A zero-inflated endemic–epidemic model with an application to measles time series in Germany

Junyi Lu  |  Sebastian Meyer

Institute of Medical Informatics, Biometry, and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Correspondence
Sebastian Meyer, Institute of Medical Informatics, Biometry, and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Waldstr. 6, 91054 Erlangen, Germany. Email: seb.meyer@fau.de

Funding information
Interdisciplinary Center for Clinical Research (IZKF), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Grant/Award Number: J75

Abstract
Count data with an excess of zeros are often encountered when modeling infectious disease occurrence. The degree of zero inflation can vary over time due to nonepidemic periods as well as by age group or region. A well-established approach to analyze multivariate incidence time series is the endemic–epidemic modeling framework, also known as the HHH approach. However, it assumes Poisson or negative binomial distributions and is thus not tailored to surveillance data with excess zeros. Here, we propose a multivariate zero-inflated endemic–epidemic model with random effects that extends HHH. Parameters of both the zero-inflation probability and the HHH part of this mixture model can be estimated jointly and efficiently via (penalized) maximum likelihood inference using analytical derivatives. We found proper convergence and good coverage of confidence intervals in simulation studies. An application to measles counts in the 16 German states, 2005–2018, showed that zero inflation is more pronounced in the Eastern states characterized by a higher vaccination coverage. Probabilistic forecasts of measles cases improved when accounting for zero inflation. We anticipate zero-inflated HHH models to be a useful extension also for other applications and provide an implementation in an R package.

KEYWORDS
epidemic modeling, measles, multivariate time series, seasonality, zero inflation

1  |  INTRODUCTION

Infectious disease models help to understand the mechanisms of disease spread and can be used to generate forecasts. The endemic–epidemic modeling approach of Held et al. (2005), also known as the HHH model (after the authors’ initials), is frequently adopted for time series of infectious disease counts. It is motivated by a branching process with immigration in that it decomposes disease incidence into endemic and autoregressive parts. The comprehensive R package surveillance (Meyer et al., 2017) provides tools for model estimation, simulation, and visualization. The HHH model has been applied to a variety of infectious diseases, including norovirus gastroenteritis (Meyer & Held, 2017), invasive pneumococcal disease (Chiavenna et al., 2019), pertussis (Munro et al., 2021), and COVID-19 (Dickson et al., 2020; Giuliani et al., 2020; Ssentongo et al., 2021).
Count data with excess zeros are often encountered in public health surveillance of rare diseases, such as syphilis in the United States of America (Yang et al., 2013), visceral leishmaniasis at block level in India (Nightingale et al., 2020), or dengue fever in China (Wang et al., 2014) or Brazil (Schmidt & Pereira, 2011). Zero-inflated (ZI) and hurdle models are typically used to analyze such data. Rose et al. (2006) conclude that the choice between a ZI and a hurdle model should generally be driven by the study purpose and the assumed underlying process. A ZI model, as a mixture model, assumes two underlying disease processes: zeros are generated both from an at-risk population (sampling zeros) and a not-at-risk population (structural zeros). This model can thus nest a conventional count data model, such as the HHH model, for the at-risk population. A hurdle model, on the other hand, typically characterizes an excess of sampling zeros without considering a not-at-risk population. For counts of infections, ZI models are more appropriate than hurdle models because a large not-at-risk population may exist, for example, those immunized by vaccination or recovery (permanently or temporarily), or a community which is detached from an outbreak area. Correspondingly, the not-at-risk proportion can vary over time; for example, it will decrease when cases are newly imported and enable local spread.

To account for the not-at-risk population, we propose an extension of the endemic–epidemic modeling approach for infectious disease time series with excess zeros: a multivariate zero-inflated HHH model. The HHH part of this mixture model uses the same endemic–epidemic decomposition as in the original HHH model. The additional zero-inflation part can incorporate both time and dynamic autoregressive effects. Region-specific random effects are allowed in both parts to account for heterogenous reporting or varying infection risks, for example, due to demographic factors not covered by covariates.

This paper is organized as follows. Section 2 describes the proposed modeling approach, which is evaluated with simulation studies in Section 3. In Section 4, we investigate different ZI models for measles time series from Germany and compare their forecast performance with classical HHH models. Section 5 concludes the paper.

### 2 | MODEL FORMULATION

#### 2.1 | Endemic–epidemic modeling

The so-called HHH model (Held et al., 2005; Paul & Held, 2011) assumes that the number of new cases $Y_{rt}$ of a (notifiable) infectious disease in unit $r, r = 1, \ldots, R$ at time $t, t = 1, \ldots, T$, given the past counts, follows a negative binomial distribution,

$$Y_{rt} | F_{t-1} \sim \text{NB}(\mu_{rt}, \psi_r),$$

with conditional mean $\mu_{rt}$ and conditional variance $\mu_{rt}(1 + \psi_r \mu_{rt})$. Here, $F_{t-1} = \sigma(Y_1, \ldots, Y_{t-1})$ represents the information up to time $t - 1$, and $\psi_r > 0$ are unit-specific overdispersion parameters, possibly $\psi_r \equiv \psi$. Units typically correspond to geographical regions, age groups, or the interaction of both (Meyer & Held, 2017), and time points often correspond to calendar weeks.

The endemic–epidemic modeling approach then decomposes the infection risk additively into an autoregressive component (for within-unit transmission), a spatiotemporal or neighborhood component (for transmission from other units), and an endemic component (for cases not directly linked to previously observed cases). The endemic component is sometimes called background risk, environmental reservoir, or immigration component. More specifically, the expected number of new cases $\mu_{rt}$ is modeled as

$$\mu_{rt} = \lambda_{rt} y_{rt-1} + \phi_{rt} \sum_{q \neq r} w_{qr} y_{qt-1} + \nu_{rt},$$

(1)

where $\lambda_{rt} > 0$ is the autoregressive parameter, $\phi_{rt} > 0$ is the spatiotemporal parameter, and $\nu_{rt} > 0$ is the endemic component. Transmission weights $w_{qr}$ quantify the “coupling” between different units and are possibly unknown (Meyer & Held, 2014). All three parameters are modeled on the log scale as

$$\log(\lambda_{rt}) = \alpha^{(\lambda)} + b_r^{(\lambda)} + x_{rt}^{(\lambda)T} \beta^{(\lambda)} + \log(\sigma_{rt}^{(\lambda)}),$$

$$\log(\phi_{rt}) = \alpha^{(\phi)} + b_r^{(\phi)} + x_{rt}^{(\phi)T} \beta^{(\phi)} + \log(\sigma_{rt}^{(\phi)}),$$

$$\log(\nu_{rt}) = \alpha^{(\nu)} + b_r^{(\nu)} + x_{rt}^{(\nu)T} \beta^{(\nu)} + \log(\sigma_{rt}^{(\nu)}),$$

with $\alpha^{(\lambda)}, \alpha^{(\phi)}, \alpha^{(\nu)}, b_r^{(\lambda)}, b_r^{(\phi)}, b_r^{(\nu)}, \sigma_{rt}^{(\lambda)}, \sigma_{rt}^{(\phi)}, \sigma_{rt}^{(\nu)}$ being the intercepts, regression coefficients, and scale parameters, respectively.
where in each component, $\alpha(\cdot)$ and $b(\cdot)$ are fixed and zero-mean random intercepts, respectively, $\beta(\cdot)$ is a vector of unknown coefficients for the covariates $x_{rt}$, and $o(\cdot)$ is an optional offset, for example, population fractions $o_{rt} = n_{rt}/n_t$ in the endemic component.

(Penalized) likelihood inference for the HHH models and many of its extensions is implemented in the R package surveillance (Meyer et al., 2017), which also contains several example data sets and vignettes for illustration.

### 2.2 Zero-inflated HHH model

To account for excess zeros in surveillance time series, we propose to extend the above HHH model to a ZI model, which we will call HHH4ZI. For this purpose, we assume the number of new cases $Y_{rt}$ given $F_{t-1}$ to follow a ZI negative binomial distribution (Yau et al., 2003). Its probability mass function is given by

$$f_{ZI}(y_{rt}; \mu_{rt}, \gamma_{rt}) = \gamma_{rt} \cdot 1_{\{0\}}(y_{rt}) + (1 - \gamma_{rt}) \cdot f(y_{rt}; \mu_{rt}, \phi_{rt}),$$

(2)

which represents a mixture of a point mass at zero and a HHH model with probability mass function $f(y_{rt}; \mu_{rt}, \psi_{rt})$, that is, an NB$(\mu_{rt}, \phi_{rt})$ distribution. The zero-inflation parameter $\gamma_{rt}$ describes the probability that a zero count comes from the point mass at zero (structural zero). If $\gamma_{rt} \equiv 0$, the mixture model would reduce to a HHH model. Otherwise, a proportion $\gamma_{rt} \in (0, 1)$ of the population is assumed to be not at risk of infection, while infections in the remaining population follow a HHH model.

Simple ZI models assume $\gamma_{rt} \equiv \gamma$ with a single parameter. However, we will typically model the logit-proportion with a linear predictor,

$$\text{logit}(\gamma_{rt}) = \alpha(\gamma) + b(\gamma) \cdot x_{rt}^T \beta(\gamma),$$

(3)

including an intercept $\alpha(\gamma)$, random unit-specific deviations $b(\gamma)$, and covariate effects, similar to the parameters $\lambda_{rt}$, $\phi_{rt}$, and $\nu_{rt}$ of the HHH mean $\mu_{rt}$. For example, to reflect varying population immunity to respiratory infections, seasonal variation of the not-at-risk proportion could be modeled via $x_{rt} = (\sin(\omega t), \cos(\omega t), \ldots, \sin(S \cdot \omega t), \cos(S \cdot \omega t))^T$, where $S$ denotes the number of harmonics and $\omega = 2\pi/52$ for weekly data, or $\omega = 2\pi/26$ for biweekly data. Note that, for any $\delta, \zeta \in \mathbb{R}$,

$$\delta \sin(\omega t) + \zeta \cos(\omega t) = A \sin(\omega t + \varphi),$$

(4)

where $A = \sqrt{\delta^2 + \zeta^2}$ is the amplitude and $\varphi = \arctan(\zeta/\delta)$ is the phase shift of a sinusoidal wave. Furthermore, we can relate the zero-inflation probability to past counts to model that locally emerging cases will put a larger population at risk. Suppose $x_{rt} = y_{rt-1}$, then Equation (3) can be rewritten in terms of the odds $\gamma_{rt}/(1 - \gamma_{rt}) = \exp(\alpha(\gamma) + b(\gamma) \cdot \exp(\beta(\gamma)) y_{rt-1})$ that the count inherits from the point mass at zero. One additional case in $y_{rt-1}$ would change the odds for excess zeros by a factor of $\exp(\beta(\gamma))$. Note that the odds for excess zeros are intrinsically driven by two opposite forces: First, with higher incidence in a region the probability of contact with an outbreak member increases, such that a larger part of the population is at risk ($\beta(\gamma) < 0$). Second, the population immunized by recovery will increase over time. Thus, a trend is a possible covariate in the zero-inflation part, reflecting a decreasing probability of epidemics due to a depletion of susceptibles.

The conditional mean and variance of the ZI model can be easily derived from a hierarchical formulation using a latent Bernoulli variable $W_{rt}$. Keeping notation simple by omitting the explicit conditioning on $F_{t-1}$ and random effects, we can write

$$W_{rt} \sim \text{Bernoulli} (\gamma_{rt}),$$

$$Y_{rt} | W_{rt} \sim \text{NB}((1 - W_{rt}) \mu_{rt}, \phi_{rt}).$$

By the laws of total expectation and variance,

$$E(Y_{rt}) = E_W(E(Y_{rt} | W_{rt})) = E_W((1 - W_{rt}) \mu_{rt}) = (1 - \gamma_{rt}) \mu_{rt}$$

(5)
and

$$\text{var}(Y_{rt}) = E_W(\text{var}(Y_{rt}|W_{rt})) + \text{var}_W(E(Y_{rt}|W_{rt}))$$

$$= (1 - \gamma_{rt})(1 + \mu_{rt} \psi_r + \gamma_{rt} \mu_{rt}) \mu_{rt}.$$  \hspace{1cm} (6)

The HHH4ZI model allows for the estimation of an effective reproduction number similar to the HHH model (Bauer & Wakefield, 2018). For this purpose, the expected non-environmental risk can be written in matrix form as $A_t y_{t-1}$, where $y_t = (y_{1t}, \ldots, y_{Rt})^T$ and $A_t$ is an $R \times R$ matrix with diagonal elements $(A_t)_{r,r} = (1 - \gamma_{rt}) \lambda_{rt}$ and $(A_t)_{r,r'} = (1 - \gamma_{rt}) \phi_{rt} w_{rr'}$ for $r \neq r'$. The model-based effective reproduction number $R_t$ is the dominant eigenvalue of $A_t$ (Diekmann et al., 2012, Part II).

We follow Paul & Held (2011) by considering two variants for the distribution of the random effects: with or without correlation between components. We denote the vector of all random effects from the four components by $b = (b(\lambda)^T, b(\phi)^T, b(\nu)^T, b(\gamma)^T)^T$, which is assumed to be multivariate normal with mean $0$ and covariance matrix $\Sigma$. The uncorrelated variant is given by

$$\Sigma = \text{blockdiag}(\sigma_\lambda^2 I_R, \sigma_\phi^2 I_R, \sigma_\nu^2 I_R, \sigma_\gamma^2 I_R),$$  \hspace{1cm} (7)

where $\sigma_\lambda^2, \sigma_\phi^2, \sigma_\nu^2, \sigma_\gamma^2$ are unknown variance parameters for each component and $I_R$ denotes the identity matrix of size $R$. A more flexible modeling assumption is to allow for within-unit correlation of the four random effects. Then the covariance matrix is defined as

$$\Sigma = \Theta \otimes I_R,$$  \hspace{1cm} (8)

where $\Theta \in \mathbb{R}^{4 \times 4}$ is an unknown covariance matrix and $\otimes$ denotes the Kronecker product. The positive definiteness of $\Sigma$ is ensured if $\Theta$ is positive definite (Horn & Johnson, 1991). The random effects are still uncorrelated between different units. In order to ensure computational efficiency and enforce the positive definiteness of $\Theta$, we use a spherical parameterization (Pinheiro & Bates, 1996; Rapisarda et al., 2007).

Parameter estimation for a HHH4ZI model follows the same (penalized) likelihood approach as used for the original HHH model (Paul & Held, 2011). The model extension is thus compatible with other HHH add-ons (e.g., Bracher & Held, 2022; Meyer & Held, 2017). Inference details including the analytical derivatives used for numerical likelihood optimization are given in the supplementary material.

### 3 | SIMULATION

We conducted simulation studies with $N = 1000$ repetitions using different time-series lengths $T \in \{50, 100, 500\}$ and two different models. In the first scenario, the data-generating process was a multivariate HHH4ZI model for the 16 German states with intercepts $\alpha(\lambda) = -0.3$, $\alpha(\phi) = 0.5$, $\alpha(\nu) = 0.5$, $\alpha(\gamma) = 0.2$, zero-inflation terms $x_{rt}^{(y)} = (\sin(2\pi t/26), \cos(2\pi t/26), y_{rt-1})^T$ with coefficients $b^{(y)} = (0.4, -0.3, -0.1)^T$, homogeneous overdispersion $\nu = 0.5$, and no random effects. We assumed normalized first-order transmission weights, that is, $w_{qr} = 1/m_q$, if state $q$ (with $m_q$ neighbors) is adjacent to state $r$, and $w_{qr} = 0$ otherwise. Furthermore, we used a state-specific offset $o(\phi)_r = n_r/n_r$ such that the rate of imported infections scales with the population size $n_r$.

Maximum likelihood estimation converged for all simulated data sets. The estimated parameter vectors are summarized by means and standard deviations in Table 1. As expected, increasing the length of the time series improves the estimates by reducing their variance. This is particularly relevant for the spatiotemporal parameter $\alpha(\phi)$, which could not be estimated reliably from short time series ($T = 50$). A possible reason is that the spatiotemporal component has a relatively small impact on the time series in the assumed model. For long time series ($T = 500$), all parameters are estimated at their true values on average. The coverage probabilities of Wald confidence intervals were approximately equal to their nominal levels (95% or 50%) for all parameters and time-series lengths.

In a second simulation study, we examined the fit of a HHH4ZI model when this is different from the true data-generating process. Specifically, we simulated from a HHH model with the same parameters as in the first scenario but
TABLE 1 True values, means, and standard deviations (SD) of parameter estimates from 1000 simulations for various time-series lengths \( T \). Coverage probabilities of 95% and 50% confidence intervals are also given.

| \( T \) | \( \hat{\alpha}(\gamma) \) | \( \hat{\alpha}(\phi) \) | \( \hat{\alpha}(\nu) \) | \( \hat{\alpha}(\psi) \) | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\psi} \) |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|       | True value     |                |                |                |                |                |                |                |
| 50    | -0.30          | 0.50           | 0.50           | 0.20           | 0.40           | -0.30          | -0.10          | 0.50           |
|       | Mean           | -0.32          | -0.13          | 0.50           | 0.20           | 0.40           | -0.31          | -0.10          | 0.49           |
|       | SD             | 0.12           | 2.79           | 0.10           | 0.13           | 0.13           | 0.14           | 0.03           | 0.10           |
|       | Coverage 95    | 95.00          | 95.60          | 94.70          | 95.50          | 95.10          | 94.70          | 94.80          | 93.40          |
|       | Coverage 50    | 49.90          | 55.40          | 51.80          | 52.10          | 50.80          | 49.10          | 49.10          | 49.90          |
| 100   | Mean           | -0.32          | 0.43           | 0.50           | 0.20           | 0.40           | -0.30          | -0.10          | 0.50           |
|       | SD             | 0.08           | 0.39           | 0.07           | 0.09           | 0.09           | 0.09           | 0.02           | 0.07           |
|       | Coverage 95    | 95.10          | 96.60          | 95.90          | 95.60          | 94.90          | 94.90          | 94.80          | 94.70          |
|       | Coverage 50    | 49.50          | 52.90          | 51.40          | 47.30          | 51.80          | 52.50          | 50.50          | 50.70          |
| 500   | Mean           | -0.30          | 0.49           | 0.50           | 0.20           | 0.40           | -0.30          | -0.10          | 0.50           |
|       | SD             | 0.04           | 0.14           | 0.03           | 0.04           | 0.04           | 0.04           | 0.01           | 0.03           |
|       | Coverage 95    | 95.90          | 96.40          | 94.40          | 94.30          | 95.60          | 94.90          | 95.30          | 94.40          |
|       | Coverage 50    | 50.50          | 50.80          | 51.80          | 48.10          | 49.80          | 49.30          | 49.20          | 48.00          |

We observed that the fit of the true HHH model on the 3 \( \times \) 1000 simulated time series has always converged, whereas the HHH4ZI fit did not converge for a few (6/3000) simulations. This relates to the poor model assumption about zero inflation in this scenario. Correspondingly, the zero-inflation probability intercept, \( \logit^{-1}(\alpha(\gamma)) \), for the average \( \alpha(\gamma) \), was only 0.05%. The power to select the true model naturally increased with the sample size: The HHH model had a lower (better) BIC than the corresponding HHH4ZI model in 98.7%, 99.7%, and 99.7% of the simulations for time-series lengths \( T = 50, 100, \) and 500, respectively.

4 | APPLICATION

4.1 | Data

We apply the proposed HHH4ZI model to count time series of reported measles cases in the 16 federal states of Germany, from 2005 to 2018. Herzog et al. (2011) have used an earlier version of these data to study the association between measles incidence and vaccination coverage. We follow their approach of aggregating the counts over biweekly intervals to approximately match the generation time of measles of around 10 days (Fine & Clarkson, 1982). Figure 1 shows that measles counts generally remained at low levels with many zeros during off-seasons. However, outbreaks with over 50 cases occurred in several states. The time series of some states, such as Berlin and Brandenburg, suggest a weak biennial cycle. A possible explanation is that susceptibles in low-immune (e.g., anthroposophic) communities (Ernst, 2011) are depleted during epidemic years. Only minor epidemics may occur throughout the following year until the susceptible population is replenished by births (Dalziel et al., 2016).

Vaccination coverage is surveyed by local health authorities by checking the receipt of the first and second doses of measles–mumps–rubella (MMR) vaccination among school starters. We assume the coverage rate among children who do not present a vaccination card at the day of the medical examination (6–14% on average for the different states) to be half of that of children presenting a vaccination card (Herzog et al., 2011). These data are currently available until 2018 (before measles vaccination became mandatory in Germany), which is why we restrict our measles analysis to the period from 2005 to 2018.

In Table 2, we compare summary statistics of case counts and vaccination coverage among the federal states. The maximum biweekly counts range from only 5 (Bremen) to 304 (North Rhine-Westphalia). Periods without any measles cases are rarest in Bavaria (66 biweeks), whereas Saarland observed zero cases in 341 (93.7%) of the 364 biweeks. Total case numbers naturally correlate with population size, but Berlin experienced relatively large epidemics. The yearly updated state
population will serve as an offset in the endemic model component. Relative population changes from the beginning to the end of the study period are in the range from −11% (Saxony-Anhalt) to +7% (Berlin). The estimated yearly vaccination coverage varies from 86% (Bremen, 2005) to above 95% (Thuringia, 2008–2012). Germany’s five eastern states (BB, MV, SN, ST, and TH, but not Berlin) already had a high initial vaccination coverage for historical reasons (measles vaccination was mandatory in the former German Democratic Republic). In the other states, vaccination coverage tended to increase during the investigated period (see Figure S1).

4.2 Models

We first consider the simple Poisson HHH model, which Herzog et al. (2011) found to provide the best fit. It included the log proportion of unvaccinated school starters as a covariate in the autoregressive component. This proportion can be regarded as a proxy for the fraction of susceptibles, assuming that one dose of the MMR vaccine already provides full

**FIGURE 1** Biweekly number of measles cases in the 16 federal states of Germany for the years 2005–2018. Note the varying y-axes.
TABLE 2  Maximum count, number of zeros, total of the 364 biweekly measles counts, (average) population, and estimated vaccination coverage in 2005 and 2018, for the 16 federal states of Germany.

| State                      | Max | Zeros | Total   | Population        | Coverage |
|----------------------------|-----|-------|---------|-------------------|----------|
| Baden-Württemberg (BW)    | 105 | 147   | 1733    | 10,773,318        | 89% ... 90% |
| Bavaria (BY)              | 175 | 66    | 3082    | 12,654,101        | 88% ... 93% |
| Berlin (BE)               | 159 | 151   | 2533    | 3,464,046         | 90% ... 93% |
| Brandenburg (BB)          | 18  | 276   | 309     | 2,498,962         | 94% ... 94% |
| Bremen (HB)               | 5   | 324   | 53      | 665,379           | 86% ... 90% |
| Hamburg (HH)              | 44  | 233   | 498     | 1,773,744         | 90% ... 94% |
| Hesse (HE)                | 41  | 181   | 859     | 6,105,864         | 91% ... 95% |
| Lower Saxon (NI)          | 22  | 181   | 504     | 7,909,337         | 91% ... 93% |
| Mecklenburg-Western Pomerania (MV) | 7 | 340   | 36      | 1,634,664         | 93% ... 93% |
| North Rhine-Westphalia (NW) | 304 | 100   | 3641    | 17,831,720        | 89% ... 94% |
| Rhineland-Palatinate (RP) | 13  | 203   | 334     | 4,033,268         | 90% ... 94% |
| Saarland (SL)             | 8   | 341   | 55      | 1,010,670         | 91% ... 93% |
| Saxony (SN)               | 57  | 278   | 495     | 4,127,619         | 95% ... 93% |
| Saxony-Anhalt (ST)        | 16  | 303   | 173     | 2,309,032         | 93% ... 93% |
| Schleswig-Holstein (SH)   | 15  | 213   | 355     | 2,841,994         | 89% ... 92% |
| Thuringia (TH)            | 56  | 321   | 302     | 2,212,890         | 94% ... 94% |

protection. The endemic component contained a sinusoidal wave \( (4) \) to capture yearly seasonality, and the standardized state population \( n_r \) as an offset. A neighborhood effect was not included due to the coarse spatial resolution. Indeed, suspected cases and their unprotected contacts are isolated immediately (WHO Regional Office for Europe, 2013), and past incidence in adjacent states is much less informative than within-state dynamics. This model, called “P0” in the following, is given by

\[
Y_{r,t} | F_{t-1} \sim Po(\lambda_{r,t} y_{r,t-1} + \nu_{r,t}),
\]

\[
\log(\lambda_{r,t}) = \alpha(\lambda) + \beta(\lambda) \log(1 - x_{r,t}),
\]

\[
\log(\nu_{r,t}) = \alpha(\nu) + \delta \sin(2\pi t / 26) + \zeta \cos(2\pi t / 26) + \log(n_{r,t}),
\]

where \( x_{r,t} \) is the estimated vaccination coverage for at least one dose. This means that the local reproduction rate \( \lambda_{r,t} \) is proportional to (a power of) the fraction of susceptibles, which makes the epidemic component conform to the mass action principle.

Fisher and Wakefield (2020) proposed an ecological Poisson model, which also incorporates MMR vaccine effectiveness. They used the conditional mean

\[
n_r(1 - x_r)(\lambda_{r,t} y_{r,t-1} / n_r + \nu_{r,t}),
\]

where \( x \) is the vaccine effect, \( \lambda_r \) is the risk of infection, and \( \nu_{r,t} \) is the endemic risk. To take the effect of the varying vaccination coverage into account, we use a similar mean component in the following model extensions. Moreover, we include both yearly and biennial seasonality in both endemic and epidemic components. To account for overdispersion, we also switch to negative binomial distributions with state-specific overdispersion parameters. This model, denoted by “NB1,” is given by

\[
Y_{r,t} | F_{t-1} \sim NB(\lambda_{r,t} y_{r,t-1} + \nu_{r,t}, \psi_{r,t}),
\]

\[
\log(\lambda_{r,t}) = \alpha(\lambda) + \sum_{s=1}^{2} \left\{ \delta_s(\lambda) \sin \left( \frac{2\pi t}{s \cdot 26} \right) + \zeta_s(\lambda) \cos \left( \frac{2\pi t}{s \cdot 26} \right) \right\} + \log(1 - x_{r,t}),
\]

\[
\log(\nu_{r,t}) = \alpha(\nu) + \sum_{s=1}^{2} \left\{ \delta_s(\nu) \sin \left( \frac{2\pi t}{s \cdot 26} \right) + \zeta_s(\nu) \cos \left( \frac{2\pi t}{s \cdot 26} \right) \right\} + \log((1 - x_{r,t}) n_{r,t}),
\]
where $\kappa = 0.92$ is the posterior median estimated by Fisher and Wakefield (2020), and $s = 1, 2$ corresponds to yearly and biennial seasonality, respectively. To account for unobserved heterogeneity between states, this model is further extended with uncorrelated (“NB2”) or correlated (“NB3”) random effects in both components.

Based on the structure of the models NB1 to NB3, we build several HHH4ZI models of increasing complexity. The simplest extension is “ZI1,” which mixes NB1 with an observation-driven zero-inflation probability $\gamma_{r,t}$ modeled via

$$\logit(\gamma_{r,t}) = \alpha(y) + \beta(y) y_{r, t-1}.$$  

Models “ZI2” and “ZI3” are ZI1 models with additional uncorrelated (“ZI2”) or correlated (“ZI3”) random effects in endemic, autoregressive, and zero-inflation components. By incorporating yearly and biennial seasonality also in the zero-inflation components of models ZI2 and ZI3, we obtain models “ZI4” and “ZI5,” respectively.

### 4.3 Results

Figure 2 summarizes the estimated parameters from the various models. The estimated seasonal amplitudes in the autoregressive part, $\hat{A}^{(2)}$, are barely affected by the model updates. The yearly pattern is stronger than the biennial cycle; the combined effect with a peak in calendar weeks 11–12 and a minimum at calendar weeks 39–40 is shown in Figure S2. It also shows that the estimated seasonality of the endemic component is very similar, but its amplitude shrinks when allowing for seasonality of the zero-inflation probability (ZI4 and ZI5). The large overdispersion $\hat{\psi}$ estimated for Thuringia is considerably reduced by accounting for state-specific zero inflation (compare ZI1 to ZI2 or ZI3).

We note that allowing for correlation between random effects increases their variance (Figure 2, especially in ZI models, compare ZI2 to ZI3 or ZI4 to ZI5). The random effects in the zero-inflation component are strongly correlated with those of the other two components (see models ZI3 and ZI5 in Table S1). For regions with excess zeros and large outbreaks, a higher zero-inflation intercept enables the ZI component to capture more zeros while the parameters of the HHH component are less affected by zeros. On the other hand, for regions with excess zeros and no outbreak, a zero-inflation component is not necessary, since an HHH part with low intercept can already capture most of the zeros. We will discuss the spatial variation of the random intercepts in the context of model ZI3 further below.

In most biweeks, the estimate of the time-varying effective reproduction number $R_t$ is slightly smaller under zero inflation (Figure 3, comparing ZI3 to NB3). The estimated $R_t$ of NB3 is smooth, whereas that of ZI3 is influenced by the observation-driven zero-inflation parameter. In both models, we observe that $R_t$ is estimated to exceed the threshold 1 in January and typically remains below one between June and December.

We assess the quality of one-step-ahead forecasts during the last 4 years. Several commonly used proper scoring rules are considered (Czado et al., 2009): logarithmic score (LS), Dawid–Sebastiani score (DSS), ranked probability score (RPS), and squared error score (SES). The latter is a classical measure of forecast performance and is proper when regarded as a score for probabilistic forecasts (Gneiting et al., 2007). However, it gives the same score to predictive distributions with the same expectations, regardless of their shapes (Bröcker & Smith, 2007). The maximum logarithmic score (maxLS) (Ray et al., 2017) is not a proper score, but is additionally used to evaluate the worst-case forecast performance. Forecast performance using these scores is compared in Table 3.

Models ZI3 and ZI5 consistently produce the best and second-best forecasts, respectively, in terms of the average LS, DSS, and RPS. Based on Monte Carlo permutation tests for differences in mean log scores, model ZI3 significantly outperforms all other models except ZI5. Adding the zero-inflation component consistently improves the aforementioned scores (NB1 to ZI1, NB2 to ZI2 and NB3 to ZI3), where the simple Poisson model ranks last. However, the SES ranks models substantially different. Model ZI1 ranks first followed by models NB3, NB1, and NB2. Model ZI5 has the worst mean SES. The root mean squared prediction errors of all models are around 4, which means that the forecasts differ from the observed counts by around four cases on average. Looking at the maxLS (worst individual forecast), models have again different ranks. Model NB1 has the best (i.e., lowest) maxLS, closely followed by model ZI1, then follow ZI4, ZI5, ZI3, and ZI2. The worst-case forecast of the Poisson model is much worse.

We now focus on model ZI3 and discuss the remaining model parameters. The estimated autoregressive parameter in the zero-inflation component is $\hat{\beta^y} = -0.82$ (95% CI: $[-1.07, -0.56]$), which means that each additional case at time $t − 1$ decreases the odds of an excess zero at time $t$ by $1 − \exp(-0.82) = 56\%$. Plots of the fitted time series in Figure S3 show that model ZI3 fits reasonably well, especially in Brandenburg, Hamburg, Saarland, Saxony-Anhalt, and Thuringia. Figure 4 shows maps of the exp-transformed region-specific random effects estimated from model ZI3. For the endemic
FIGURE 2 Estimated parameters of fitted models. Circles represent point estimates and bars are corresponding 95% Wald confidence intervals. Dots are region-specific estimates.
**TABLE 3** Performance of the 103 one-step-ahead forecasts in terms of proper scoring rules: mean log score (LS), maximum log score (maxLS), mean Dawid–Sebastiani score (DSS), mean ranked probability score (RPS), and mean squared error score (SES). Lower scores are better; ranks are shown in parantheses. The Monte Carlo $p$-values for differences in mean log scores are based on 9999 random permutations, comparing each model against the best model ZI3.

| Model | LS    | $p$-Value | maxLS | DSS   | RPS   | SES   |
|-------|-------|-----------|-------|-------|-------|-------|
| P0    | 1.71(9)| 0.00      | 8.51(9)| 3.24(9)| 1.20(8)| 17.71(7)|
| NB1   | 1.36(8)| 0.00      | 3.85(1)| 1.88(6)| 1.22(9)| 15.89(3)|
| NB2   | 1.35(5)| 0.01      | 4.07(8)| 1.94(8)| 1.19(6)| 16.23(4)|
| NB3   | 1.35(6)| 0.01      | 4.06(7)| 1.94(7)| 1.19(7)| 15.86(2)|
| ZI1   | 1.36(7)| 0.00      | 3.86(2)| 1.85(3)| 1.18(5)| 15.47(1)|
| ZI2   | 1.34(3)| 0.02      | 4.05(6)| 1.87(5)| 1.13(3)| 17.66(6)|
| ZI3   | 1.33(1)|          | 4.04(5)| 1.82(1)| 1.13(1)| 17.21(5)|
| ZI4   | 1.34(4)| 0.03      | 3.96(3)| 1.86(4)| 1.14(4)| 18.05(8)|
| ZI5   | 1.34(2)| 0.06      | 3.97(4)| 1.82(2)| 1.13(2)| 18.08(9)|

**FIGURE 3** Time-varying effective reproduction number $R_t$ estimated by model ZI3 compared to model NB3.

**FIGURE 4** Maps of exp-transformed region-specific intercepts estimated in model ZI3. These correspond to rate ratios (left, center) and odds ratios (right), respectively.
and autoregressive components these correspond to rate ratios taking the regional average as a reference, whereas for the zero-inflation component the values correspond to odds ratios. Five states have a relatively high zero inflation (odds ratio larger than 3): Brandenburg (BB), Saarland (SL), Saxony (SN), Saxony-Anhalt (ST), and Thuringia (TH). Except for Saarland (SL), these states are in East Germany. This can be explained by the relatively high vaccination coverage in these states throughout the study period (facilitating long periods without any cases) combined with the occurrence of a few large outbreaks (Figure 1). This pattern can be accommodated by increasing both the zero inflation and the autoregressive parameter. Correspondingly, these five states have the largest autoregressive random effects (rate ratios above 1.18). Autoregression is also relatively strong in Bavaria (BY) and North Rhine-Westphalia (NW). The relative contributions of the endemic and epidemic components by state are shown in Figures S3 and S4. The baseline ZI probability, \( \logit^{-1}(\alpha(\gamma) + \beta(\gamma) \gamma) \) (i.e., given no previous cases), for Thuringia (TH), for example, is estimated as 85.1%. However, this may not always be a reliable proxy for the not-at-risk proportion of the population. The corresponding value for Mecklenburg-Western Pomerania (MV), which is also in East Germany with a high vaccination coverage, is only 2.2%. As discussed above, the lack of outbreaks in that state impedes the estimation of a not-at-risk population; such a time series with only sporadic cases (see Figure 1) actually conforms with a simple endemic model with a low intercept.

5 | CONCLUSION

We have proposed a multivariate ZI endemic–epidemic model for infectious disease counts with excess zeros. This model consists of two parts: a zero-inflation part, which represents the not-at-risk population, and a conventional HHH part, which models the at-risk population via an endemic–epidemic decomposition. Both parts can incorporate an observation-driven term and seasonality or other covariates, as well as random effects to account for heterogeneity across units.

We applied this model to state-level biweekly measles counts in Germany. Both yearly and biennial seasonality were included in our models to capture a potential biennial cycle of measles epidemics. The effect of spatiotemporally varying vaccination coverage was accounted for by assuming the incidence to be proportional to the unvaccinated proportion of school starters. We found more pronounced zero inflation in East Germany, where also the vaccination coverage was higher. With a smaller at-risk population, epidemics were less likely and more contained. However, in states with only sporadic cases and no outbreaks, zero inflation was negligible. Such time series are compatible with a simple endemic-only HHH model and thus do not provide enough information to estimate the not-at-risk population. Of course, there is a trade-off between identifiability and a more detailed, disaggregated model. Depending on the research question, an additional stratification in epidemiologically relevant age groups (Meyer & Held, 2017) could be investigated, also to reduce within-unit heterogeneity.

Having said that, the interpretation of the model components becomes more challenging with an increasing number of layers. We believe that the main advantages of the model extension are the greater ability to fit heterogeneous regions in a joint model, improving short-term forecasts of disease incidence. Indeed, in comparison with negative binomial models with the same HHH part, zero-inflated HHH models can better capture time series where both low-incidence periods and large outbreaks occur. In our case study of measles incidence, we found the extended models to consistently improve forecast performance. Zero-inflated HHH models will be useful also in other settings and can be easily applied given our reference implementation in the accompanying R package hhh4ZI.

ACKNOWLEDGEMENTS

This work was financially supported by the Interdisciplinary Center for Clinical Research (IZKF) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany [project J75]. Junyi Lu performed the present work in partial fulfilment of the requirements for obtaining the degree “Dr. rer. biol. hum.” at the FAU.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study were derived from the following resources available in the public domain: case counts from SurvStat@RKI 2.0, Robert Koch Institute (https://survstat.rki.de, accessed March 26, 2021), vaccination
coverage from the Information System of the Federal Health Monitoring (https://www.gbe-bund.de/, accessed February 23, 2021), and population data from the Federal Statistical Office of Germany (Statistisches Bundesamt, https://www.destatis.de/, accessed April 10, 2021). The derived data sets are part of the dedicated R package hhh4ZI available at https://github.com/Junyi-L/hhh4ZI. Code to reproduce all results using that package is provided in the supplementary material; code for the application is also available as demo(“measles”, package = “hhh4ZI”). We intend to merge the extended HHH implementation into future versions of the surveillance package.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID
Sebastian Meyer https://orcid.org/0000-0002-1791-9449

REFERENCES
Bauer, C., & Wakefield, J. (2018). Stratified space-time infectious disease modelling, with an application to hand, foot and mouth disease in China. Journal of the Royal Statistical Society Series C, Applied Statistics, 67(5), 1379–1398.

Bracher, J., & Held, L. (2022). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. International Journal of Forecasting, 38(3), 1221–1233.

Bröcker, J., & Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting, 22(2), 382–388.

Chiavenna, C., Presanis, A. M., Charlett, A., de Lusignan, S., Ladhani, S., Pebody, R. G., & De Angelis, D. (2019). Estimating age-stratified influenza-associated invasive pneumococcal disease in England: A time-series model based on population surveillance data. PLOS Medicine, 16(6), 1–21.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. Biometrics, 65(4), 1254–1261.

Dalziel, B. D., Bjernstad, O. N., van Panhuis, W. G., Burke, D. S., Metcalf, C. J. E., & Grenfell, B. T. (2016). Persistent chaos of measles epidemics in the pre-vaccination United States caused by a small change in seasonal transmission patterns. PLOS Computational Biology, 12(2), e1004655.

Dickson, M. M., Espa, G., Giuliani, D., Santi, F., & Savadori, L. (2020). Assessing the effect of containment measures on the spatio-temporal dynamic of COVID-19 in Italy. Nonlinear Dynamics, 101, 1833–1846.

Diekmann, O., Heesterbeek, H., & Britton, T. (2012). Mathematical tools for understanding infectious disease dynamics. Princeton Series in Theoretical and Computational Biology. Princeton University Press.

Ernst, E. (2011). Anthroposophy: A risk factor for noncompliance with measles immunization. Pediatric Infectious Disease Journal, 30(3), 187–189.

Fine, P. E. M., & Clarkson, J. A. (1982). Measles in England and Wales—I: An analysis of factors underlying seasonal patterns. International Journal of Epidemiology, 11(1), 5–14.

Fisher, I. H., & Wakefield, J. (2020). Ecological inference for infectious disease data, with application to vaccination strategies. Statistica Medica, 39(3), 220–238.

Giuliani, D., Dickson, M. M., Espa, G., & Santi, F. (2020). Modelling and predicting the spatio-temporal spread of COVID-19 in Italy. BMC Infectious Diseases, 20(1), 1–10.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society Series B, Statistical Methodology, 69(2), 243–268.

Held, L., Höhle, M., & Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling, 5(3), 187–199.

Herzog, S. A., Paul, M., & Held, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. Epidemiology and Infection, 139(4), 505–515.

Horn, R. A., & Johnson, C. R. (1991). Matrix equations and the Kronecker product. In Topics in matrix analysis (pp. 239–297). Cambridge University Press.

Meyer, S., & Held, L. (2014). Power-law models for infectious disease spread. Annals of Applied Statistics, 8(3), 1612–1639.

Meyer, S., & Held, L. (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread. Biostatistics, 18(2), 338–351.

Meyer, S., Held, L., & Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software, 77(11), 1–55.

Munro, A. D., Smallman-Raynor, M., & Algar, A. C. (2021). Long-term changes in endemic threshold populations for pertussis in England and Wales: A spatiotemporal analysis of Lancashire and South Wales, 1940–69. Social Science & Medicine, 288, 113295.

Nightingale, E. S., Chapman, L. A., Srikantiah, S., Subramanian, S., Jambulingam, P., Bracher, J., Cameron, M. M., & Medley, G. F. (2020). A spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India. PLoS Neglected Tropical Diseases, 14(7), e0008422.
Paul, M., & Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistica Medica*, 30, 1118–1136.

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistical Computation*, 6, 289–296.

Rapisarda, F., Brigo, D., & Mercurio, F. (2007). Parameterizing correlations: A geometric interpretation. *IMA Journal of Management Mathematics*, 18(1), 55–73.

Ray, E. L., Sakrejda, K., Lauer, S. A., Johansson, M. A., & Reich, N. G. (2017). Infectious disease prediction with kernel conditional density estimation. *Statistica Medica*, 36(30), 4908–4929.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*, 16(4), 463–481.

Schmidt, A. M., & Pereira, J. B. M. (2011). Modelling time series of counts in epidemiology. *International Statistical Review*, 79(1), 48–69.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.

Ssentongo, P., Fronterre, C., Geronimo, A., Greybush, S. J., Mbabazi, P. K., Muvawala, J., Nahalamba, S. B., Omadi, P. O., Opar, B. T., Sinnar, S. A., Wang, Y., Whalen, A. J., Held, L., Jewell, C., Muwanguzi, A. J. B., Greatrex, H., Norton, M. M., Diggle, P. J., & Schiff, S. J. (2021). Pan-African evolution of within- and between-country COVID-19 dynamics. *Proceedings of the National Academy of Sciences of the USA*, 118(28), e2026664118.

Wang, C., Jiang, B., Fan, J., Wang, F., & Liu, Q. (2014). A study of the dengue epidemic and meteorological factors in Guangzhou, China, by using a zero-inflated Poisson regression model. *Asia-Pacific Academic Consortium for Public Health*, 26(1), 48–57.

WHO Regional Office for Europe. (2013). *Guidelines for measles and rubella outbreak investigation and response in the WHO European Region*. [https://www.euro.who.int/__data/assets/pdf_file/0003/217164/OutbreakGuidelines-updated.pdf](https://www.euro.who.int/__data/assets/pdf_file/0003/217164/OutbreakGuidelines-updated.pdf)

Yang, M., Zamba, G. K., & Cavanaugh, J. E. (2013). Markov regression models for count time series with excess zeros: A partial likelihood approach. *Statistical Methodology*, 14, 26–38.

Yau, K. K. W., Wang, K., & Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4), 437–452.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article**: Lu, J., & Meyer, S. (2023). A zero-inflated endemic–epidemic model with an application to measles time series in Germany. *Biometrical Journal*, 65, 2100408. [https://doi.org/10.1002/bimj.202100408](https://doi.org/10.1002/bimj.202100408)