

An Endemic-Epidemic Beta Model for Time Series of Infectious Disease Proportions

Junyi Lu and Sebastian Meyer*

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

Abstract

Time series of proportions of infected patients or positive specimens are frequently encountered in disease control and prevention. Since proportions are bounded and often asymmetrically distributed, conventional Gaussian time series models only apply to suitably transformed proportions. Here we borrow both from beta regression and from the well-established HHH model for infectious disease counts to propose an endemic-epidemic beta model for proportion time series. It accommodates the asymmetric shape and heteroskedasticity of proportion distributions and is consistent for complementary proportions. Coefficients can be interpreted in terms of odds ratios. A multivariate formulation with spatial power-law weights enables the joint estimation of model parameters from multiple regions. In our application to a flu activity index in the USA, we find that the endemic-epidemic beta model provides a better fit than a seasonal ARIMA model for the logit-transformed proportions. Furthermore, a multivariate approach can improve regional forecasts and reduce model complexity in comparison to univariate beta models stratified by region.

Keywords: multivariate time series, epidemic modelling, seasonality, beta regression, influenza-like illness

*E-mail: seb.meyer@fau.de

1 Introduction

The current coronavirus pandemic has triggered the development of complex mechanistic models to simulate how the disease might spread through populations worldwide under various scenarios of interventions. Such models try to capture the different disease stages (e.g., susceptible, infectious, symptomatic, hospitalized, recovered) and their interplay at the population level in great detail, but the choice of suitable parameter values is not straightforward, especially for emerging infectious diseases. Provided with a longer history of public health surveillance data, empirical models become increasingly important to estimate seasonal, sociodemographic or environmental effects on disease spread. A frequently adopted approach for (multivariate) time series of case counts is the so-called *endemic-epidemic* regression model “HHH” of Held *et al.* [14] (named after the authors’ initials). It has been applied to a wide range of infectious diseases, including norovirus gastroenteritis [23], invasive pneumococcal disease [7], and most recently, COVID-19 [10, 34]. Being a dynamic statistical model, it can even be used to generate probabilistic forecasts of case counts [15].

In some applications, however, the *proportion* of infected individuals is of more direct interest than their absolute number. An important example is the national weighted influenza-like illness (wILI) index used by the Centers for Disease Control and Prevention (CDC) in the USA to monitor flu activity. The wILI index is the proportion of outpatient visits for influenza-like illness (ILI) weighted by state population, where ILI is defined as “fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a known cause other than influenza” [6]. Neither HHH with its Poisson or negative binomial distribution nor standard time series models with Gaussian errors are designed to model this proportion over time. Autoregressive integrated moving average (ARIMA) models only apply to suitably transformed proportions, e.g., to the log-scale [28], trying to improve the Gaussian approximation.

In this work, we thus propose an adaptation of the HHH approach to infectious disease proportions: a beta model with endemic and epidemic components capturing baseline seasonality and autoregressive effects, respectively. We formulate a distributional regression model for the mean and precision parameters of the conditional beta distribution along the lines of beta regression models [11, 33]. Coefficients of the mean model can be interpreted in terms of odds ratios. The autoregressive structure of the mean model is inspired by an ARMA model for proportions [2]. A multivariate version with spatio-temporal effects similar to the HHH model [26] is proposed to analyse proportion time series from different regions.

This paper is organized as follows. Section 2 describes the beta model and its features. Its multivariate version is formulated in Section 3. In Section 4, we investigate different beta models for both national and regional wILI time series in the USA. Section 5 concludes the paper with a discussion.

2 Endemic-epidemic beta model

Held *et al.* [14] proposed a modelling framework for time series of infectious disease surveillance counts, which has subsequently been termed the HHH framework. It can be motivated from the discrete-time SIR model [35] and is based on a branching process approximation for the number of newly infected individuals Y_t in time periods $t = 1, \dots, T$ in a large population, where the infected proportion stays small (rare disease assumption). An advantage of this approximation is that the model does not require information on the number of disease susceptibles.

A proportion time series $X_t = Y_t/N_t$ measures the proportion of new cases over time, where the population N_t could refer to all disease susceptibles, or all laboratory tests or doctor visits from a sentinel surveillance system. In principle, the nominator counts Y_t could still be modelled using the HHH framework, treating N_t as a fixed offset for the mean of Y_t . This approach is particularly attractive if there is negligible variation in N_t . Otherwise, forecasting the proportion would require an extra model for N_t . So for volatile N_t , such as laboratory tests or doctor visits, a natural alternative is to model the proportions directly.

Proportions are usually distributed asymmetrically and heteroskedastic, with less variation at the boundaries of the unit interval [8]. These characteristics can be accommodated by assuming a beta distribution for the response variable, i.e., a beta regression model [11, 33]. In this context, the beta density is parametrized by the mean $\mu \in (0, 1)$ and a precision parameter $\phi > 0$:

$$f(x; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1} (1-x)^{(1-\mu)\phi-1}, \quad 0 < x < 1. \quad (1)$$

For $X \sim \text{Beta}(\mu, \phi)$ in this parametrization, $E(X) = \mu$ and $\text{var}(X) = \mu(1-\mu)/(1+\phi)$, so a beta model is naturally heteroskedastic.

Inspired by HHH and related endemic-epidemic approaches [24], we propose a dynamic model for time series of proportions assuming a beta distribution as the conditional distribution. We consider the proportion $X_t \in (0, 1)$ of infections with a particular disease in time period $t = 1, \dots, T$ as a measure of disease activity. The corresponding denominator could be all doctor visits or all sampled specimens from a population at time t . Conditional on past observations, $\mathcal{F}_{t-1} = \sigma(X_1, \dots, X_{t-1})$, we assume this proportion X_t to follow a beta distribution:

$$X_t | \mathcal{F}_{t-1} \sim \text{Beta}(\mu_t, \phi_t). \quad (2)$$

Following the HHH approach, the mean proportion μ_t will be modelled by an endemic predictor and an autoregressive term.

When incorporating autoregressive terms into the mean model, the main difficulty is that the unit interval $I = (0, 1)$ with standard addition and multiplication is not even a vector space. However, I becomes a Hilbert space by introducing two basic operations

and distance on I following simplicial geometry [1, 27], and the logit transformation is an isometry from I to \mathbb{R} [27]. We thus use the logit link for the mean and the autoregressive term, i.e.,

$$g(\mu_t) = \nu_t + \beta g(X_{t-1}), \quad (3)$$

where $\beta > 0$ quantifies the epidemic effect of the last observation, and $g(x) = \log(x/(1-x))$ is the logit link function.

The endemic component is modelled as

$$\nu_t = \alpha^{(v)} + \beta^{(v)\top} z_t^{(v)}, \quad (4)$$

where $\beta^{(v)}$ is a vector of regression parameters associated with covariates $z_t^{(v)}$. In the absence of covariates, $g^{-1}(\alpha^{(v)})$ corresponds to the expected proportion given a prior value of 50%.

Equation (3) actually describes a model for the log-odds of infection. Exp-transformation to odds gives $o_t := \mu_t/(1-\mu_t) = e^{\nu_t} \hat{o}_{t-1}^\beta$, where \hat{o}_{t-1} denotes observed odds at time $t-1$. Thus, exponentiated $\beta^{(v)}$ parameters can be interpreted in terms of odds ratios. Suppose that the value of the i -th covariate increases by one unit and all other covariates remain unchanged. This change multiplies the odds by $e^{\beta_i^{(v)}}$. Similarly, when the endemic component remains unchanged, but the observed prior odds \hat{o}_{t-1} change by a factor of c , the odds o_t get multiplied by c^β .

The precision parameter $\phi_t > 0$ is also allowed to be time-varying with

$$\log(\phi_t) = \alpha^{(\phi)} + \beta^{(\phi)\top} z_t^{(\phi)}. \quad (5)$$

Seasonality in both the endemic component (4) and the log-precision (5) can be modelled as a Fourier series via $z_t = (\sin(\omega t), \cos(\omega t), \dots, \sin(S \cdot \omega t), \cos(S \cdot \omega t))^\top$, where S denotes the number of harmonics and $\omega = 2\pi/52$ for weekly data [16]. Holiday effects for specific calendar weeks can be captured via dummy variables.

The above autoregressive model for the mean proportion (3) can be extended to include higher-order lags [4]:

$$g(\mu_t) = \nu_t + \sum_{k=1}^p \beta_k g(X_{t-k}). \quad (6)$$

This extended model is denoted by Beta(p), where p is the maximum lag for which an autoregressive effect is considered. When $p > 1$, the epidemic effect of past observations is more complicated, and $\beta_k, k = 1, \dots, p$ can be either greater or less than 0.

When modelling proportion time series using the beta model, the distribution of complementary proportions is also taken into account. The time series of the complementary proportion $\tilde{X}_t := 1 - X_t$ describes the proportion of individuals who are not infected by the disease under consideration, turning to the doctor for other reasons (depending on the definition of the denominator). Suppose the time series of proportions X_t follows a Beta(p)

model with conditional mean μ_t and precision ϕ_t . Since $f(x; \mu, \phi) = f(1 - x; 1 - \mu, \phi)$ for the beta density in (1), the time series of the complementary proportions \tilde{X}_t also follows a conditional beta distribution, now with mean $\tilde{\mu}_t := 1 - \mu_t$ and the same precision ϕ_t , where, by $g(x) = -g(1 - x)$,

$$\begin{aligned} g(\tilde{\mu}_t) &= -g(\mu_t) \\ &= -v_t - \sum_{k=1}^p \beta_k g(X_{t-k}) \\ &= -v_t + \sum_{k=1}^p \beta_k g(\tilde{X}_{t-k}). \end{aligned} \tag{7}$$

Thus, \tilde{X}_t follows a Beta(p) model with an opposite endemic component $-v_t$ and an autoregressive component $\beta_k g(\tilde{X}_{t-k})$ having the same parameters as the X_t model. Modelling either of X_t or \tilde{X}_t and then deriving the corresponding parameters for X_t will give equivalent results, so Beta(p) is a symmetric model.

3 Multivariate formulation

Disease surveillance systems typically provide multivariate infectious disease data, for example disaggregated by region. The endemic-epidemic beta model can be extended along the lines of the multivariate HHH model [14, 26]. We denote the proportion of infections in region r , $r = 1, \dots, R$, by $X_{r,t}$ and assume conditionally independent beta distributions

$$X_{r,t} | \mathcal{F}_{t-1} \sim \text{Beta}(\mu_{r,t}, \phi_{r,t}). \tag{8}$$

In the most general formulation, the mean model can be formulated as

$$g(\mu_{r,t}) = v_{r,t} + \sum_{k=1}^p \beta_{r,k} g(X_{r,t-k}) + \gamma_r \sum_{r' \neq r} [\omega_{r',r}] g(X_{r',t-1}), \tag{9}$$

with $v_{r,t}$ modelled as in Equations (4) and (5), and normalized transmission weight $[\omega_{r',r}] = \omega_{r',r} / \sum_k \omega_{r',k}$. We denote this model by mBeta(p).

The parameter γ_r quantifies the effect of past proportions in other regions on the proportion in region r . The transmission weight $\omega_{r',r}$ models the disease transmission from region r' to r . Note that we use the normalized transmission weights, so that $\sum_r [\omega_{r',r}] = 1$. A possible parametric model for transmission weights is the power-law model [22], with $\omega_{r',r} = o_{r',r}^{-\rho}$ where $o_{r',r}$ is the adjacency order and $\rho > 0$ is the decay parameter. This captures both short-range dependence of direct neighbours ($o_{r',r} = 1$) as well as occasional long-range transmissions (large $o_{r',r}$). The larger the decay ρ , the less relevant is the latter. The limit $\rho \rightarrow \infty$ is equivalent to binary adjacency weights $\omega_{r',r} = \mathbf{1}(r' \sim r)$, a simple and frequently used alternative to model only first-order dependencies [26]. In contrast, $\rho = 0$

would assign equal weight to all regions.

Besides accounting for spatial dependence, the multivariate approach allows for more efficient parameter estimates than R independent univariate models by region. If the effects of covariates or seasonality can be assumed homogeneous across (subsets of) regions, a single $m\text{Beta}(p)$ fit is preferable over R $\text{Beta}(p)$ fits.

All parameters of the proposed distributional regression model can be estimated via (conditional) maximum likelihood inference, where we build on functions from the R package `betareg` [8]. This uses the quasi-Newton optimizer implemented in the R function `optim()` to maximize the log-likelihood and estimates the variance-covariance matrix of the parameters as the inverse of the expected Fisher information. We additionally use profile likelihood inference to estimate parametric transmission weights in multivariate models. Specifically, we maximize the maximum log-likelihood of the $m\text{Beta}$ model with fixed weight matrix as a function of the log-transformed decay parameter (to ensure $\rho > 0$). A similar approach was previously used to estimate contact weights [23] and higher-order autoregressive effects [4] in HHH models.

4 Applications

4.1 National influenza-like illness in the USA

The Centers for Disease Control and Prevention (CDC) in the USA monitor flu activity through their "outpatient influenza-like illness surveillance network" (ILINet). Flu activity levels are based on the national weighted influenza-like illness (wILI) index, which is the proportion of patients with influenza-like illness (ILI) among all outpatient visits, weighted by state population [6]. We used CDC's weekly wILI data from season 1998/1999 to season 2018/2019 (Figure 1), indexing seasons from week 31 (season week 1) to week 30 in the next year. The two seasons affected by the H1N1 pandemic (2008/2009 and 2009/2010) were excluded from our analysis.

[Figure 1 about here.]

Figure 1(b) shows that a peak or secondary peak often occurs in season week 22 (calendar week 52). The observed bump during winter holidays is potentially driven by two opposite effects. First, mildly ill patients are less likely to visit a doctor than usual, thus reducing the number of non-ILI visits and consequently increasing the wILI index [5]. Second, influenza transmission is hampered due to a reduction of work and school contacts [17, 5], thus decreasing wILI in the subsequent week (season week 23). To capture this pattern in our model, we included two dummy variables for season weeks 22 and 23, respectively, in the endemic component.

The wILI proportion has more variation during high incidence periods than in the off-season. This pattern is naturally accounted for by the mean-dependent variance of the beta

distribution, but we also consider more flexible models with harmonic regression terms in the precision parameter. Furthermore, we observe a decreasing fluctuation (increasing precision) of wILI over the first few years of ILINet surveillance (Figure 1(a)). This is likely due to an increase in the number of healthcare data providers over time; there was a particularly large increase after the two pandemic seasons, leading to more accurate proportion data. We thus include a trend $T_S = \min\{13, S\}$ in the precision model, where S is the season index and the 13th season is the first season after the two pandemic seasons.

We used the corrected Akaike information criterion (AICc) to select the number p of autoregressive lags (1 to 5) and the number of harmonics (up to 5) to capture yearly seasonality in both the mean and the precision [18]. This exhaustive search over 180 candidate models resulted in a Beta($p = 4$) model with three harmonics in the mean and four harmonics in the precision. The estimated seasonal patterns are shown in Figure 2. The estimated effect of season week 22 is 0.075 (95% CI: [0.007, 0.143]), which means that the odds for ILI visits increase by $e^{0.075} - 1 \approx 8\%$ compared to the usual flu activity at that time of the year. For season week 23, the corresponding estimate is -0.260 (95% CI: $[-0.332, -0.187]$), meaning a reduction of the odds for ILI visits by roughly 23%. The estimated autoregressive parameters are $\hat{\beta}_1 = 0.990$, $\hat{\beta}_2 = 0.036$, $\hat{\beta}_3 = -0.072$, and $\hat{\beta}_4 = -0.068$, respectively. The estimated parameter of the trend in the precision part is 0.184, corresponding to a 20% increase by season.

[Figure 2 about here.]

To explore the relative contributions of the different model components, we compared the following variants:

- (a) the full Beta(4) model as described above
- (b) model (a) without trend in the precision
- (c) model (a) with constant precision: $\phi_t = \phi$
- (d) model (a) without autoregression: $\beta_k = 0$
- (e) model (a) with identity link for autoregressive effects: $g(X_{t-k}) = X_{t-k}$

We also benchmarked the above models against a seasonal ARIMA (SARIMA) model of the logit-transformed proportion time series, using the R package `logitnorm` [36] to calculate the log-likelihood. The order of the SARIMA model was selected by the AICc-based search implemented in the `auto.arima` function in the R package `forecast` [19]. This resulted in a SARIMA(2,0,0)(1,1,0)[52] model. Note that the selected SARIMA model includes seasonal differencing and thus only fits a subset of the observations. When comparing the different model fits (Table 1), we restricted the likelihood-based scores to the common set of observations (24 observations were omitted).

[Table 1 about here.]

It turns out that the full beta model (a) performs best in terms of all criteria, followed by model (b), which omits the trend parameter in the precision. The non-dynamic beta model (d) fits considerably worse than all other models. Similarly, using an identity link for past observations (e) hugely deteriorates the original performance. The middle ranks are shared by the beta model (c) with constant precision and the SARIMA model.

Plots of Pearson residuals against fitted values for both the beta model (a) and the SARIMA model (not shown) do not reveal any heteroskedasticity, which means that both approaches can accommodate a mean-dependent variance. The SARIMA model for the logit proportions actually assumes a constant variance, but it becomes mean-dependent via back-transformation to the original scale. The beta model is naturally heteroskedastic but becomes even more flexible via the distributional regression approach with seasonal effects in the precision parameter.

The estimated autocorrelation function of the conditional Pearson residuals (Figure 3) indicates that the beta models (a)–(c) have some residual autocorrelation at lag 52. Not allowing for seasonal variation of the precision as in variant (c) seems to slightly increase residual autocorrelation. The SARIMA model has more pronounced residual autocorrelation at lower lags. The non-dynamic model (d) has large, slowly decreasing, residual autocorrelations, which indicates a trend in the residuals. As the wILI proportions mirror an epidemic process, they highly depend on the recent past; adding an autoregressive component can remove the trend in the residuals. Model (e) with an identity link for the autoregressive effects has large residual autocorrelations but also includes periodic patterns. This indicates that the residuals have both a trend and seasonality, although the model contains an epidemic and endemic part. Thus using logit-transformed past observations together with the logit link for the mean can remove the trend and seasonality in residuals.

[Figure 3 about here.]

4.2 Regional influenza-like illness in the USA

Besides the national ILI monitoring, the ILINet system in the USA also reports the wILI index for the 10 Health & Human Service (HHS) regions. A map of these regions and the disaggregated wILI time series from season 1998/1999 to 2018/2019 can be found in the supplementary material (Figures S1 and S2). We consider several multivariate beta models of varying complexity and compare their goodness-of-fit as well as the quality of their rolling one-week-ahead forecasts during the last four seasons (test data).

The basic temporal components of our initial mBeta model were taken to be the same as in the selected univariate beta model for the aggregated national wILI time series in the previous section, but assuming that all effects are region-specific (saturated model).

This means we used a maximum autoregressive lag of $p = 4$ weeks, three harmonics in the mean, four harmonics in the precision including a trend T_S , and dummy variables for season weeks 22 and 23. For the neighbourhood weights, we applied the power-law model.

Estimated parameters and corresponding 95% Wald confidence intervals from this mBeta model are shown in Figure 4. There is obvious heterogeneity across regions, especially for the precision parameters. For the parameters of the mean, most regions have overlapping confidence intervals, but region 1 (Boston) stands out with a relatively high association with past proportions in neighbouring regions (the only direct neighbour being region 2, New York). The estimated power-law decay parameter is $\hat{\rho} = 0.55$ (95% CI: [0.20, 1.49]), which corresponds to neighbourhood weights ($\omega_{r',r}$) of 68%, 55%, 47%, and 41% for adjacency orders 2 to 5, respectively. This relatively weak decay (also compared to a decay of 1.80 for a district-level HHH model of influenza in Southern Germany [22]) could be related to the large size of the HHS regions; even travelling between directly adjacent regions largely reflects occasional long-range contacts.

[Figure 4 about here.]

To find a more parsimonious model, we considered the following variants, ordered by increasing complexity:

- (M1) Only the intercepts of the mean and precision models are region-specific, but all effects are shared across regions: $\beta_{r,k} = \beta_k, \beta_r^{(v)} = \beta^{(v)}, \beta_r^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$
- (M2) M1 with region-specific autoregressive parameters
- (M3) M2 with region-specific seasonality, but excluding the neighbourhood effect, i.e., $\gamma_r = 0$. This model is equivalent to univariate beta models stratified by region.
- (M4) M3 with a homogeneous neighbourhood effect, i.e., $\gamma_r = \gamma$
- (M5) Saturated mBeta model with all effects being region-specific

The goodness of fit of these models is compared in Table 2. The most complex model M5 with a total of 241 parameters has the best AIC and AICc. The performance of the models with respect to these scores decreases from M5 to M1 as more and more effects are assumed homogeneous across regions. By assuming same neighbourhood influence across regions, model M4 saves 9 parameters, and for model M1 only 43 parameters remain, but it also has the worst fit in terms of AIC and AICc. Interestingly, the Bayesian information criterion (BIC) gives almost inverse model ranks. The simplest model M1 has the best BIC and the stratified approach (M3) has the worst BIC. AIC (or AICc) assesses predictive performance more adequately [32], so we would choose the full model M5 if the goal is to produce one-step-ahead forecasts.

[Table 2 about here.]

To verify AICc-driven model choice for time series prediction, we generated rolling one-week-ahead probabilistic forecasts of the above models over the test period. To avoid excessive runtimes, we skipped refitting the decay parameter at each time point and instead kept it fixed at the estimate from the first test time point. We applied proper scoring rules to assess their relative performance, including both the logarithmic score [12] and the Dawid-Sebastiani score [9]. We considered two subsets of the test period for forecast comparison. In the "all weeks" subset, the scores are averaged over the whole test period. The "high incidence" subset only includes season weeks 10 to 42 (or 43 when a season has 53 weeks), where flu activity is relatively high (see Figure 1). Forecasts during this second subset are more important from a public health perspective. In addition to the mean scores, we also report the respective maximum log score as a measure of worst-case performance [28]. We also performed pairwise Monte Carlo permutation tests for differences in mean log scores using the best model in the corresponding subset as the reference.

[Table 3 about here.]

Table 3 shows the obtained scores of forecast performance. In both the "all weeks" and the "high incidence" assessment, model M5 consistently achieves the best average scores and the best maximum log score. Second-best (nearly identical) scores are consistently achieved by model M4. The model ranks in the "all weeks" subset are actually identical to the AICc-based goodness-of-fit assessment from Table 2. In either subset, we see that accounting for neighbourhood effects as in models M4 or M5 improves forecast performance compared to independent univariate forecasts stratified by region (M3). However, this improvement is not statistically significant. Interestingly, even the simplest model M1 with much less parameters than model M5 does not perform significantly worse than model M5. A parsimonious joint model of regional flu activity (M1) could be preferable over fitting multiple univariate models (M3).

Figure S3 in the supplementary material shows fan plots and PIT histograms of the forecasts from the best-performing model M5. For regions 1 and 8, there are many large PIT values, meaning that too many observations fall in the upper tail of the predictive distribution (underprediction). The PIT plots for regions 3, 9 and 10 have a hump shape, indicating overdispersed predictive distributions [12]. The supplementary material also contains forecast results when assuming simple first-order adjacency weights. The rankings of the five models are almost the same as described above, but the first-order models always perform worse than the respective power-law models.

5 Discussion

We proposed an endemic-epidemic beta model for time series of infectious disease proportions and a multivariate extension. It assumes a beta distribution with parameters for

the mean and precision, and can be regarded as a distributional regression model. Autoregressive terms enter on the logit scale to account for the boundedness of proportions. An appealing property of this model formulation is that the complementary proportion process also follows a beta model and will give equivalent results. Furthermore, the beta distribution naturally adapts to the asymmetric shape and heteroskedasticity of proportion distributions. Building on functionality from the R package `betareg`, model estimation is relatively straightforward and fast.

The application to regional flu activity showed that a multivariate modelling approach can improve both the goodness of fit as well as one-step-ahead forecasts of the proportions compared to a stratified analysis by region. Furthermore, when the different regions show similar effects for some covariates, model complexity can be reduced by assuming shared parameters between regions, also to obtain more efficient parameter estimates, in particular for short time series. In our application, such a simplification did not significantly reduce forecast performance.

Other approaches using the beta distribution to model time series of proportions exist. However, to the best of our knowledge, these are all limited to univariate time series analyses. For example, Rocha and Cribari-Neto [29] proposed the beta autoregressive moving average model (β ARMA), which has recently been applied to hydrologic data [31, supplemented with R code]. Guolo and Varin [13] proposed a marginal beta regression time series model, for which likelihood inference is implemented in the R package `gcmr` [21]. Unfortunately, we could not incorporate these two modelling approaches in our comparison in Section 4.1, as their currently available implementations cannot handle missing values in the time series. Our applications of beta regression models benefit from the proper handling of missing values by the R package `betareg`. An assessment of how the univariate beta model performs compared to other readily available forecasting models is described elsewhere [20]. A main result is that the Beta(p) model is competitive in producing short-term forecasts of wILI in the USA, while having a relatively simple model structure and short run time.

There is still room for improvements and extensions for the beta model. First, seasonality in the endemic part could be modelled via stochastic time variable parameters [37], since the amplitude and timing of the peak fluctuate across different years. Second, the beta model could be extended to include weighting schemes for past observations [4], while still using the logit transformation for past observations and the conditional mean. Finally, the beta model could be extended for proportion data with so-called essential zeros or ones by using inflated beta regression models [25, 3].

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

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.

Data availability statement

The data that support the findings of this study were derived from the following resources available in the public domain: *Weekly U.S. Influenza Surveillance Report* (<https://www.cdc.gov/flu/weekly/index.htm>), accessed via the R package `cdcfluview` [30]. The derived datasets as well as the code to reproduce all results are openly available at <https://github.com/Junyi-L/mBeta/>.

ORCID

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

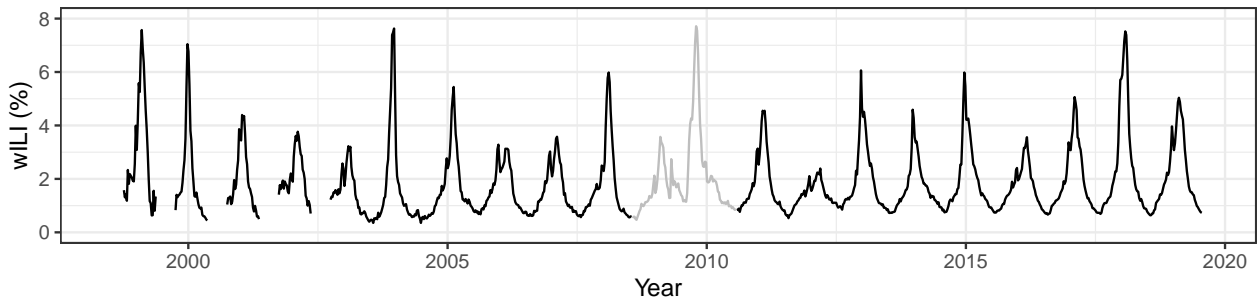
References

- [1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London, UK, 1986.
- [2] C. Barceló-Vidal and L. Aguilar. Time series of proportions: A compositional approach. In A. W. Bowman, editor, *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, UK, 2010.
- [3] C. L. Bayes and L. Valdivieso. A beta inflated mean regression model for fractional response variables. *J Appl Stat*, 43(10):1814–1830, 2016.
- [4] J. Bracher and L. Held. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *Int J Forecast*, 2020. In press.
- [5] L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, and R. Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput Biol*, 14(6):1–29, 2018.
- [6] Centers for Disease Control and Prevention. U.S. influenza surveillance system: Purpose and methods, 2019. Available at <https://www.cdc.gov/flu/weekly/overview.htm> (accessed: 2019-10-30).

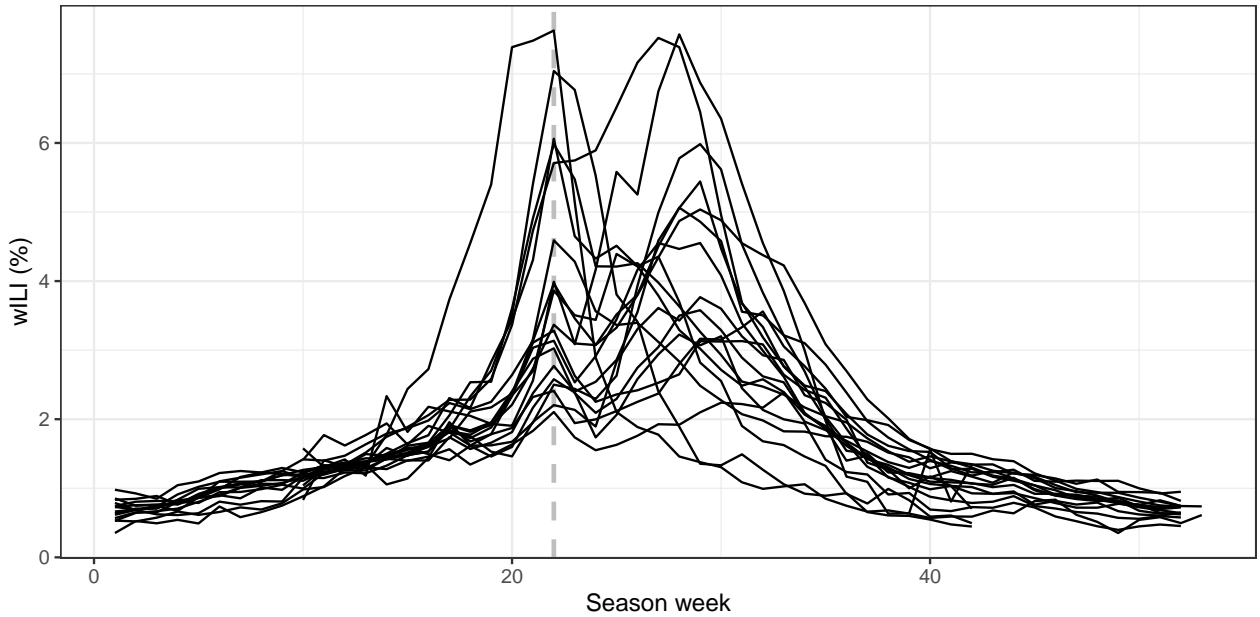
- [7] C. Chiavenna, A. M. Presanis, A. Charlett, S. de Lusignan, S. Ladhani, R. G. Pebody, and D. De Angelis. Estimating age-stratified influenza-associated invasive pneumococcal disease in England: A time-series model based on population surveillance data. *PLoS Med*, 16(6):1–21, 2019.
- [8] F. Cribari-Neto and A. Zeileis. Beta regression in R. *J Stat Softw*, 34(2):1–24, 2010.
- [9] A. P. Dawid and P. Sebastiani. Coherent dispersion criteria for optimal experimental design. *Ann Stat*, 27(1):65–81, 1999.
- [10] M. M. Dickson, G. Espa, D. Giuliani, F. Santi, and L. Savadori. Assessing the effect of containment measures on the spatio-temporal dynamic of COVID-19 in Italy. *Nonlinear Dyn*, 101:1833–1846, 2020.
- [11] S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *J Appl Stat*, 31(7):799–815, 2004.
- [12] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *J Royal Stat Soc Ser B (Stat Methodol)*, 69(2):243–268, 2007.
- [13] A. Guolo and C. Varin. Beta regression for time series analysis of bounded data, with application to Canada Google[®] Flu Trends. *Ann Appl Stat*, 8(1):74–88, 2014.
- [14] L. Held, M. Höhle, and M. Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Stat Model*, 5(3):187–199, 2005.
- [15] L. Held, S. Meyer, and J. Bracher. Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Stat Med*, 36(22):3443–3460, 2017.
- [16] L. Held and M. Paul. Modeling seasonality in space-time infectious disease surveillance data. *Biometrical J*, 54(6):824–843, 2012.
- [17] N. Hens, G. Ayele, N. Goeyvaerts, M. Aerts, J. Mossong, J. Edmunds, and P. Beutels. Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infect Dis*, 9(1):187, 2009.
- [18] C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.
- [19] R. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *J Stat Softw*, 27(3):1–22, 2008.
- [20] J. Lu and S. Meyer. Forecasting flu activity in the United States: Benchmarking an endemic-epidemic beta model. *Int J Environ Res Public Health*, 17(4):1381, 2020.

- [21] G. Masarotto and C. Varin. Gaussian copula regression in R. *J Stat Softw*, 77(8):1–26, 2017.
- [22] S. Meyer and L. Held. Power-law models for infectious disease spread. *Ann Appl Stat*, 8(3):1612–1639, 2014.
- [23] S. Meyer and L. Held. Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*, 18(2):338–351, 2017.
- [24] S. Meyer, L. Held, and M. Höhle. Spatio-temporal analysis of epidemic phenomena using the R package *surveillance*. *J Stat Softw*, 77(11):1–55, 2017.
- [25] R. Ospina and S. L. Ferrari. A general class of zero-or-one inflated beta regression models. *Comput Stat Data An*, 56(6):1609–1623, 2012.
- [26] M. Paul, L. Held, and A. Toschke. Multivariate modelling of infectious disease surveillance data. *Stat Med*, 27(29):6250–6267, 2008.
- [27] V. Pawlowsky-Glahn and J. J. Egozcue. Geometric approach to statistical analysis on the simplex. *Stoch Env Res Risk A*, 15(5):384–398, 2001.
- [28] E. L. Ray, K. Sakrejda, S. A. Lauer, M. A. Johansson, and N. G. Reich. Infectious disease prediction with kernel conditional density estimation. *Stat Med*, 36(30):4908–4929, 2017.
- [29] A. V. Rocha and F. Cribari-Neto. Beta autoregressive moving average models. *Test*, 18(3):529–545, 2008.
- [30] B. Rudis. *cdcfluview: Retrieve Flu Season Data from the United States Centers for Disease Control and Prevention ('CDC') 'FluView' Portal*, 2019. R package version 0.9.0. Software available at <https://CRAN.R-project.org/package=cdcfluview>.
- [31] V. T. Scher, F. Cribari-Neto, G. Pumi, and F. M. Bayer. Goodness-of-fit tests for β ARMA hydrological time series modeling. *Environmetrics*, 31(3):e2607, 2020.
- [32] G. Shmueli. To explain or to predict? *Stat Sci*, 25(3):289–310, 2010.
- [33] A. B. Simas, W. Barreto-Souza, and A. V. Rocha. Improved estimators for a general class of beta regression models. *Comput Stat Data An*, 54(2):348–366, 2010.
- [34] P. Ssentongo, C. Fronterre, A. Geronimo, S. J. Greybush, P. K. Mbabazi, J. Muvawala, S. B. Nahalamba, P. O. Omadi, B. T. Opar, S. A. Sinnar, Y. Wang, A. J. Whalen, L. Held, C. Jewell, A. J. B. Muwanguzi, H. Greatrex, M. M. Norton, P. J. Diggle, and S. J. Schiff. Pan-African evolution of within- and between-country COVID-19 dynamics. *Proc Natl Acad Sci*, 118(28):e2026664118, 2021.

- [35] J. Wakefield, T. Q. Dong, and V. N. Minin. Spatio-temporal analysis of surveillance data. In L. Held, N. Hens, P. D. O'Neill, and J. Wallinga, editors, *Handbook of Infectious Disease Data Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, chapter 23, pages 455–476. Chapman & Hall/CRC, Boca Raton, Florida, USA, 2019.
- [36] T. Wutzler. *logitnorm: Functions for the Logitnormal Distribution*, 2018. R package version 0.8.37. Software available at <https://CRAN.R-project.org/package=logitnorm>.
- [37] P. C. Young, D. J. Pedregal, and W. Tych. Dynamic harmonic regression. *J Forecasting*, 18(6):369–394, 1999.



(a) Weekly wILI time series. Excluded seasons are in grey. Year-round data is only provided since 2004.



(b) Seasonal wILI time series, excluding seasons 2008/2009 and 2009/2010. Season week 22 (calendar week 52) is indicated with a vertical dashed line, where a peak or secondary peak occurs in most seasons.

Figure 1: Weekly weighted national influenza-like illness (wILI) in the USA for flu seasons 1998/1999 through 2018/2019.

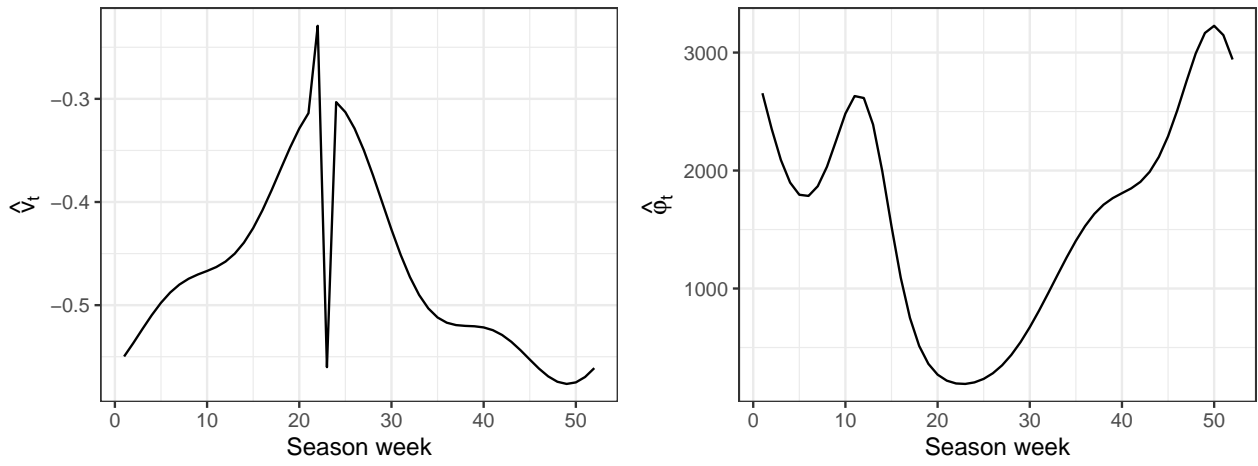


Figure 2: Estimated endemic component ν_t (left), and estimated seasonal variation of the precision parameter ϕ_t without trend (right).

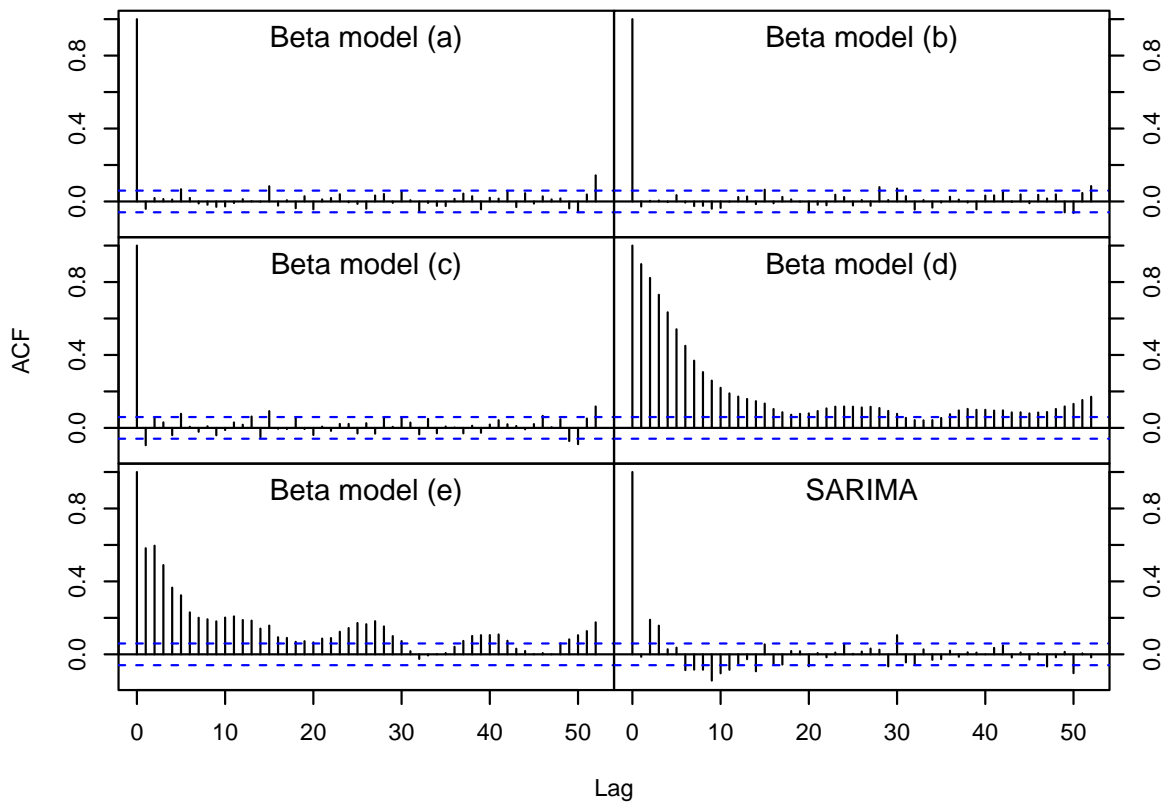


Figure 3: Estimated autocorrelation function (ACF) of conditional Pearson residuals. The dotted lines are thresholds for significance at the 5% level for an uncorrelated time series.

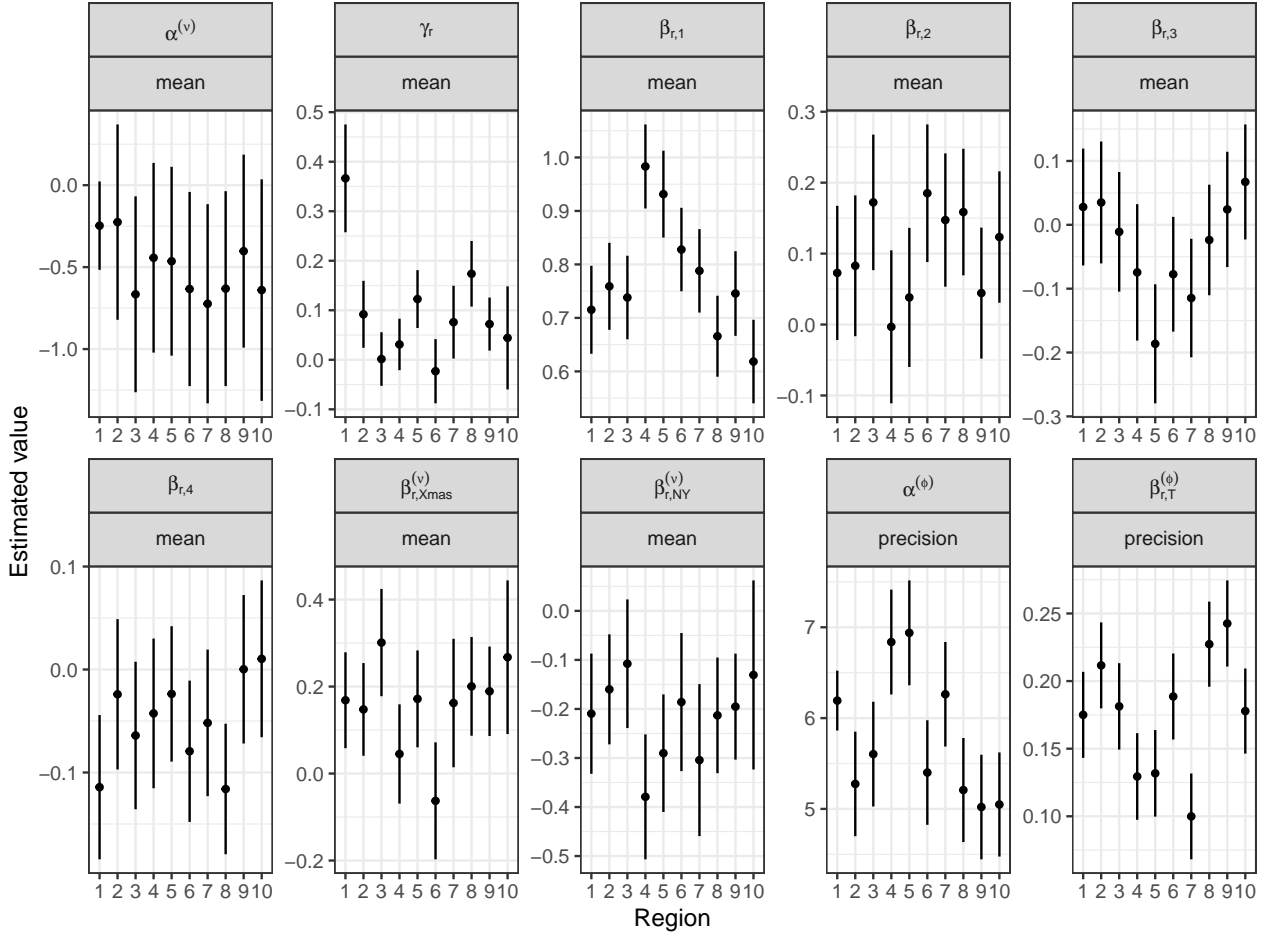


Figure 4: Estimated region-specific parameters and corresponding 95% Wald confidence intervals from the saturated mBeta model (M5). The parameters $\alpha^{(\nu)}$ and $\alpha^{(\phi)}$ denote the intercepts of the mean and precision models, respectively. The parameters γ_r , $\beta_{r,NY}^{(\nu)}$, and $\beta_{r,Xmas}^{(\nu)}$ refer to the neighbourhood, New Year, and Christmas effects, respectively. The parameters $\beta_{r,1}$ to $\beta_{r,4}$ are the autoregressive coefficients. The parameter $\beta_{r,T}^{(\phi)}$ refers to the trend coefficient in the precision. Coefficients of the harmonics are not shown.

Model	Variant	LL	npar	AIC	AICc	BIC
Beta	(a) full model	4489 (1)	23	-8933 (1)	-8931 (1)	-8823 (1)
Beta	(b) no trend in ϕ_t	4380 (2)	22	-8716 (2)	-8715 (2)	-8611 (2)
Beta	(c) $\phi_t = \phi$	4223 (3)	14	-8418 (3)	-8417 (3)	-8351 (4)
Beta	(d) $\beta_k = 0$	3700 (6)	19	-7361 (6)	-7360 (6)	-7270 (6)
Beta	(e) $g(X_{t-k}) = X_{t-k}$	4138 (5)	23	-8231 (5)	-8230 (5)	-8121 (5)
SARIMA	(2,0,0)(1,1,0)[52]	4197 (4)	5	-8384 (4)	-8384 (4)	-8360 (3)

Table 1: Goodness-of-fit criteria for univariate models of the national wILI time series. Ranks are shown in parantheses. The log-likelihood (LL) is ranked descending, and AIC, AICc, BIC are ranked ascending. The "npar" column gives the number of estimated parameters.

Model	LL	AIC	AICc	BIC	npar
M1: $\beta_{r,k} = \beta_k, \beta_r^{(v)} = \beta^{(v)}, \beta_r^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$	26119 (5)	-52152 (5)	-52151 (5)	-51864 (1)	43
M2: $\beta_r^{(v)} = \beta^{(v)}, \beta^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$	26212 (4)	-52266 (4)	-52264 (4)	-51738 (2)	79
M3: $\gamma_r = 0$	26459 (3)	-52458 (3)	-52440 (3)	-50919 (5)	230
M4: $\gamma_r = \gamma$	26485 (2)	-52505 (2)	-52487 (2)	-50953 (3)	232
M5: full model	26508 (1)	-52533 (1)	-52513 (1)	-50921 (4)	241

Table 2: Goodness-of-fit criteria for different mBeta models of the regional wILI time series. Models are ordered by complexity (number of estimated parameters, "npar"). Ranks are shown in parantheses. The log-likelihood (LL) is ranked descending, and AIC, AICc, BIC are ranked ascending.

Model	Subset	LS	p-value	maxLS	DSS
M1: $\beta_{r,k} = \beta_k, \beta_r^{(v)} = \beta^{(v)}, \beta_r^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$	All weeks	-47.13 (5)	0.25	-45.58 (5)	-112.22 (5)
M2: $\beta_r^{(v)} = \beta^{(v)}, \beta_r^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$		-47.20 (4)	0.59	-45.84 (4)	-112.34 (4)
M3: $\gamma_r = 0$		-47.21 (3)	0.36	-46.02 (3)	-112.37 (3)
M4: $\gamma_r = \gamma$		-47.25 (2)	0.68	-46.11 (2)	-112.43 (2)
M5: full model		-47.26 (1)		-46.16 (1)	-112.43 (1)
M1: $\beta_{r,k} = \beta_k, \beta_r^{(v)} = \beta^{(v)}, \beta_r^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$	High Incidence	-43.89 (5)	0.46	-42.34 (5)	-105.74 (5)
M2: $\beta_r^{(v)} = \beta^{(v)}, \beta_r^{(\phi)} = \beta^{(\phi)}, \gamma_r = \gamma$		-43.98 (3)	0.85	-42.63 (4)	-105.92 (3)
M3: $\gamma_r = 0$		-43.93 (4)	0.29	-42.93 (3)	-105.87 (4)
M4: $\gamma_r = \gamma$		-44.01 (2)	0.98	-43.12 (2)	-106.04 (2)
M5: full model		-44.01 (1)		-43.17 (1)	-106.06 (1)

Table 3: Model performance in terms of mean log score (LS), mean Dawid-Sebastiani score (DSS), and maximum log score (maxLS) for one-week-ahead forecasts. Ranks are shown in bracket. The "all weeks" group shows average scores over the whole test period (208 weeks), whereas the "High Incidence" group shows averages over the high incidence periods only (132 weeks). Models are ordered by model complexity. The Monte Carlo p-values for differences in mean log scores are based on 9999 random permutations, comparing each model against the best model (M5) in each subset.