, 0.934)</td>
</tr>
<tr>
<td>log median household income</td>
<td>0.0284 (-0.224, 0.281)</td>
<td>-0.5573 (-1.451, 0.337)</td>
<td>-0.6099 (-1.062, -0.158)</td>
<td>-0.7041 (-1.178, -0.231)</td>
</tr>
<tr>
<td>percent owner occupied housing</td>
<td>0.3923 (0.013, 0.772)</td>
<td>1.0397 (-0.412, 2.451)</td>
<td>0.7434 (0.382, 1.104)</td>
<td>0.6334 (0.28, 0.987)</td>
</tr>
<tr>
<td>percent hispanic</td>
<td>-0.5473 (-1.065, -0.029)</td>
<td>-1.0773 (-1.787, -0.368)</td>
<td>0.6337 (0.038, 1.23)</td>
<td>0.6459 (0.039, 1.253)</td>
</tr>
<tr>
<td>percent less than HS education</td>
<td>0.6412 (0.307, 0.976)</td>
<td>0.3586 (-0.505, 1.182)</td>
<td>0.0844 (-0.433, 0.602)</td>
<td>0.1517 (-0.349, 0.653)</td>
</tr>
<tr>
<td>2nd quintile pop. density</td>
<td>0.0648 (0.004, 0.125)</td>
<td>0.3405 (-0.309, -0.072)</td>
<td>0.0505 (0.004, 0.125)</td>
<td>0.1860 (-0.027, 0.4)</td>
</tr>
<tr>
<td>3rd quintile pop. density</td>
<td>0.1330 (0.079, 0.187)</td>
<td>0.0999 (-0.045, 0.085)</td>
<td>0.1226 (0.005, 0.23)</td>
<td>0.0422 (-0.056, 0.14)</td>
</tr>
<tr>
<td>4th quintile pop. density</td>
<td>0.1266 (0.075, 0.166)</td>
<td>0.0310 (-0.037, 0.099)</td>
<td>0.1916 (0.032, 0.351)</td>
<td>0.0882 (-0.056, 0.232)</td>
</tr>
<tr>
<td>5th quintile pop. density</td>
<td>0.1499 (0.099, 0.201)</td>
<td>0.0685 (-0.07, 0.207)</td>
<td>0.2359 (0.067, 0.405)</td>
<td>0.1170 (-0.035, 0.269)</td>
</tr>
<tr>
<td>( \log (y_{i,t}^*/m_i) )</td>
<td>0.6822 (0.646, 0.719)</td>
<td>0.6433 (0.61, 0.676)</td>
<td>0.5010 (0.461, 0.541)</td>
<td>0.5015 (0.464, 0.539)</td>
</tr>
</tbody>
</table>

Table 4. Coefficients and 95% confidence intervals based on robust standard errors for the excess zero model.

<table>
<thead>
<tr>
<th></th>
<th>Hurdle-NB</th>
<th>GEE-ZIOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda )</td>
<td>CI</td>
<td>CI</td>
</tr>
<tr>
<td>holiday tomorrow</td>
<td>-9.5816 (-10.5353, -8.628)</td>
<td>-5.4241 (-6.46, -4.3882)</td>
</tr>
<tr>
<td>holiday today</td>
<td>-0.5704 (-1.1793, 0.0386)</td>
<td>-0.3106 (-1.2357, 0.6145)</td>
</tr>
<tr>
<td>holiday yesterday</td>
<td>-0.2420 (-0.9197, 0.4357)</td>
<td>-0.5812 (-1.1974, 0.0349)</td>
</tr>
<tr>
<td>weekend</td>
<td>-0.3922 (-1.0473, 0.2629)</td>
<td>0.3125 (-1.1039, 1.7289)</td>
</tr>
<tr>
<td>poverty rate</td>
<td>0.3045 (0.0086, 0.6003)</td>
<td>0.4318 (-0.0858, 0.9495)</td>
</tr>
<tr>
<td>percent owner occupied housing</td>
<td>6.5020 (5.1133, 7.8907)</td>
<td>7.3636 (5.8934, 8.8338)</td>
</tr>
<tr>
<td>percent hispanic</td>
<td>8.8344 (7.6426,10.0262)</td>
<td>3.8253 (2.9493, 4.7013)</td>
</tr>
<tr>
<td>( \mathbb{I}(y_{i,t-1} = 0) )</td>
<td>2.4586 (2.2854, 2.6318)</td>
<td>1.5558 (1.4174, 1.6942)</td>
</tr>
</tbody>
</table>

GEE-ZIOP: \( \tau_Z = 197439.4 \)

Hurdle-NB: negative binomial Hurdle model, without spatial correlation

GEE-ZIOP: zero-inflated over-dispersed Poisson marginal model, with spatial correlation

Discussion

One of the goals of this analysis was to understand the relationship between holiday timing and surges in COVID-19 cases while accounting for zero-inflation, temporal effects, and spatial correlations in case counts from neighboring areas. We implemented an ES algorithm for the estimation of our zero-inflated Poisson model as a mixture of reporting zero cases. This confounding may be due to logistical challenges in reporting cases at a daily frequency and less community transmission due to sparse population density in rural areas. In addition, high poverty rate was significantly associated with an increased probability of reporting zero cases in both models. By testing the effects of socioeconomics factors on case reporting, we found that zero reported cases could be confounded by disease surveillance resources of local municipalities. Rural and poor counties may lack necessary resources to reliably report cases leading to excess zeros which can ultimately hinder the modeling COVID-19 incidence rates if not addressed.
Fig 2. Incidence rate ratio with robust 95% confidence interval for holiday effects. Estimated IRRs corresponding to the 15 holiday covariates, for each model. 95% confidence intervals were calculated using robust standard errors.

marginal models. We used a semivariogram for the spatial working correlation and update it in the S-step, after each Newton Raphson iteration. We take advantage of fast and robust semivariogram estimation as a sub-routine in the greater ES algorithm. Considering the zero-inflated nature of the data, carried forward imputation and time-lagged covariates were used to account for the temporal pattern of case reporting. After the estimations have converged, we used the robust covariance estimator, which is consistent under misspecified spatial correlation, to compute standard errors.

We analyzed data from March 20, 2020 to January 23, 2021, before widespread vaccine distribution, COVID-19 variants and at-home testing. Our analysis suggests a statistically significant surge in reported cases, an IRR of 1.3 to 1.41, 6-8 days after a holiday, with the surge of cases gradually tapering off afterward. This is in line with the anticipated timing of 4–5 days to symptom onset and up to 3 days for PCR testing results. Although the models estimated the surge in reported cases to be 6-8 days after a holiday, transmission may have occurred days prior to the holiday. It’s important to note that a calendar holiday date is not the exact date of COVID-19 transmission. For example, traveling may commence the Friday before a Monday holiday or many people may also extend their holidays so people may travel and congregate days after a holiday, e.g., Thanksgiving and Friday holidays. The reality is that there’s a window of days surrounding a holiday which the transmission likely occurred. In contrast, we observed negative effects 2–3 days after a holiday which could be explained by testing lag associated with holidays. If one contracts COVID-19 when traveling near the calendar holiday date, it takes on average 4–5 days for symptoms to appear. We believe that the negative effect timing could be explained by the asymptomatic incubation period right after contracting COVID-19 when people have yet to be tested.

All but the hurdle model, estimates in Fig 2 confirm a drop in the case counts on a holiday, with many of the backlogged cases being reported the next day, evident by a large positive effect one day after a holiday. In our GEE-ZIOP model, the IRR is
estimated to be 0.86 when the current is a holiday and 1.31 one day after a holiday. This trend is aligned with case reporting guidelines. Note that all four models (GLM, Hurdle-NB, GEE-OP and GEE-ZIOP) share similar conclusions regarding holiday effect, indicating the surge in reported cases occurs 6-8 after a holiday, suggesting that these results are robust under different assumptions.

Comparing different models, the GEE-ZIOP model was associated with an increase in the coefficient estimates, but at the expense of efficiency. When using GEE-ZIOP, weighting each data point by the zero-inflation membership probability decreases the effective sample size. However, even when considering this trade-off, there are often still meaningful results; our GEE-ZIOP estimate for the coefficients of HL6, HL7, and HL8 are significantly greater than zero. Furthermore, the GEE-ZIOP model detected an IRR significantly greater than 1, 6 days after a holiday, while GEE-OP did not. Several holidays (Memorial, Labor, Columbus) fall on a Monday, thus 6 days after a Monday is a Sunday and Sundays are likely dates where zero cases are reported. GEE-ZIOP allows us to down weight these excess zeros, revealing that surges in reported cases can occur as early as 6 days after a holiday.

In summary, we expect there to be a surge in county level reported cases one week after a holiday in the state of Pennsylvania. While our results reaffirm the common intuition regarding the early COVID-19 pandemic, our regression model elucidates the temporal trends of post-holiday case counts over a two-week period. In conclusion, we present results that illustrate the timing of a post-holiday surge while considering federal holidays and election day. Understanding the post-holiday surges can inform public policy and can be used to improve public health programs. We believe that our modeling approach and conclusions based on the rich data gathered through rigorous surveillance during the beginning of the pandemic, can guide future disease modeling analyses and are applicable to a wide range of epidemiological settings.

Supporting information

This work was supported by grants from the National Institute of Environmental Health Sciences: P30-ES013508.

Table 5. Temperature coefficients and 95% confidence intervals.

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>Hurdle-NB</th>
<th>GEE-OP</th>
<th>GEE-ZIOP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta )</td>
<td>CI</td>
<td>( \beta )</td>
<td>CI</td>
</tr>
<tr>
<td>TL0</td>
<td>-0.0018</td>
<td>(-0.006, 0.002)</td>
<td>-0.0061</td>
<td>(-0.01, -0.002)</td>
</tr>
<tr>
<td>TL1</td>
<td>-0.0030</td>
<td>(-0.008, 0.003)</td>
<td>-0.0010</td>
<td>(-0.006, 0.004)</td>
</tr>
<tr>
<td>TL2</td>
<td>-0.0016</td>
<td>(-0.007, 0.003)</td>
<td>-0.0021</td>
<td>(-0.007, 0.003)</td>
</tr>
<tr>
<td>TL3</td>
<td>-0.0085</td>
<td>(-0.014, -0.003)</td>
<td>-0.0076</td>
<td>(-0.013, -0.002)</td>
</tr>
<tr>
<td>TL4</td>
<td>0.0026</td>
<td>(-0.001, 0.007)</td>
<td>-0.0016</td>
<td>(-0.007, 0.004)</td>
</tr>
<tr>
<td>TL5</td>
<td>-0.0007</td>
<td>(-0.007, 0.004)</td>
<td>0.0030</td>
<td>(-0.003, 0.009)</td>
</tr>
<tr>
<td>TL6</td>
<td>-0.0018</td>
<td>(-0.006, 0.003)</td>
<td>-0.0017</td>
<td>(-0.007, 0.003)</td>
</tr>
<tr>
<td>TL7</td>
<td>-0.0021</td>
<td>(-0.006, 0.002)</td>
<td>-0.0029</td>
<td>(-0.007, 0.001)</td>
</tr>
<tr>
<td>TL8</td>
<td>0.0014</td>
<td>(-0.003, 0.006)</td>
<td>0.0005</td>
<td>(-0.005, 0.006)</td>
</tr>
<tr>
<td>TL9</td>
<td>-0.0038</td>
<td>(-0.009, 0.001)</td>
<td>-0.0023</td>
<td>(-0.008, 0.003)</td>
</tr>
<tr>
<td>TL10</td>
<td>-0.0020</td>
<td>(-0.006, 0.002)</td>
<td>-0.0038</td>
<td>(-0.009, 0.001)</td>
</tr>
<tr>
<td>TL11</td>
<td>0.0011</td>
<td>(-0.003, 0.005)</td>
<td>0.0037</td>
<td>(-0.001, 0.008)</td>
</tr>
<tr>
<td>TL12</td>
<td>0.0019</td>
<td>(-0.003, 0.007)</td>
<td>0.0019</td>
<td>(-0.003, 0.007)</td>
</tr>
<tr>
<td>TL13</td>
<td>-0.0015</td>
<td>(-0.006, 0.003)</td>
<td>-0.0060</td>
<td>(-0.01, -0.002)</td>
</tr>
<tr>
<td>TL14</td>
<td>0.0027</td>
<td>(-0.001, 0.007)</td>
<td>0.0054</td>
<td>(0.002, 0.009)</td>
</tr>
</tbody>
</table>
### Table 6. Precipitation coefficients and 95% confidence intervals.

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>Hurdle-NB</th>
<th>GEE-OP</th>
<th>GEE-ZIOP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>CI</td>
<td>β</td>
<td>CI</td>
</tr>
<tr>
<td>PL0</td>
<td>-0.0176</td>
<td>(-0.067, 0.032)</td>
<td>-0.0641</td>
<td>(-0.116, -0.012)</td>
</tr>
<tr>
<td>PL1</td>
<td>-0.0158</td>
<td>(-0.072, 0.04)</td>
<td>0.0182</td>
<td>(-0.034, 0.071)</td>
</tr>
<tr>
<td>PL2</td>
<td>-0.0142</td>
<td>(-0.07, 0.041)</td>
<td>-0.0185</td>
<td>(-0.072, 0.035)</td>
</tr>
<tr>
<td>PL3</td>
<td>0.0081</td>
<td>(-0.046, 0.062)</td>
<td>-0.0275</td>
<td>(-0.081, 0.026)</td>
</tr>
<tr>
<td>PL4</td>
<td>-0.0246</td>
<td>(-0.082, 0.033)</td>
<td>-0.0146</td>
<td>(-0.071, 0.041)</td>
</tr>
<tr>
<td>PL5</td>
<td>-0.0743</td>
<td>(-0.128, -0.02)</td>
<td>-0.0968</td>
<td>(-0.148, -0.045)</td>
</tr>
<tr>
<td>PL6</td>
<td>0.0212</td>
<td>(-0.032, 0.074)</td>
<td>-0.0196</td>
<td>(-0.079, 0.04)</td>
</tr>
<tr>
<td>PL7</td>
<td>-0.0885</td>
<td>(-0.143, -0.034)</td>
<td>-0.0714</td>
<td>(-0.129, -0.014)</td>
</tr>
<tr>
<td>PL8</td>
<td>-0.0359</td>
<td>(-0.093, 0.021)</td>
<td>-0.0231</td>
<td>(-0.076, 0.03)</td>
</tr>
<tr>
<td>PL9</td>
<td>0.0258</td>
<td>(-0.034, 0.086)</td>
<td>0.0113</td>
<td>(-0.045, 0.068)</td>
</tr>
<tr>
<td>PL10</td>
<td>-0.0635</td>
<td>(-0.126, -0.001)</td>
<td>-0.0724</td>
<td>(-0.128, -0.017)</td>
</tr>
<tr>
<td>PL11</td>
<td>0.0577</td>
<td>(-0.005, 0.12)</td>
<td>0.0177</td>
<td>(-0.035, 0.071)</td>
</tr>
<tr>
<td>PL12</td>
<td>-0.0633</td>
<td>(-0.117, -0.01)</td>
<td>-0.0336</td>
<td>(-0.084, 0.016)</td>
</tr>
<tr>
<td>PL13</td>
<td>0.0535</td>
<td>(-0.003, 0.11)</td>
<td>-0.0262</td>
<td>(-0.083, 0.03)</td>
</tr>
<tr>
<td>PL14</td>
<td>-0.0323</td>
<td>(-0.09, 0.026)</td>
<td>-0.0800</td>
<td>(-0.128, -0.032)</td>
</tr>
</tbody>
</table>

**References**


21. Lawson, A. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. (Chapman, 2018)


35. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. vol. 998. (John Wiley & Sons, 2012)


