Spatio-temporal forecasting and infectious disease count data

Master Thesis in Biostatistics (STA495)

by

Kelly Reeve
15-721-053

supervised by

Prof. Dr. Leonhard Held

Zurich, August 23, 2017

Abstract

The purpose of this thesis is to demonstrate the use of an analytical method for producing long-term predictions (Held et al., 2017) within the HHH model framework (Held et al., 2005) while studying the effect of spatio-temporal model components on prediction of infectious disease. Of interest is whether inclusion of more sophisticated spatial model components improves model prediction, and if these improvements are still detected when aggregating over space to obtain weekly count predictions or over space and time for final count predictions. This exercise is threefold: (1) to analyze 2008 southern Germany influenza incidence data in order to make direct comparisons to the results of Meyer and Held (2014), which were obtained using one-step-ahead and simulated predictions, (2) to broaden the scope of models considered in Meyer and Held (2014) by including border effect covariates and gravity measures other than population in order to further study the effects of spatial components in modeling, and (3) to evaluate an updated dataset containing counts through the 2017 epidemic season and thereby increasing the training set on which the models are fit.

This research has shown that power law models with and without gravity components often perform better than baseline models when evaluated with the multivariate Dawid-Sebastiani score. The top models indicated are corroborated by univariate results and results from weekly and regional scores. Unusually large variance is often observed in autoregressive and first order model predictions, and even some power model predictions. The sensitivity of models including space to small changes in conditioning week is also noteworthy and should be researched further.

Notation

The notation for this thesis closely follows that of Meyer and Held (2014). Scalar parameters, such as $\alpha^{(\nu)}$, are represented by lowercase letters. Vectors, e.g. $\boldsymbol{\mu_P}$, and matrices, e.g. $\boldsymbol{\Sigma_P}$, are printed in bold, with matrices also represented by uppercase letters. The transpose of a vector or matrix is denoted by $^\top$, as in $(\boldsymbol{x} - \boldsymbol{\mu_P})^\top$, for example.

Software

All analyses were performed in `R` (version 3.3.3 (2017-03-06)), a free software environment for statistical computing and graphics which is available at `https://www.r-project.org/`. The following model-specific packages were used: `surveillance` (version 1.13.1), `hhh4predict` (version 0.1.0.7), and `hhh4contacts` (version 0.12.1). In addition to base packages, `ggplot2` (2.2.1), `knitr` (1.15.1), and `parallel` (3.3.3) were used in analysis and reporting. The computing environment had the following specifications: Ubuntu 16.04.2 LTS (Xenial Xerus) and Intel Xeon E312xx (Sandy Bridge) x 6 (Processor).

Acknowledgements

I would like to thank my thesis supervisor Prof. Dr. Leonhard Held for giving me the opportunity to work on a topic in the field of infectious disease modeling. Sincere thanks also go to Johannes Bracher for all his time and thorough explanations. I've learned so much under their guidance and through their discussions. A big thanks goes to Chelsea Little for her comments. I would like to thank all of the people who make the Master Program in Biostatistics possible and Dr. Eva Furrer in particular.

Thanks are also due to my fellow students. Thank you for the full day study sessions and computer lab companionship. My friends outside the program are also deserving of innumerable thanks for all the ways they have supported me. Thank you Steve Towler. Thank you Becki Witt, Claire Beck, and Nicola Hautle. Finally, the biggest of thanks to Damian Kozbur for his support and encouragement, his explanations and corrections, and for his time.

Contents

1 **Introduction** 4

2 **Materials and Methods** 6

2.1 Data 6

2.2 Models 7

2.3 Prediction 11

2.4 Evaluation 12

2.5 Aggregated data versus “sliced” data 14

3 **Exploratory data analysis** 15

3.1 Spatio-temporal characteristics of influenza in Baden-Württemberg and Bayern 15

3.2 Comparison of gravity measures 18

4 **Influenza surveillance data (`fluBYBW`) from 2001-2008** 21

4.1 Applying analytical method for long-term forecasts to `fluBYBW` 21

4.2 Extending Meyer and Held (2014)’s `fluBYBW` analysis 25

5 **Updated influenza surveillance data for 2001-2017** 34

5.1 Analysis by year 34

5.2 Effect of changes in conditioning week 56

6 **Conclusion** 57

7 **Appendix** 62

7.1 Fitting the models 62

7.2 Incidence maps for each year from 2001-2017 66

7.3 Fanplots for models predicting 2008 incidence 70

7.4 Maps of predicted incidence in 2008 for each model 73

7.5 Long-term predictions based on various conditioning weeks for selected models. 74

7.6 Additional tables 78

1 Introduction

Modern mathematical modeling of infectious disease now has over a century-long history. These models have often been used to understand the mechanisms of spread of disease and to estimate the duration and size of epidemics. However, another major and increasingly important purpose of infectious disease modeling is prediction of disease outbreak and quantification of the uncertainty related to that prediction. Accurate predictions give health professionals and policymakers the information needed to prepare for upcoming epidemics.

Models for infectious diseases have been employed as early as the 18th century in works by Bernoulli and Laplace, however, it was not until Ross (1911) that systematic development of these models started to occur (Siettos and Russo, 2013). Following up on this work, Kermack and McKendrick (1927) are credited with creation of the SIR model, variants of which are still used to this day to understand disease dynamics and explore the possible effects of interventions. This model framework compartmentalizes populations into susceptibles, infectives, and recovereds and describes their dynamics through a set of differential equations. Conclusions are drawn from numerical solutions to these equations. While deterministic models are very useful for understanding disease dynamics, they do not incorporate the uncertainty found in epidemics.

To address the issue of uncertainty, numerous extensions of the SIR model with stochastic components have been proposed, but general statistical models have also been developed to account for uncertainty while explaining disease dynamics and making predictions. Such statistical models include: general linear models (Goldstein et al., 2011; Andersson et al., 2008; Mooney et al., 2002), regression trees (Sočan et al., 2012), classification models (Nsoesie et al., 2011), Bayesian networks (Sebastiani et al., 2006), and time series models (Box et al., 2016; Held et al., 2005).

Even before John Snow and the Broad Street pump, people have understood the importance of space to infectious disease. But now, with increasing computing power and availability of spatio-temporal data, simplistic assumptions about disease transmission in space can be dropped for more complex, realistic spatial models which can improve the epidemic forecasts so important to public health and policy. However, of the thirty-five studies included in Chretien et al. (2014)'s review of current research on influenza forecasting, only seven made forecasts in time and space, while twenty-eight employed only temporal forecasting. It is clear from the review that further research should be done on spatio-temporal forecasting.

Two of the studies making spatio-temporal forecasts of influenza mentioned in Chretien et al. (2014) use the HHH model, the multivariate time-series model framework for aggregated surveillance data first proposed by Held et al. (2005). Since first being proposed, this model has been further developed in Paul et al. (2008), Paul and Held (2011), and Meyer and Held (2014). Additionally, an analytical method for obtaining the first two moments of multivariate path forecasts produced within the HHH framework is presented in Held et al. (2017). Meyer and Held (2014) propose modeling infection transmission in space using a power law for short-time human travel. Together, these studies allow for long-term predictions of infectious disease in both space and time.

The review of Chretien et al. (2014) also highlights the need for good practices in prediction and assessment of model calibration. Until recently there was very little in the epidemiological literature about prediction evaluation measures (Funk et al., 2016; Chretien et al., 2014; Gneiting et al., 2007). Across studies, numerous metrics for forecast validation are used, with mean/median absolute error and mean absolute percent error the most common, but varying from comparison of number of observations in the predicted credible intervals to correlation and t-tests to no quantitative metric at all (Chretien et al., 2014). Several studies argue for the use of absolute error metrics which reduce the impact of outliers and increase interpretability, and present forecast evaluation frameworks based on them (Hyndman and Koehler, 2006; Reich et al., 2016).

However, Gneiting (2011) illustrates how such techniques can lead to grossly misguided inference when the use of any one evaluation metric becomes automatic for all types of prediction. He argues that target functionals, such as expectations or quantiles, should be specified and scoring functions should be used that are consistent for that target functional. Scoring functions (also scoring rules) assess the quality of a forecast by assigning a numerical score based on the predictive distribution and observations. Gneiting and Raftery (2007) and Gneiting et al. (2007) propose the use of proper scoring rules in predictive performance evaluation, with a focus on maximizing sharpness subject to forecast calibration. Applications to count data are further described by Czado et al. (2009) and Wei and Held (2014). Most of these methods are based on the full predictive distribution and some only on moments, but all take into account the uncertainty in the predictions.

A related problem in the current literature is the lack of research on the contributions of different model components to improved prediction (Johansson et al., 2016). One exception is the study of Yang et al. (2016), in which they examine whether the inclusion of spatial dynamics improves influenza forecast at different observation scales (borough or neighborhood) in New York City. They find spatial network data to improve borough-level predictions but degrade neighborhood-level predictions, possibly due to poor signal to noise ratios at this finer granularity. Additionally, Meyer and Held (2014) find power models to produce better predictions of final counts, epidemic curves, and regional final counts. Held et al. (2017) also find power model predictions to perform better than predictions from models with simpler spatial assumptions at several aggregation levels. Further research exploring predictions at different spatial scales, however, is needed.

In this study we investigate the contribution of more complex spatio-temporal components in the HHH model framework to forecast performance. This exercise is threefold: (1) to analyze 2008 southern Germany influenza incidence data in order to make direct comparisons to the results of Meyer and Held (2014), which were obtained using one-step-ahead and simulated predictions, (2) to broaden the scope of models considered in Meyer and Held (2014) by including border effect covariates and gravity measures other than population in order to further study the effects of spatial components in modeling, and (3) to evaluate an updated dataset containing counts through the 2017 epidemic season and thereby increasing the training set on which the models are fit. Our study can then be seen as an extension of Meyer and Held (2014), in which we evaluate analytical forecast distributions by multivariate proper scoring rules (instead of univariate score evaluation of simulated forecasts) at various aggregation levels. Of interest is whether inclusion of more sophisticated spatial model components improves not only model fit, but also predictions, and at what aggregation levels these improvements are observed. This includes whether the incorporation of complex spatio-temporal components significantly contributes to prediction accuracy, even when aggregating over space to obtain weekly count predictions or over space and time for final count predictions. In this way, we contribute to a growing body of knowledge about application of proper scoring rules to forecast evaluation in the epidemiological setting and the utility of spatio-temporal model components to various types of forecasts, from those at the finest granularity to final counts for entire epidemic seasons.

2 Materials and Methods

Using the HHH multivariate time-series model framework for aggregated surveillance data, as presented in Meyer and Held (2014), and the analytical method for obtaining the first two moments of multivariate path forecasts presented in Held et al. (2017), long-term forecasts of influenza are made.

First, the models and data of Meyer and Held (2014) are considered. In Section 4, the models are trained on data from years 2001 to 2007 and 2008 data acts as a hold-out sample with which we evaluate model predictive performance. To study the effect of further data aggregation on predictive performance, the spatio-temporal data is aggregated over region, time, and then over both dimensions to yield weekly counts, regional final counts, and a total final count. The predictions are compared to the actual observations of the hold-out sample period and proper scoring methods allow predictive performance of different models to be compared.

Secondly, we broaden the scope of models considered in Meyer and Held (2014) by considering (1) two additional gravity measures and (2) border effects, described in Subsection 2.2. Additionally, mean weekly scores and mean regional scores are calculated by considering only the predicted count, observed count, and covariance for a given week/region. This method is further described in Subsection 2.5.

Thirdly, an updated influenza count dataset is considered. Counts for years 2009 through the epidemic season of 2017 have been added, and counts from the original study period of 2001 to 2008 have been updated. In Section 5, a recursive sampling approach is used on this data. First, models are trained on updated 2001 to 2007 data with updated 2008 data as the hold-out sample, then the models are refit using this 2001 to 2008 data and 2009 is used as the hold-out sample. This is repeated so that each additional year of data is first used as the hold-out sample and then incorporated into the training data when the following year is used as the hold-out sample.

2.1 Data

Data on incidence of notifiable diseases in Germany can be freely obtained from the Robert Koch Institute's (RKI) web query (SurvSat@RKI 2.0[1]). RKI is the center for disease surveillance and prevention under the German Federal Ministry of Health. It is responsible for maintaining databases of notifiable diseases reported under the German 'Protection against Infection Act' (*Infektionsschutzgesetz*, IfSG). This law was enacted in 2000 to enable early detection and control of infection. Under IfSG, influenza is classified as a notifiable disease, but only confirmed cases (i.e., laboratory-confirmed) must be reported.

[1] `https://survstat.rki.de/Default.aspx`

Seasonal influenza is an acute respiratory infection which circulates in all parts of the world. The influenza virus spreads easily from person to person via infectious droplets. It is characterized by fever, cough, sore throat, runny nose, headache, muscle/joint pains, and fatigue, which take most people less than a week to recover from without medical attention. Symptoms begin, on average, 2 days after exposure and the estimated time between onset of symptoms in a primary and secondary case is 3.6 days (95% CI 2.9-4.3 days) (Cowling et al., 2009). In temperate climates, such as that found in southern Germany, seasonal epidemics tend to occur during the winter (World Health Organization, 2016).

The number of influenza cases reported weekly from January 2001 through December 2008 from each of the 140 counties of the southern German states Bayern and Baden-Württemberg was obtained by Paul and Held (2011) and updated by Meyer and Held (2014). It is available as the ready-to-use dataset `fluBYBW` in the `R` package `surveillance` (Höhle et al., 2016) and is analyzed in Section 4. This enables comparison with the results of Meyer and Held (2014). The finest available granularity of the incidence data is at the week-county level. Additional county data, including area and populations of largest towns (as of 31.12.2001), are obtained from the Federal Statistical Office of Germany. Area and town population are used when creating the additional gravity measures density and urbanicity.

Since the last retrieval from RKI's SurvStat, data for several additional years has become available. In Section 5, `fluBYBW` is supplemented with data from 2009 through the epidemic season of 2017, allowing for analysis through 2017. Figure 2 displays the time series of influenza incidence for 2001 to 2017. Analysis of the additional data starts with predictions of the 2011 epidemic season. This is due to the 2009 influenza pandemic. Extremely high influenza counts are observed during the second half of 2009, followed by a lack of epidemic season in 2010. For model fitting purposes, the 2009 and 2010 data is joined by subsetting the first half of 2009 and the second half of 2010 and considering the combination of these two subsets to be a single year.

2.2 Models

As in the work of Meyer and Held (2014), we use the HHH multivariate time-series model to describe disease incidence from aggregated spatio-temporal surveillance data. This model additively decomposes incidence into endemic and epidemic components to account for exogenous factors and infectiousness, respectively. The epidemic components are driven by past counts in the region of interest and in neighboring regions, while the endemic component captures exogenous factors such as seasonality and population demographics.

Here we describe the model more formally. Disease counts in periods $t = 1, \dots, T$ and regions $i = 1, \dots, I$ are denoted by Y_{it} and counts in the previous time period by $\boldsymbol{Y}_{\cdot,t-1}$. The counts are assumed to be conditionally negative binomially distributed

$$Y_{it} | \boldsymbol{Y}_{\cdot,t-1} \sim \text{NegBin}(\mu_{it}, \psi)$$

with mean

$$\mu_{it} = \nu_{it} e_{it} + \lambda_{it} Y_{i,t-1} + \phi_{it} \Sigma_{j \neq i} w_{ji} Y_{j,t-1}$$

and overdispersion parameter ψ. Note that this implies

$$\operatorname{Var}(Y_{it}|\boldsymbol{Y}_{\cdot,t-1}) = \mu_{it}(1 + \psi\mu_{it}).$$

The unknown quantities ν_{it}, λ_{it}, and ϕ_{it} are log-linear predictors of the form

$$\log(\nu_{it}) = \alpha^{(\nu)} + b_i^{(\nu)} + \beta^{(\nu)\top}\boldsymbol{z}_{it}^{(\nu)};$$

$$\log(\lambda_{it}) = \alpha^{(\lambda)} + b_i^{(\lambda)} + \beta^{(\lambda)\top}\boldsymbol{z}_{it}^{(\lambda)};$$

$$\log(\phi_{it}) = \alpha^{(\phi)} + b_i^{(\phi)} + \beta^{(\phi)\top}\boldsymbol{z}_{it}^{(\phi)}.$$

This form allows for fixed intercepts $\alpha^{(\cdot)}$, region-specific intercepts $b_i^{(\cdot)}$, and exogenous covariates $\boldsymbol{z}_{it}^{(\cdot)}$ in each model compartment. Population fraction, population density, urbanicity, and border effects are used as covariates and explained in detail below. The region-specific intercepts are random effects accounting for heterogeneity between the regions, such as from differences in case reporting. The random effects are assumed to be independent and identically distributed across i, but can be correlated across different model components. They follow a trivariate normal distribution

$$\boldsymbol{b}_i := (b_i^{(\lambda)}, b_i^{(\phi)}, b_i^{(\nu)})^\top \sim \mathcal{N}_3(0, \boldsymbol{\Sigma}_{\boldsymbol{b}})$$

with mean zero and unknown covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}}$.

Several submodels are worth consideration. First we describe our three baseline models: the endemic model, the endemic+autoregressive model, and the first order model. The simplest baseline model, the endemic model, contains just the endemic component $e_{it}\nu_{it}$, where ν_{it} is the endemic mean and e_{it} is a multiplicative unit-specific offset for population fraction. The log-linear predictor includes a linear trend and an $S = 3$ harmonic wave with respect to time. The next baseline model contains both the endemic component and first-order autoregressive component, where λ_{it} is a log-linear predictor with an $S = 1$ sinusoidal wave and $Y_{i,t-1}$ are counts in region i at the previous timepoint. The most complex baseline model, the first order model, contains all three components: the endemic, autoregressive, and neighborhood components. This last component models infection transmission from neighboring regions ($j \neq i$). The transmission weights ω_{ji} model the flow of infection from j to i and in this model are assumed to be known. The first order model is limited to first order, or adjacent, neighbors with weights

$$w_{ji} = \mathbb{1}(j \sim i)/|k \sim j|.$$

Here "$\sim$" means "is adjacent to", so the normalized weights allow the neighbors j of i to distribute their cases uniformly to all their neighbors k.

Now we describe several models of interest.

Power-law

Motivated by Brockmann et al. (2006), who found human travel behavior to be well described by a decreasing power law of the distance $f(x) \propto x^{-d}$, Meyer and Held (2014) propose inclusion of a power law for distance in the neighborhood component of the model in order to consider transmission from neighbors of all adjacency orders. Its characteristic heavy tail, or slow convergence to zero, allows for occasional long range infection transmission in addition to the more probable short distance transmissions. Unlike in the first order model, within the power law model higher order, or nonadjacent, neighbors are taken into account and weights w_{ji} are parametrically estimated.

Brockmann and Helbing (2013) found that effective distances can reveal patterns in spatio-temporal spread not easily seen when using geographic distances. As in Meyer and Held (2014), distance here is measured discretely by the neighborhood order. Two regions are k-th order neighbors if the shortest route between them crosses into k distinct regions. Neighborhood order can then be expressed as a symmetric square matrix with dimensions equal to the number of regions and with zeros on the diagonal. The row-normalized weight matrix, generalized from first-order to higher order, is now made up of

$$w_{ji} = \frac{o_{ji}^{-d}}{\sum_{k=1}^{I} o_{jk}^{-d}}, \text{ for } j \neq i.$$

The weights represent the strength of transmission from j to i and are a function of adjacency order o_{ji}. The power law decay parameter d is considered unknown and estimated from the data. As d increases, the influence of higher order neighbors decreases.

Gravity models

Gravity models, common in social sciences such as economics and human geography, measure interactions between all possible location pairs and assume the flows between them are a function of the location attributes. In the past, spatial coupling was assumed to be an inverse function of distance, however, this may be too simplistic for the complex interactions between humans. Xia et al. (2004) propose a gravity model for regional spread of infectious disease and this motivates Meyer and Held (2014) to account for commuter-driven spread by scaling region i's susceptibility (neighborhood component) by its population fraction. The main idea is that larger cities tend to attract more people. The population data is incorporated as covariate $z_{it}^{(\phi)}$ in the power model.

Inspired by Kafadar and Tukey (1993) and Goodall et al. (1998), in addition to population, we use two other "gravity measures" to model "attractiveness" across regions. Kafadar and Tukey (1993) propose a new urbanicity measure based on the size of the largest place within a region for characterizing urban effects on epidemiological variables or for controlling for urban effects in analyses. Because typical measures, such as total population, population density, and percent urban, are not based on cities, their results are sometimes inconsistent with what one might consider a reasonable urban-rural scale. For example, the population fractions of the urban county Fürth (0.004851897) and the rural county Kelheim (0.004837113) are similarly ranked at 90th and 91st in southern Germany, but their densities are 1756.2 and 104 people per square kilometer, respectively. In this case, population fraction fails to distinguish between an urban center and a county with nature parks and farms.

In this study we compare the following gravity measures: population fraction, population density, and urbanicity. Population fraction is given by the function `population()` in the package `surveillance`. Population density was calculated by dividing the population fraction by area data obtained from *Statistisches Bundesamt* (Destatis)[2].Goodall et al. (1998) suggest three variations on urbanicity: the population of the largest subunit in each unit, the square root of the sum of the squared population of the top three largest subunits in each unit, and the square root of the sum of the squares of all subunit populations. In this study, our units are the 140 counties (*die Landkreise, die Stadtkreise*) of southern Germany and the subunits are the towns (*die Gemeinden*). Some counties (*die Stadtkreise*), such as the city of Munich, contain no smaller subunits and so we use the population of the largest subunit as our measure of urbanicity.

The actual code to fit each of these models can be found in the appendix (Section 7.1).

Inclusion of border effects

Extending the work of Meyer and Held (2014), we also include border covariates. Motivation can be found in the 2016 regional yearbook produced by the Statistical Office of the European Union. The number of employed people in the EU region (EU-28) reached 220.7 million in 2015. Although most of these workers lived in the same region as their place of work, 8.1% commuted to another region, with 0.9% commuting internationally. Of the top 20 regions with the largest number of national commuter outflows, 7 are in Germany. One such region is Rheinhessen-Pfalz, which lends 37.0% of its commuters to Karlsruhe, found in the northwestern section of our observation region. Additionally, the German NUTS 2 region of Freiburg is characterized by high numbers of cross-border outbound commuters. It is clear that in spatial analysis of infectious disease within Germany, border effects may play a significant role.

Apart from sharing a border with the rest of Germany, Bayern shares a border with the Czech Republic and Austria and Baden-Württemberg shares a border with Switzerland and France. While the neighborhood component of our model explicitly accounts for influenza transmission from neighboring counties within the observed region, possible unobserved transmission from the bordering countries are only accounted for in the endemic component in a broad way, along with other exogenous transmissions. Meyer and Held (2014) found the pattern in the estimated region-specific endemic random effects of invasive meningococcal disease models to support border effects, but the estimates for the influenza data, which we also use here, did not. Nonetheless, a detailed investigation of such effects here may shed light on general patterns that could arise in other applications.

For each of the 3 baseline and 4 power models, distance to the border is included as covariate z_{it} in either the endemic or neighborhood component (except in the endemic and autoregressive models in which no neighborhood component exists). The covariate takes one of the following forms: a discrete index, a log discrete index, inverse of the discrete index, or a binary indicator. The discrete index quantifies the distance of a county to the border in terms of the least number of counties that must be crossed to reach the border, similar to the order of neighborhood.

The binary indicator and the order to the border can be generated and checked with R-code 1.The function `surveillance::polyAtBorder` determines which counties have coordinates in common with the outline of the entire region. The function after determines the order a county is removed from the border. Visual confirmation for each covariate is

[2] `https://www-genesis.destatis.de/genesis/online`, code: 11111-0002, Size of territory: Administrative districts, reference date

found in Figure 1.

```
## make the binary indicator for being at border (Meyer and Held, 2014)
atBorder <- polyAtBorder(fluBYBW@map, snap=1e-5, method="polyclip")
plot(fluBYBW@map, col=atBorder) # visual check

## get the order from the border, written by J. Bracher
distance_from_margin <- function(ne_matrix){
  atBorder <- polyAtBorder(fluBYBW@map, snap=1e-5, method="polyclip")
     # atBorder=TRUE/FALSE for each region
  i <- 1 # start
  # choose sufficiently large number so that i always < number
  while(any(atBorder == 0) & i < 50){
    # add 1 to regions of distance < i to border
    atBorder <- atBorder + ((ne_matrix %*% atBorder) > 0)
    i <- i + 1 # next value
  }
  return(max(atBorder) - atBorder) # reverse scale
}
dist <- distance_from_margin(nbOrder1)
plot(fluBYBW@map, col = dist) # visual check
```

Code Example 1: This code generates a binary variable for whether a county is at the observation border and a variable for the discrete order of a county to the border and maps these variables for visual verification.

2.3 Prediction

Primarily, long-term predictions are made for the first 20 weeks of 2008 using models fit to the data from 2001 through 2007 and conditioning on the last week of 2007. This is accomplished using the function `longterm.prediction.hhh4` from package `hhh4predict`, which generates the prediction based on the first two moments of the fitted model. For example, R-code 2 makes predictions for the first 20 weeks of 2008 for a list of model fits conditioned on the final week of 2007 (week 364 of the data).

```
flupreds <- mclapply(flufits, longterm_prediction_hhh4,
                     t_condition=364, lgt=20,
                     mc.preschedule=FALSE, mc.cores=4)
```

Code Example 2: Long-term predictions with a length of 20 weeks and conditioned on week 364 (calendar week 52 of 2007).

The first two conditional moments are the conditional expected value $\mathbb{E}(Y_{it}|\boldsymbol{Y}_{t^*})$ and the conditional covariance $\text{Cov}(Y_{it}, Y_{i,t-1}|\boldsymbol{Y}_{t^*})$. Here, week is represented by $t = 1, \ldots, 20$ and county by $i = 1, \ldots, 140$. We have counts Y_{it} in county i at week t of the prediction

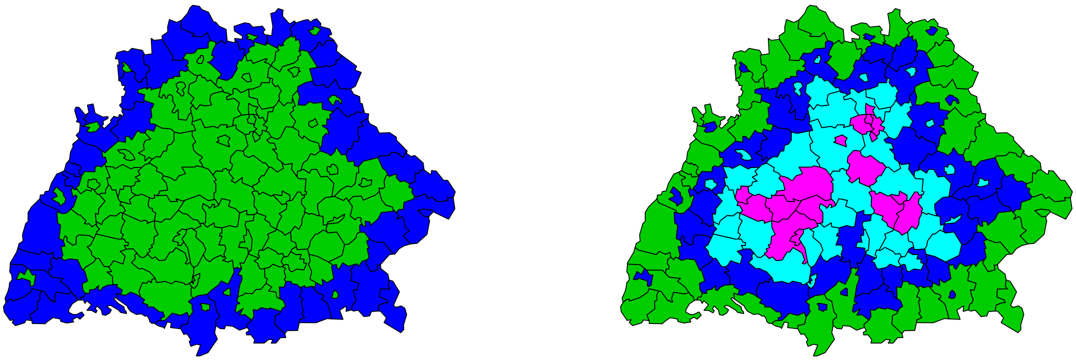

Figure 1: Visual confirmation of border covariates: binary indicator of whether a county is at the border or not (left) and the order the county is removed from the border (right).

and conditioning week $\boldsymbol{Y}_{t^*} = (Y_{1t^*}, Y_{2t^*}, \ldots, Y_{It^*})$, which contains the counts from all counties during the last week in the training data and is the period directly preceding the first week of the prediction. The means are iteratively calculated, meaning we first calculate $\mathbb{E}(Y_{it^*+1}|\boldsymbol{Y}_{t^*})$, then $\mathbb{E}(Y_{it^*+2}|\boldsymbol{Y}_{t^*})$, etc., using the law of total expectation and model-specific formulas. We obtain means

$$\begin{aligned}\mathbb{E}(Y_{it}|\boldsymbol{Y}_{t^*}) &= \mathbb{E}(\mathbb{E}(Y_{it}|\boldsymbol{Y}_{t-1})|\boldsymbol{Y}_{t^*}) \\ &= \nu_{it}e_{it} + \lambda_{it}\mathbb{E}(Y_{i,t-1}|\boldsymbol{Y}_{t^*}) + \phi_{it}\Sigma_{j\neq i}w_{ji}\mathbb{E}(Y_{j,t-1}|\boldsymbol{Y}_{t^*}).\end{aligned}$$

The second moments are obtained similarly, this time calculating $\mathbb{E}(Y_{it}^2|\boldsymbol{Y}_{t^*})$ and using the relationship $\mathrm{Var}(Y_{it}|\boldsymbol{Y}_{t^*}) = \mathbb{E}(Y_{it}^2|\boldsymbol{Y}_{t^*}) - \mathbb{E}(Y_{it}|\boldsymbol{Y}_{t^*})^2$. These calculations are also initialized by conditioning week $\boldsymbol{Y}_{t^*}$.

For the derivation of the first two moments of the multivariate path forecast, see Appendix A of Held et al. (2017).

2.4 Evaluation

Forecast evaluation here is guided by the idea that sharpness of a predictive distribution should be maximized subject to calibration (Gneiting et al., 2007). In other words, the predictive distribution should be statistically consistent with the observations (calibrated) and be sufficiently concentrated (sharp). To this end, proper scoring rules are employed at several aggregation levels: (i) raw space-time data, (ii) the data aggregated over region yielding weekly counts, (iii) the data aggregated over time yielding region-specific total counts, and (iv) data aggregated over time and space yielding final counts for the entire observation region and 20 weeks.

Scoring rules are used in evaluation to measure the quality of probabilistic forecasts and to rank competing forecast methods by assigning a numerical score based on the predictive distribution and the observations (Gneiting and Raftery, 2007). They are considered summary measures because they address both calibration and sharpness simultaneously. The score $S(\boldsymbol{P}, \boldsymbol{x})$ assigned to the predictive distribution $\boldsymbol{P}$ when we observe $\boldsymbol{x}$ can be considered a penalty of the statistical difference between observations and predictions which one wishes to minimize. Propriety of scoring rules encourages honest prediction and strict propriety ensures that both calibration and sharpness are addressed (Gneiting and Raftery, 2007; Winkler et al., 1996).

Propriety A forecaster's best judgement is the predictive distribution $\boldsymbol{P}$, as opposed to any other distribution $\boldsymbol{Q}$. Under propriety, there is no incentive to quote any distribution $\boldsymbol{Q}$ that is not $\boldsymbol{P}$ ($\boldsymbol{Q} \neq \boldsymbol{P}$) because $S(\boldsymbol{P}, \boldsymbol{P}) \leq S(\boldsymbol{Q}, \boldsymbol{P})$ for all $\boldsymbol{Q}$ and $\boldsymbol{P}$. If $S(\boldsymbol{P}, \boldsymbol{P}) = S(\boldsymbol{Q}, \boldsymbol{P})$, which occurs only when $\boldsymbol{Q} = \boldsymbol{P}$, then the score is strictly proper (Czado et al., 2009). In other words, a scoring rule is considered proper if its expected value for an observation from some distribution $\boldsymbol{G}$ is minimized when the predictive distribution $\boldsymbol{F}$ is equal to distribution $\boldsymbol{G}$ and strictly proper when this minimum is unique Gneiting et al. (2007).

Proper scoring rules compare predictive distributions, not just point predictions. As these methods are non-parametric and do not depend on nested models, they are widely applicable. See Gneiting et al. (2007) for further discussion of their advantages. Here we employ the multivariate Dawid-Sebastiani score (1999), which we denote as mDSS,

$$mDSS(\boldsymbol{P}, \boldsymbol{x}) = \log|\boldsymbol{\Sigma_P}| + (\boldsymbol{x} - \boldsymbol{\mu_P})^\top \boldsymbol{\Sigma_P}^{-1}(\boldsymbol{x} - \boldsymbol{\mu_P})$$

because it depends only on the mean vector $\boldsymbol{\mu_P}$ and covariance matrix $\boldsymbol{\Sigma_P}$ of the predictive distribution $\boldsymbol{P}$, making it easy to compute based on our long-term predictions. In this analysis, the standardized score is reported, reducing problems from unnecessarily large values:

$$mDSS(\boldsymbol{P}, \boldsymbol{x}) = [\log|\boldsymbol{\Sigma_P}| + (\boldsymbol{x} - \boldsymbol{\mu_P})^\top \boldsymbol{\Sigma_P}^{-1}(\boldsymbol{x} - \boldsymbol{\mu_P})]/(2d),$$

where d is the dimension of covariance matrix $\boldsymbol{\Sigma_P}$.

Like other scores, mDSS is negatively oriented, with lower scores indicating better predictive performance. As outlined in Held et al. (2017), p-values for forecast validity can be approximated based on the mDSS. When fitting our models we specify a negative binomial distribution. For the predictive distribution we know only the first two moments. Sometimes we approximate the predictive distribution using a negative binomial distribution, as in the fan plots, and other times with a normal distribution. In order to calculate p-values for forecast validity we assume the observations are approximately normal. Then, under the null hypothesis $H_0 : \boldsymbol{X} \sim \boldsymbol{P} = \mathcal{N}_d(\boldsymbol{\mu_P}, \boldsymbol{\Sigma_P})$, the mDSS follows a χ^2-distribution with d degrees of freedom but shifted by constant $\log|\boldsymbol{\Sigma_P}|$:

$$mDSS(\boldsymbol{X}, \boldsymbol{P}) - \log|\boldsymbol{\Sigma_P}| \sim \chi^2(d).$$

The mDSS is closely related to the determinant sharpness, denoted by DS,

$$DS = |\boldsymbol{\Sigma_P}|^{1/(2d)},$$

a negatively-oriented multivariate measure of sharpness suggested by Gneiting et al. (2008). Here we report

$$\mathrm{DS} = \log|\boldsymbol{\Sigma}_{\boldsymbol{P}}|/(2d)$$

to avoid unnecessarily large numbers, as in Held et al. (2017).

Calculation of these values is easily accomplished with the function `hhh4predict::ds_score_hhh4`, which provides scaled and unscaled mDSS and DS values when the option `detailed` is set to `true`. `R`-code 3 calculates mDSS, DS, and p-values for forecast validity for our list of flu predictions.

```r
DSSraw <-as.data.frame(do.call(rbind,
            lapply(flupreds, ds_score_hhh4, detailed=TRUE)))
mDSS.pval <- apply(DSSraw, 1, function(df) pchisq(df["term2"],
                 df=length(flupreds$endemic$mu), lower.tail=FALSE))
```

Code Example 3: Calculation of multivariate Dawid-Sebastiani score and determinant sharpness (using `hhh4predict::ds_score_hhh4`) and p-values for model validity.

To aid in interpretation, predictive performance is ranked by mDSS scores, with 1 indicating the best model.

2.5 Aggregated data versus "sliced" data

The scoring methods explained in Subsection 2.4 are applied to several different forms of the data. First, mDSS and DS are applied to the data at different aggregation levels, as in Meyer and Held (2014). The finest aggregation level available in this study is the space-time level or counts for each week in each county. When aggregating over only one dimension, one obtains either weekly (time) level data by summing over all regions per week or county (space) level data by summing over all weeks in each region. The data can be completely collapsed into the coarse final count aggregation level by summing over all weeks and all regions to yield a single number. The data is first aggregated and then all of the data available at that aggregation level is considered at once in calculating mDSS and DS for each model. In this way, we analyze whether models perform in the same manner across different aggregation levels. This is important because in practice predictions at only a single aggregation level may be of interest.

Second, to further investigate model performance with respect to time and space, mDSS and DS are also computed weekly and per region. Unlike in the procedure mentioned above in which data is first aggregated and then a single score calculated, here we subset the data by week or county, compute mDSS and DS for each week or county, and then average the weekly or county scores. This is a novel method and here we refer to the weekly/regional subsets as data "slices." This allows for a closer look at model performance in relation to weekly and regional characteristics.

For weekly scores, this requires subsetting $i \times i$ blocks from the diagonal of the $it \times it$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{P}}$ for each week to be scored. Each $i \times i$ block is the covariance matrix for only that point t in the prediction period. Each row of the $t \times i$ realization matrix and predicted mean matrix give the observed incidence and predictions for time t of the prediction period. `R`-code 4 computes the weekly mDSS scores. For regional scores, each column of the realization and predicted mean matrices gives the observed and predicted counts in a specific region i. The covariance matrix for each county is subset as shown in R-code 4.

```
# weekly score
 for (ind in weeks){
  w <- ((140*(ind-1))+1):(140*ind)
  term1[ind] <- determinant(pred$Sigma[w, w] ,
              logarithm = TRUE)$modulus[1]
  term2[ind] <- t(pred$realizations_matrix[ind,] - pred$mu_matrix[ind,])
   %*% solve( pred$Sigma[w, w]) %*%
   (pred$realizations_matrix[ind,] - pred$mu_matrix[ind,])
  det_sharpness[ind] <- term1[ind] / (2 * length(pred$mu_matrix[ind,]))
  scaled_DSS[ind] <- (term1[ind] + term2[ind]) /
    (2 * length(pred$mu_matrix[ind,]))
 }
# regional score
 for (z in 1:140){
  inds <- z+0:19*140
  term1[z] <- determinant(pred$Sigma[inds, inds], log = TRUE)$modulus[1]
  term2[z] <- t(pred$realizations_matrix[,z]-pred$mu_matrix[,z]) %*%
   solve( pred$Sigma[inds, inds]) %*%
   (pred$realizations_matrix[,z]-pred$mu_matrix[,z])
  det_sharpness[z] <- term1[z] / (2*length(pred$mu_matrix[,z]))
  scaled_DSS[z] <- (term1[z]+term2[z]) / (2*length(pred$mu_matrix[,z]))
 }
```

Code Example 4: This code shows snippets of functions used to extract relevant parts of the mean and covariance matrices for computing weekly and regional mDSS.

3 Exploratory data analysis

3.1 Spatio-temporal characteristics of influenza in Baden-Württemberg and Bayern

The plotted time series for 2001 through 2017 (Figure 2) reveals a general seasonality to the data, with incidence peaking toward the end of the first quarter/beginning of the second quarter, i.e., in winter, as expected. However, the counts each year vary in several ways: final size, peak size, and duration (Table 1). The data for 2009 and 2010 (in red) represent the extreme nature of the H1N1 "swine flu" pandemic. Incidence peaks much higher, at 8772 counts, than at any other time in the data set and there is only one week of 2009 with zero cases. The peak week of 2010 is listed as -8 because the 2010 epidemic season is really the end of the second epidemic season of 2009, which peaked 8 weeks before the start of 2010. The season persisted for 13 weeks starting from the first week of January 2010. However, counting from the beginning of consecutive weeks of elevated incidence (mid 2009), this epidemic season lasts for 48 consecutive weeks - almost an entire year. This extreme nature makes 2009-2010 unsuitable for this analysis; it is transformed (see Subsection 2.1) for use in model fitting but not evaluated in Section 5.

Peak height ranges from 124 counts in 2001 and 2002 to 5036 counts in 2017. The epidemic season peaks between weeks 5 and 13, with a median peak week of 9. Final size increases dramatically over the observation period, with just 642 counts in 2001, but 27,143 counts in 2017. The higher counts in the later years are suspected to come from improvements in reporting practices since the IfSG was enacted in 2000. Consecutive weeks of elevated counts (> 0), both counting from the beginning of the calendar year and from the beginning of elevated counts at the end of the previous year, have also increased. This also indicates that reporting has improved, especially outside of the epidemic season.

	Peak height	Peak week	Final size	Consecutive weeks (Jan.)	Consecutive weeks (season)
2001	124	7	642	15	15
2002	124	11	745	17	17
2003	517	9	2559	18	18
2004	127	7	937	18	27
2005	645	9	3797	16	20
2006	269	13	1303	22	22
2007	1181	9	6315	18	26
2008	791	5	6286	21	27
2009	8772	6	61596	16	29
2010	507	-8	1250	13	48
2011	1881	6	11570	21	33
2012	1042	11	6431	24	33
2013	2374	8	18440	28	40
2014	370	11	3892	31	45
2015	4152	8	26078	25	36
2016	1587	11	12880	23	45
2017	5036	6	27143	21	41

Table 1: Characteristics of the epidemic season for each year from 2001 through 2017.

From the map of the final incidence counts in Figure 3 one can see that the largest number of cases tend to be found in and near the two largest cities, Munich and Stuttgart. Passau (LK) also reports large numbers of cases, surpassing even the major city it envelops. However, a breakdown of this information by year is more informative (see Figure 36 and Figure 37 in appendix). Some regions consistently report zero (SK Memmingen) or few counts while the region with the highest yearly incidence is fairly stable throughout the observation period. Several counties report zero counts at the start of the observation period, but by year 2008, almost all counties are reporting some incidence every year.

Heterogeneity of counts exists in both time and space. According to the 'Protection Against Infection' statute (IfSG, BGBl. I S. 1045), only verified influenza cases must be reported, leaving considerable room for underreporting at different rates across regions.

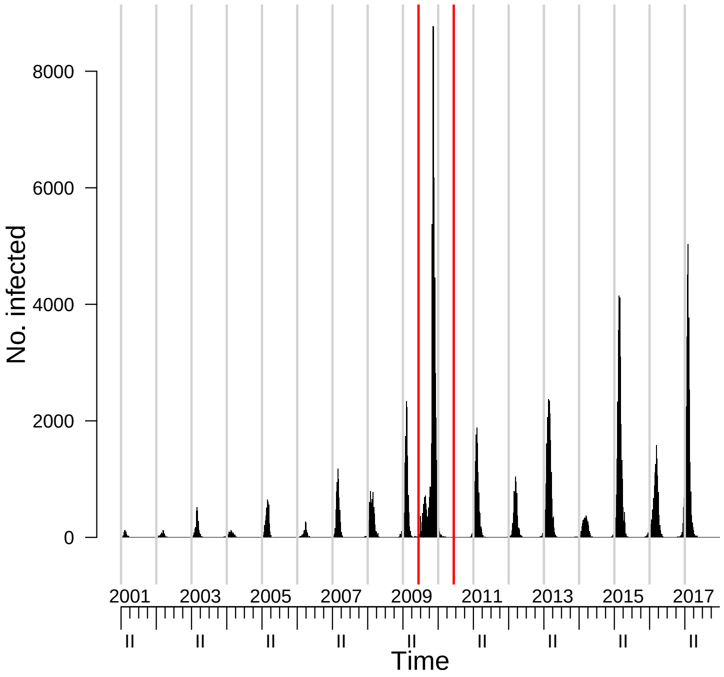

Figure 2: Time series of influenza counts in Bayern and Baden-Württemberg from 2001 to 2017. The counts for the period between the red lines are a result of the 2009 pandemic and are not considered in model evaluation.

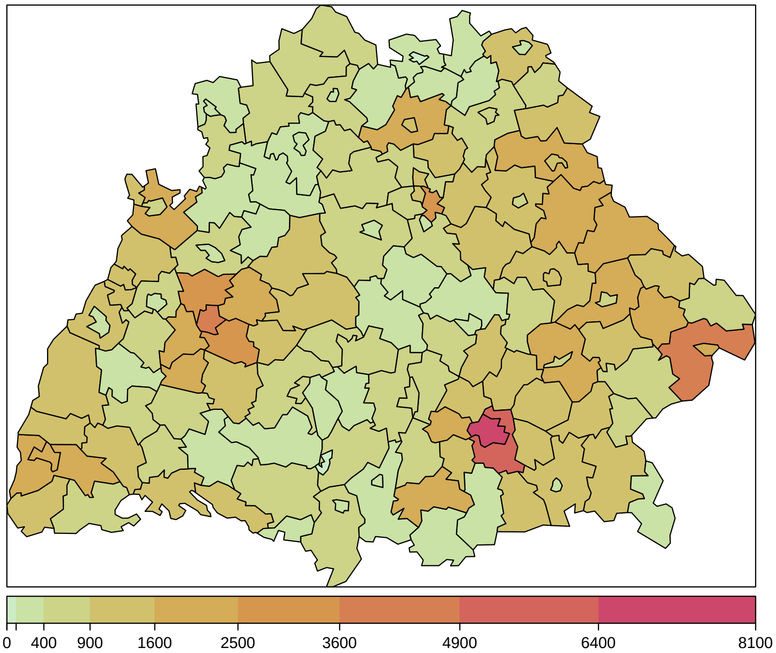

Figure 3: Total incidence per county in Bayern and Baden-Württemberg from 2001 to 2017 using the most recent data. The color scale, from green to red, represents increasing total incidence.

Differences between `fluBYBW` and newly-retrieved data

The first exercise in this study (Section 4) is to make predictions for the first 20 weeks of 2008. This enables comparison with the results of Meyer and Held (2014). Since the data was last retrieved, RKI has updated the counts for the 2001-2008 period. The results in Table 1 are for the updated counts. According to the original `fluBYBW` data the peak size ranges from 111 in 2002 to 1158 in 2007 and the final count for each year ranges from 612 in 2001 to 6136 in 2007. Consecutive weeks with elevated counts range from 15 to 22 per year, or 15 to 27 per epidemic season.

For the 2008 prediction period, the largest weekly change in counts was 60, occurring in week 5, corresponding to the time of the unexpected early first peak. The average change was an increase of 8.75 counts and no decreases were observed by week (see Figure 3.1). In Figure 5, the change in counts is only discernible in 2 regions: Pfaffenhofen a.d.Ilm and Dillingen a.d.Donau. These regions saw increases of 28 and 53 counts, respectively. The mean change per region, however, is only an increase of 1.25 counts.

3.2 Comparison of gravity measures

In addition to population fraction, we use population density and urbanicity as further "gravity" measures. The log of all of these measures are used for comparison. Figure 6 displays the distributions of population fraction, population density, and urbanicity on the diagonal, the correlations between them in the upper triangle and scatter plots in the lower triangle. For both density and urbanicity, a comparison with population fraction results in two separate clusters for *Stadtkreise* and *Landkreise*, where *Stadtkreise* are counties composed of one large city and *Landkreise* are counties made up of several subunits, such as towns. The two clusters exist in the scatter plot between density and urbanicity but are part of a single continuous pattern, suggesting that density and urbanicity measure the urban nature of a region more similarly than either compared to population fraction. *Stadtkreise* are clearly more dense and more urban than *Landkreise* given any population fraction, suggesting that population fraction alone may not sufficiently capture the idea of urbanicity. When comparing population fraction and urbanicity, the *Stadtkreise* form a straight line because *Stadtkreise* do not have subunits; for *Stadtkreise* population fraction and urbanicity are equivalent. These units contribute their own populations toward the measure of urbanicity, while urbanicity for *Landkreise* contain only partial population information for a given unit. Density and urbanicity correlate highly (r=0.86) with each other, but less so with population (r=0.12 and r=0.38).

Table 2 provides a comparison of the highest and lowest ten ranking counties by population fraction, population density, and urbanicity. All three measures rank Munich and Stuttgart as the top counties. The top four rankings by density and urbanicity are identical, as well as 4 of the 5 lowest ranking counties for these measures. Population fraction is unlike the other measures at the low end of the rankings because small *Stadtkreise* account for the lowest 17 ranks (lowest 12.1%). By density and urbanicity, however, these same regions have ranks between eight and fifty-seven. As observed in the scatter plots, density and urbanicity are more related to each other than to population fraction.

In our analysis, regions most likely to draw commuters to them should be considered the most urban. Using population fraction alone may allow regions with high total populations but low relative density and low largest populations to be more attractive in the

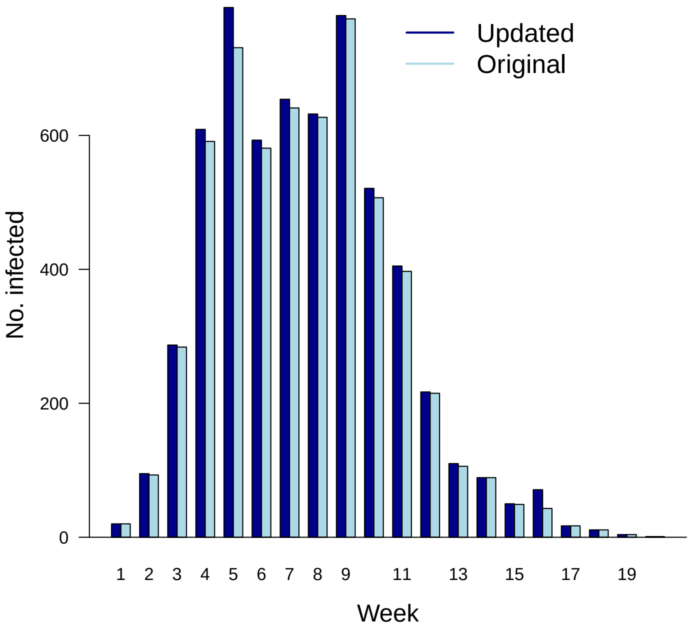

Figure 4: Comparison of original (light blue) and updated (dark blue) 2008 counts per week during the epidemic season.

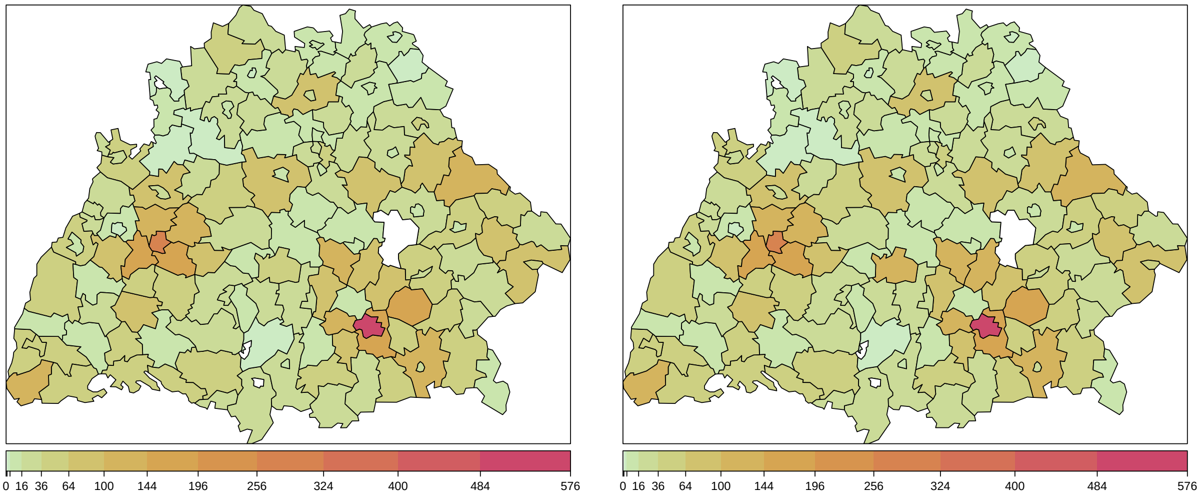

Figure 5: Maps of 2008 influenza season counts based on the original (left) and updated (right) data. The redder the color, the higher the incidence. White regions reported no cases. The most noticeable changes occur in counties to the southwest of the largest white county.

models than they deserve to be.

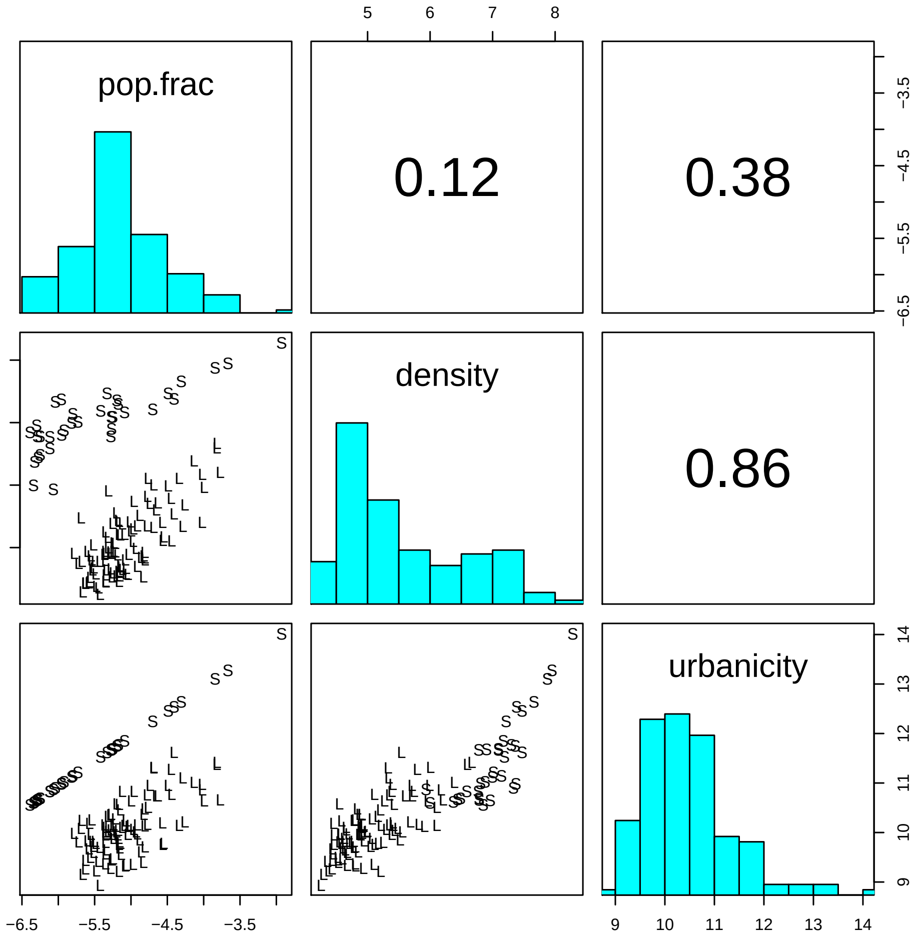

Figure 6: Scatter plots, distributions, and correlations for the three gravity measures on the log scale: population fraction, population density, and urbanicity. Each ‘S’ represents a *Stadtkreise* and ‘L’ a *Landkreise*.

Population rank	Density rank	Urbanicity rank
SK München	SK München	SK München
SK Stuttgart	SK Stuttgart	SK Stuttgart
LK Rhein-Neckar-Kreis	SK Nürnberg	SK Nürnberg
LK Esslingen	SK Mannheim	SK Mannheim
LK Ludwigsburg	SK Augsburg	SK Karlsruhe
SK Nürnberg	SK Fürth	SK Augsburg
LK Karlsruhe	SK Karlsruhe	SK Freiburg i.Breisgau
LK Rems-Murr-Kreis	SK Rosenheim	SK Heidelberg
LK Ortenaukreis	SK Regensburg	SK Würzburg
LK Böblingen	SK Schweinfurt	SK Regensburg
SK Passau	LK Amberg-Sulzbach	LK Hof
SK Hof	LK Garmisch-Partenkirchen	LK Würzburg
SK Straubing	LK Bayreuth	LK Landshut
SK Amberg	LK Rhön-Grabfeld	LK Bamberg
SK Weiden i.d.OPf.	LK Regen	LK Schweinfurt
SK Coburg	LK Freyung-Grafenau	LK Freyung-Grafenau
SK Kaufbeuren	LK Straubing-Bogen	LK Straubing-Bogen
SK Memmingen	LK Neustadt/Aisch-Bad Windsheim	LK Miltenberg
SK Ansbach	LK Tirschenreuth	LK Tirschenreuth
SK Schwabach	LK Neustadt a.d.Waldnaab	LK Neustadt a.d.Waldnaab

Table 2: Top ten and lowest ten ranking counties by population fraction, population density, and urbanicity.

4 Influenza surveillance data (fluBYBW) from 2001-2008

4.1 Applying analytical method for long-term forecasts to fluBYBW

Figure 8 shows the predictive means of the final size (in red) of the 2008 epidemic season for each model with the first through ninety-ninth predictive quantiles represented by the green-gray scale. The observed final size depicted by the horizontal dashed line is 5781. As in Meyer and Held (2014), power-law configurations overestimate the final size while the first order and autoregressive models underestimate it. Similar to the results of Held et al. (2017) with regards to norovirus, the endemic model is the only model which does not include the actual final size in its 95% prediction interval, meaning the observed final count is not supported by the prediction. Prediction interval lengths increase in size with the inclusion of epidemic potential and to a lesser extent with the number of parameters in the model. The endemic model is simply a negative binomial regression model for independent observations and so it has the narrowest interval. The power law configurations have relatively large, but similarly sized 95% prediction intervals, but large prediction intervals are common for long-term predictions. The autoregressive model and power law model without gravity components come closest to the observed count, underestimating by 290 and overestimating by 294 respectively.

Figure 7, the time series of observed and predicted weekly counts, more clearly depicts differences between the models. Predictions are similar across models at the season start and end, when counts are low. The biggest differences are observed when counts are highest. As in Figure 8, the endemic model overestimates the most and the autoregressive and first order models have the lowest predicted counts. Power law formulations are similar to each other and lie between the simpler models, as observed in Meyer and Held (2014). The observed counts (black) peak twice, once at week 5 and again at week 9, however, none of the models capture this. All models peak only once, and not until week 9 or 10. Failing to predict the early start of the 2008 epidemic season may be a consequence of later starts in the epidemic seasons of the training data and to the inflexibility in the seasonality added to the model. Occurrence of one peak instead of two is a result of training on data with only single peaks. Such a large spread between peaks is only observed again in the 2009 pandemic.

The models can be visually evaluated further with the fanplots in Figure 40 and the prediction maps in Figure 41, both found in the appendix. The endemic model predictions plotted in Figure 40 are least like the observed counts, reaffirming the trend seen in Figure 8. While this trend is difficult to see in Figure 41, one can see that all models are predicting the highest counts in the two largest cities, Munich and Stuttgart, as observed, but are less consistent with the observations when no (as in Kelheim, Memmingen, and Kempten) or low counts are recorded. Surprisingly high counts are predicted by all models for Schwäbisch Hall and Neustadt an der Waldnaab. Both cities have quite different ranks according to the three different urban measures. Schwäbisch Hall, for example, ranks 35th by population fraction, 59th by urbanicity, but 99th by density.

To summarize and rank the predictive performance of each model, we use the multivariate Dawid-Sebastiani score (mDSS) and determinant sharpness (DS). In Table 3 these scores are given, along with p-values quantifying evidence against the null hypothesis, H_0 : model calibrated. These scores are computed at various levels of data aggregation: (1) space-time - the data is unaggregated with counts for each region at each week; (2) time - the weekly counts in each region are summed up for total weekly counts in the entire observation region; (3) space - total counts in each region for the entire time period are obtained by summing the weekly counts; (4) final counts - the data is aggregated over week and region for a single total count. Each model is ranked according to its mDSS, with lowest scores ranked best.

As found in Meyer and Held (2014) (Table 5), the power law models perform better than the baseline models at the raw, time, and space aggregation levels. In contrast to these findings, however, the autoregressive and first order models are best-ranked at the final count level. There is evidence, though, that all models are miscalibrated at the space-time and time levels (p-values < 0.001). The DS indicates the power law models are also sharper than the baseline models at all levels except the final count level, even though the fan plots in Figure 40 are wider for these models. This is because determinant sharpness also reflects the strength of the implied autocorrelations. In Figure 10 we see the power models have the highest autocorrelations for almost the entire prediction period. Held et al. (2017) also find a power law model with a gravity component to be sharper than baseline models for the epidemic curve because of high autocorrelations. Another similarity is that models with the largest DS (first order model shown) have the smallest correlations between regions (Figure 11).

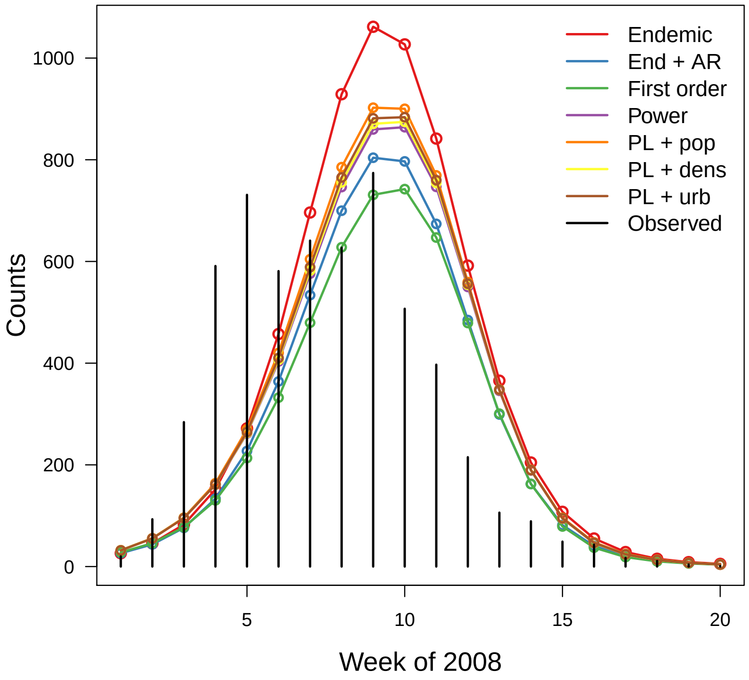

Figure 7: Time series of observed and predictive mean counts.

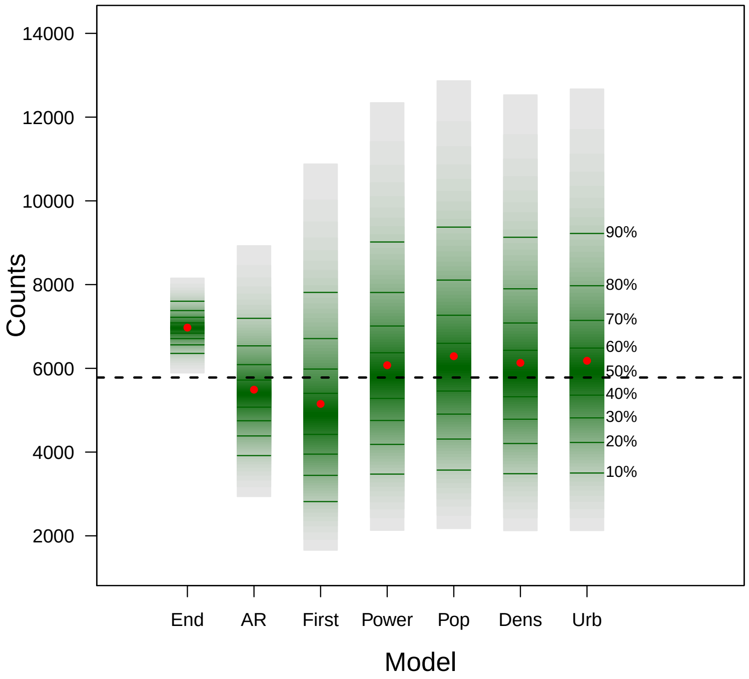

Figure 8: Final size predictions with 95% prediction intervals for the first 20 weeks of 2008. The dashed vertical line is at the actual final count size, 5781.

Of note are the large p-values (1.00 and 0.74) for the autoregressive and first order models at the regional count level. If we restrict our overdispersion parameter to $\psi = 0$, i.e., a poisson model instead of a negative binomial, the mDSS increases (3.9780 and 4.6027), but the DS decreases (2.5625 and 2.3070), i.e., the new models do not perform as well but they are sharper. Both p-values have shrunk to < 0.0001. This indicates that the problem has to do with large estimated overdispersion leading to large variance, as displayed in Figure 9. For large enough variance estimates, the models cannot be considered miscalibrated.

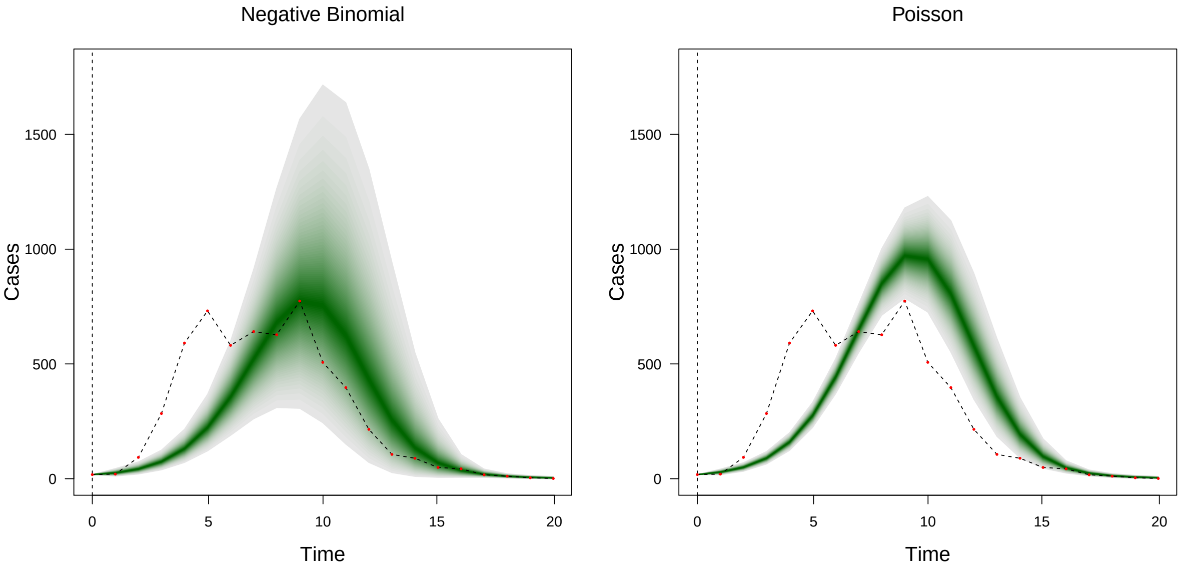

Figure 9: Fanplots of long-term predictions of the 2008 epidemic season using the original negative binomial autoregressive model (left) and a poisson autoregressive model (right).

At the final count level, the endemic model is sharpest, but there is much more evidence against calibration of the endemic model than against the other models. Models should be sharp, but subject to first being calibrated. Even at the final count level, there is still some evidence of miscalibration of the endemic model (p-value 0.014). This is confirmed by Figure 8, in which the endemic model is the only model whose 95% prediction interval does not include the observed final size. At the space level, the population model performs best by mDSS and is the only complex model for which there is no evidence against calibration. In general, the scores of the power law models are all close and there is no clear best across all aggregation levels. This is expected because when plotting the estimated weights against neighborhood order, one observes similar weights for all power models.

Table 4 displays the univariate Dawid-Sebastiani score (uDSS) for these models and is most comparable to Meyer and Held's (2014) Table 5. Both use the univariate Dawid-Sebastiani score, however Meyer and Held apply it to simulated predictions and here it is applied to predictive distributions characterized by the analytical mean and covariance. Ranks of models common to both tables are the same, except for the power models at the

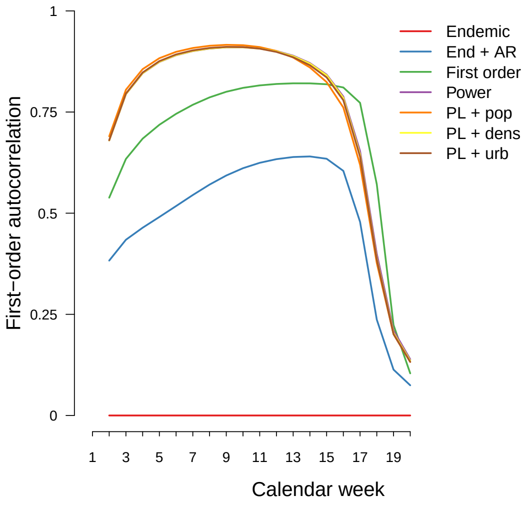

Figure 10: First order autocorrelations between weeks.

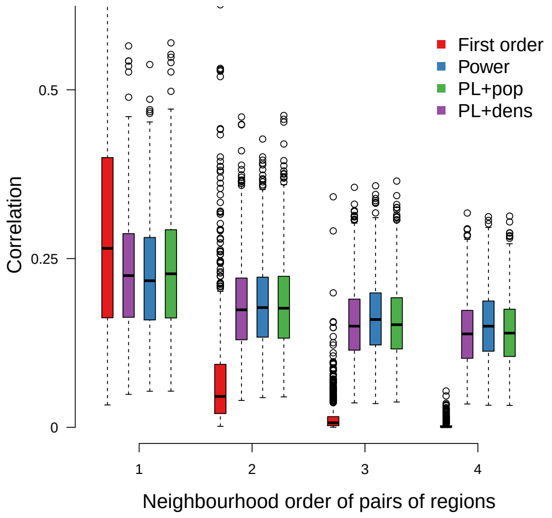

Figure 11: Correlations between regions implied by several models.

space-time level. The mDSS detected no difference between the simulated power model and power model with population (both 2.29), but the analytical results show the power model to be slightly better (2.24 and 2.27). The results are also similar to the multivariate results: power law formulations consistently perform better than the simpler models until final count level (not reported in (Meyer and Held, 2014)). When ranking the power law models only, the results are the same for the multivariate and univariate scores, with the power law or population models always best.

The updated counts do not alter the evaluation of the 2008 predictions much (Table 5). Even with the updated counts, power models are still performing better than the baseline models at all data aggregation levels except for final counts. The endemic+autoregressive model still ranks best at the final count level, but now the power model outperforms the first order model. There is a large amount of evidence against calibration of all models at the space-time and time levels, as before, but now there is more evidence against calibration at the space level, as well. The p-values for the power, density, and urbanicity models were 0.019, 0.017, and 0.016, but with the updated data are 0.0016, 0.0018, and 0.0023. The sharpness ranks are unchanged for the top 2 models at any aggregation level.

4.2 Extending Meyer and Held (2014)'s `fluBYBW` analysis

Weekly and regional "slices"

The mean weekly mDSS and mean regional mDSS (Table 6) indicate that the power law models perform better on average across the twenty weeks of prediction and across the 140 regions. The mean scores separate models with and without the power law, with power law configurations scoring better. Amongst the better scoring models, the power

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	1.4294	0.3116	<0.0001	7	13.4979	3.5427	<0.0001	7
End + AR	1.0945	0.4427	<0.0001	6	10.3229	3.5789	<0.0001	6
First order	1.0721	0.2057	<0.0001	5	9.9263	3.5724	<0.0001	5
Power	0.9570	0.1939	<0.0001	1	8.3223	3.3702	<0.0001	4
PL + pop	0.9734	0.2056	<0.0001	4	8.2217	3.3837	<0.0001	1
PL + dens	0.9640	0.1942	<0.0001	2	8.2572	3.3800	<0.0001	3
PL + urb	0.9705	0.1962	<0.0001	3	8.2417	3.3891	<0.0001	2
	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	3.9218	2.7416	<0.0001	7	9.1885	6.1874	0.014	7
End + AR	3.9046	3.7952	1.00	6	7.1878	7.1625	0.82	1
First order	3.6682	3.2072	0.74	5	7.6475	7.5976	0.75	2
Power	3.6469	3.0149	0.019	2	7.7079	7.6991	0.89	3
PL + pop	3.6221	3.0299	0.067	1	7.7694	7.7451	0.83	6
PL + dens	3.6536	3.0191	0.017	3	7.7301	7.7179	0.88	4
PL + urb	3.6574	3.0213	0.016	4	7.7464	7.7310	0.86	5

Table 3: Multivariate Dawid-Sebastiani scores, determinant sharpness, p-values, and ranks for the models at various aggregation levels (`fluBYBW` data).

	Space-time		Time		Space		Final	
	mDSS	rank	mDSS	rank	mDSS	rank	mDSS	rank
Endemic	2.8589	7	26.9959	5	7.8437	7	18.3770	7
End + AR	2.5014	6	30.6564	7	7.8092	6	14.3757	1
First order	2.4727	5	27.6817	6	7.4771	5	15.2950	2
Power	2.2431	1	16.0305	4	7.2923	2	15.4159	3
PL + pop	2.2743	4	15.2788	1	7.2506	1	15.5389	6
PL + dens	2.2569	2	15.8731	3	7.3023	3	15.4601	4
PL + urb	2.2712	3	15.7301	2	7.3075	4	15.4929	5

Table 4: Univariate Dawid-Sebastiani scores and ranks for the models at various aggregation levels (`fluBYBW` data).

model without a gravity component scores best. The weekly and regional endemic model scores are exactly the same; this is expected because the covariance matrix for this model is diagonal. The weekly and regional rankings align with the space-time rankings better than with the time or space rankings of Table 3. Rather than scores on data aggregated over time or space, the weekly and regional scores are scores on smaller subsets of the predictions that are averaged together. The univariate scores also indicate that the power law model performs best, followed by the power law models with gravity components. The univariate weekly and regional scores are the same; they are averages of the same data just taken in a different order.

The weekly scores are expected to be relatively low when counts are low as, for ex-

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	1.7177	0.3405	<0.0001	7	14.0064	3.5620	<0.0001	7
End + AR	1.4570	0.4753	<0.0001	5	11.1083	3.6253	<0.0001	6
First order	1.5778	0.2358	<0.0001	6	10.3459	3.6975	<0.0001	5
Power	1.3175	0.2266	<0.0001	1	8.8283	3.4717	<0.0001	4
PL + pop	1.3307	0.2388	<0.0001	4	8.7032	3.4882	<0.0001	1
PL + dens	1.3259	0.2278	<0.0001	2	8.7563	3.4758	<0.0001	3
PL + urb	1.3290	0.2305	<0.0001	3	8.7279	3.4849	<0.0001	2

	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	4.0443	2.7789	<0.0001	7	8.8770	6.2107	0.021	7
End + AR	3.9503	3.8273	1.00	6	7.2427	7.1965	0.76	1
First order	3.7902	3.2432	0.21	5	7.7785	7.7444	0.79	3
Power	3.7482	3.0531	0.0016	2	7.7758	7.7693	0.91	2
PL + pop	3.7063	3.0700	0.016	1	7.8319	7.8121	0.84	6
PL + dens	3.7510	3.0586	0.0018	4	7.7939	7.7847	0.89	4
PL + urb	3.7489	3.0626	0.0023	3	7.8058	7.7939	0.88	5

Table 5: Evaluation of 2008 (updated count data) predictions including multivariate Dawid-Sebastiani score, determinant sharpness, p-value, and rank for models at various aggregation levels.

ample, at week 1 and again at the end of the epidemic season. In Figure 12 low scores are seen at week 1, with no large differences between models. Differences can already be seen at week 2, with trajectories already starting to separate. All model scores peak at week 4, due to the fact that all models failed to predict the early first peak of the 2008 epidemic season. The models did predict the second half of the epidemic, and the scores all decrease at this time. After level scores for several weeks, a decrease can be seen as the epidemic dies out. The local peak at week 18 is partially explained by an abnormally high observed count in Freising at this time. The scores for the baseline models are higher (i.e., worse predictions) than the scores for the power law formulations at the week 4 and 18 peaks and in general over the whole prediction period, except at weeks 5 and 6, when the autoregressive model scores best. The differences between the power law models are only vaguely perceptible over the entire prediction period. Figure 13, the plot of the univariate scores, shows a very similar pattern except on a slightly larger scale.

The regional scores are expected to be relatively low for less urban regions, where counts are usually low, and higher for urban regions, which may have much greater variation in counts. In Table 7 the scores for the ten most and least dense regions are given for each model. The two most dense regions, Munich and Stuttgart, do have noticeably larger scores than the other regions, but surprisingly large scores are also found for several of the most rural regions, such as Freyung-Grafenau and Neustadt a.d. Waldnaab. The counts of these rural regions with high scores do not peak until after week 9, when the total weekly counts start dropping. The power models score better than the simpler models more consistently for the more rural regions, whereas results for the most urban

	Multivariate mean scores						
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Week	1.4294	1.2507	1.1747	1.0400	1.0549	1.0467	1.0531
Region	1.4294	1.0945	1.0729	0.9703	0.9870	0.9770	0.9835
	Univariate mean scores						
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Week	2.8589	2.5014	2.4727	2.2431	2.2743	2.2569	2.2712
Region	2.8589	2.5014	2.4727	2.2431	2.2743	2.2569	2.2712

Table 6: Mean weekly and regional multivariate Dawid-Sebastiani scores.

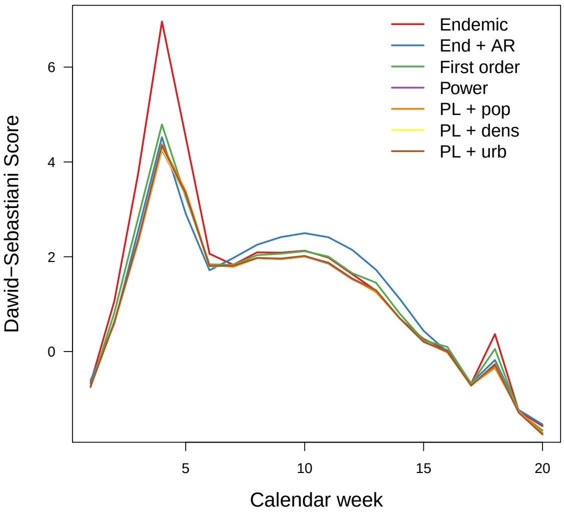

Figure 12: Weekly multivariate Dawid-Sebastiani scores for each model.

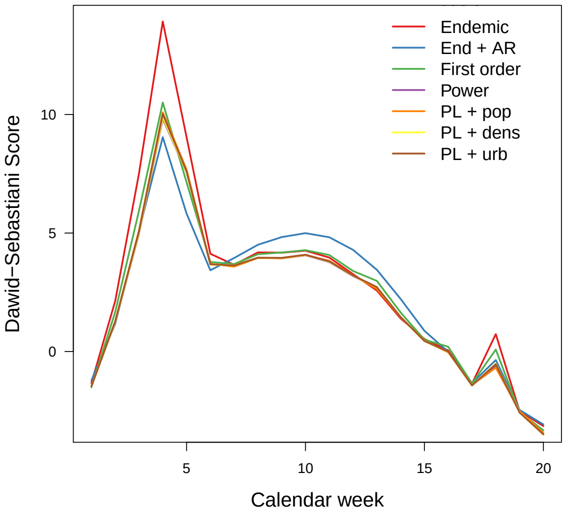

Figure 13: Weekly univariate Dawid-Sebastiani scores for each model.

regions are more mixed. In urban counties most counts are expected to come from within and therefore the autoregressive component would be more important to these areas. For small communities, incidence in neighbors is expected to be more important and so the neighborhood component would be more important.

Border effects

For each of the 3 baseline and 4 power models, distance to the border was included as a covariate as either a discrete index, the log of the discrete index, the inverse of the discrete index, or a binary indicator. The discrete index quantifies the distance of a county to the border in terms of the least number of districts that must be crossed to reach the border, similar to the order of neighborhood. The covariate was added to either the endemic component or the neighborhood component (when it exists).

	Greatest population density						
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
SK München	3.5779	3.5206	3.4968	3.3287	3.3015	3.2978	3.2722
SK Stuttgart	2.6315	2.4769	2.2677	2.2078	2.2211	2.2070	2.2049
SK Nürnberg	1.8924	2.0444	1.7825	1.7325	1.7370	1.7337	1.7311
SK Mannheim	1.5010	0.8277	0.8742	1.0243	1.0189	1.0199	1.0156
SK Augsburg	1.3413	1.1343	0.9006	0.9576	0.9706	0.9628	0.9654
SK Fürth	0.2830	0.3302	0.3240	0.2857	0.2864	0.2576	0.2497
SK Karlsruhe	1.0480	0.9448	0.6932	0.6497	0.6861	0.6536	0.6584
SK Rosenheim	0.6580	0.4895	0.8845	0.6429	0.7134	0.5633	0.5777
SK Regensburg	0.5803	0.5627	0.4650	0.3222	0.3107	0.3289	0.3286
SK Schweinfurt	0.3932	0.2635	0.0890	0.0146	0.0184	0.0216	0.0220
	Smallest population density						
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
LK Amberg-Sulzbach	0.5345	0.6247	0.6696	0.5802	0.5656	0.5908	0.5928
LK Garmisch	0.1964	0.4131	0.2415	0.1253	0.1487	0.1292	0.1307
LK Bayreuth	0.2876	0.1046	0.1334	0.1725	0.2052	0.1939	0.2047
LK Rhön-Grabfeld	0.4736	0.4861	0.2202	0.2128	0.2067	0.2068	0.2030
LK Regen	1.2110	1.2151	1.2066	0.8667	0.8759	0.8748	0.8794
LK Freyung-Grafenau	3.3338	1.3671	1.8010	1.6578	1.6612	1.6932	1.6638
SK Straubing	0.4496	0.3106	0.2967	0.0955	0.1739	0.0798	0.0836
LK Neustadt	0.7726	0.9296	0.5993	0.5016	0.5070	0.5032	0.5043
LK Tirschenreuth	0.3275	0.3045	0.7229	0.2562	0.1924	0.2631	0.2555
LK Neustadt a.d.W.	2.1019	2.2296	2.1306	1.9396	1.9593	1.9861	2.0221

Table 7: Regional multivariate Dawid-Sebastiani scores for the most and least dense regions.

The results for baseline (Table 8) and power models (Table 9) are quite different. For baseline models, the models without any border effect covariate perform best in terms of mDSS in 7 rankings out of 12 and perform second best in 2 more rankings. However, the endemic and autoregressive models aggregated for weekly counts perform better with the covariate. Again, testing the calibration of the autoregressive model and to a lesser extent the first order model results in p-values which are unusually large due to extremely large values in the covariance matrix. The first order model without a border covariate ranks best, however, when such a covariate is added, it tends to perform better in the neighborhood compartment. No such statements can be made for the endemic and autoregressive models because neither contain a neighborhood component.

By contrast, when ranking power models with a border covariate against those without, the model without border effects only ranks best in 1 of the set of 16 rankings, specifically for the population model at the regional count level. For the density and urbanicity power models, models with a border covariate in the neighborhood component perform better than those with the covariate in the endemic component in almost all cases at all aggregation levels. Results for the power and population models are mixed. The top power model always includes the border effect covariate in the neighborhood component, but the rest of the neighborhood component border covariate power models do not always beat the models with the covariate in the endemic component. As opposed to all other models, results for the population model suggest better performance with the covariate in the endemic component, at least at the raw and final count levels.

	Space-time				Time				Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic model																
no BE	1.4294	0.3116	<0.0001	1	13.4979	3.5427	<0.0001	5	3.9218	2.7416	<0.0001	1	9.1885	6.1874	0.014	1
log Ind	1.4298	0.3115	<0.0001	3	13.4864	3.5434	<0.0001	2	3.9243	2.7415	<0.0001	4	9.1967	6.1883	0.014	4
Ind	1.4297	0.3115	<0.0001	2	13.4828	3.5436	<0.0001	1	3.9252	2.7415	<0.0001	5	9.2000	6.1885	0.014	5
Inv Ind	1.4298	0.3115	<0.0001	4	13.4892	3.5432	<0.0001	3	3.9236	2.7415	<0.0001	3	9.1947	6.1881	0.014	3
Bin Ind	1.4298	0.3115	<0.0001	5	13.4925	3.5430	<0.0001	4	3.9227	2.7415	<0.0001	2	9.1935	6.1879	0.014	2
Endemic + autoregressive model																
no BE	1.0945	0.4427	<0.0001	1	10.3229	3.5789	<0.0001	4	3.9046	3.7952	1.00	5	7.1878	7.1625	0.82	2
log Ind	1.0964	0.4388	<0.0001	4	10.3143	3.5840	<0.0001	2	3.8984	3.7872	1.00	2	7.1923	7.1683	0.83	4
Ind	1.0964	0.4375	<0.0001	5	10.3085	3.5873	<0.0001	1	3.8964	3.7845	1.00	1	7.1960	7.1728	0.83	5
Inv Ind	1.0963	0.4400	<0.0001	3	10.3186	3.5813	<0.0001	3	3.9003	3.7897	1.00	3	7.1894	7.1648	0.82	3
Bin Ind	1.0960	0.4416	<0.0001	2	10.3237	3.5780	<0.0001	5	3.9028	3.7930	1.00	4	7.1860	7.1607	0.82	1
First order model																
no BE	1.0721	0.2057	<0.0001	2	9.9263	3.5724	<0.0001	1	3.6682	3.2072	0.74	1	7.6475	7.5976	0.75	1
log Ind	1.0771	0.2061	<0.0001	8	9.9354	3.5858	<0.0001	6	3.6757	3.2080	0.70	8	7.6659	7.6203	0.76	7
Ind	1.0777	0.2061	<0.0001	9	9.9283	3.5855	<0.0001	4	3.6770	3.2081	0.69	9	7.6661	7.6208	0.76	8
Inv Ind	1.0764	0.2060	<0.0001	7	9.9400	3.5851	<0.0001	8	3.6743	3.2079	0.70	7	7.6646	7.6185	0.76	6
Bin Ind	1.0752	0.2059	<0.0001	6	9.9439	3.5829	<0.0001	9	3.6718	3.2076	0.72	6	7.6611	7.6140	0.76	5
NE log Ind	1.0723	0.2062	<0.0001	4	9.9275	3.5784	<0.0001	2	3.6693	3.2080	0.73	3	7.6544	7.6062	0.76	3
NE Ind	1.0723	0.2057	<0.0001	5	9.9277	3.5756	<0.0001	3	3.6690	3.2069	0.73	2	7.6506	7.6014	0.75	2
NE Inv Ind	1.0722	0.2068	<0.0001	3	9.9297	3.5829	<0.0001	5	3.6699	3.2092	0.74	4	7.6601	7.6131	0.76	4
NE Bin Ind	1.0720	0.2077	<0.0001	1	9.9395	3.5935	<0.0001	7	3.6713	3.2111	0.74	5	7.6732	7.6287	0.77	9

Table 8: Multivariate Dawid-Sebastiani scores for baseline model predictions (endemic, autoregressive, and first order) accounting for effects of proximity to border in different compartments for the 2008 epidemic season at various levels of aggregation. “NE” denotes the covariate is in the neighborhood component, otherwise the covariate is in endemic component.

	Space-time				Time				Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Power law model																
no BE	0.9570	0.1939	<0.0001	6	8.3223	3.3702	<0.0001	6	3.6469	3.0149	0.019	4	7.7079	7.6991	0.89	7
log Ind	0.9573	0.1928	<0.0001	7	8.3220	3.3700	<0.0001	5	3.6501	3.0130	0.016	8	7.7055	7.6972	0.90	2
Ind	0.9578	0.1925	<0.0001	9	8.3185	3.3698	<0.0001	2	3.6514	3.0128	0.015	9	7.7060	7.6978	0.90	5
Inv Ind	0.9570	0.1931	<0.0001	5	8.3241	3.3704	<0.0001	7	3.6493	3.0132	0.016	7	7.7055	7.6972	0.90	3
Bin Ind	0.9568	0.1933	<0.0001	4	8.3256	3.3704	<0.0001	9	3.6480	3.0136	0.017	6	7.7058	7.6974	0.90	4
NE log Ind	0.9568	0.1942	<0.0001	3	8.3198	3.3705	<0.0001	3	3.6468	3.0153	0.019	3	7.7084	7.6994	0.89	8
NE Ind	0.9574	0.1944	<0.0001	8	8.3181	3.3715	<0.0001	1	3.6477	3.0153	0.019	5	7.7101	7.7008	0.89	9
NE Inv Ind	0.9562	0.1940	<0.0001	2	8.3216	3.3695	<0.0001	4	3.6462	3.0152	0.019	1	7.7070	7.6982	0.89	6
NE Bin Ind	0.9556	0.1936	<0.0001	1	8.3245	3.3677	<0.0001	8	3.6463	3.0147	0.019	2	7.7052	7.6967	0.90	1
Power law + population model																
no BE	0.9734	0.2056	<0.0001	9	8.2217	3.3837	<0.0001	5	3.6221	3.0299	0.067	1	7.7694	7.7451	0.83	9
log Ind	0.9721	0.2041	<0.0001	1	8.2244	3.3828	<0.0001	7	3.6251	3.0266	0.056	8	7.7625	7.7400	0.83	2
Ind	0.9722	0.2035	<0.0001	2	8.2197	3.3823	<0.0001	1	3.6260	3.0260	0.053	9	7.7616	7.7395	0.83	1
Inv Ind	0.9723	0.2046	<0.0001	3	8.2265	3.3832	<0.0001	9	3.6242	3.0273	0.059	7	7.7638	7.7409	0.83	3
Bin Ind	0.9728	0.2051	<0.0001	6	8.2262	3.3834	<0.0001	8	3.6229	3.0285	0.063	6	7.7662	7.7427	0.83	4
NE log Ind	0.9729	0.2057	<0.0001	7	8.2204	3.3838	<0.0001	3	3.6224	3.0298	0.067	3	7.7691	7.7447	0.83	7
NE Ind	0.9730	0.2057	<0.0001	8	8.2199	3.3840	<0.0001	2	3.6223	3.0299	0.067	2	7.7693	7.7449	0.83	8
NE Inv Ind	0.9728	0.2057	<0.0001	5	8.2210	3.3837	<0.0001	4	3.6225	3.0297	0.066	4	7.7688	7.7445	0.83	6
NE Bin Ind	0.9725	0.2056	<0.0001	4	8.2224	3.3833	<0.0001	6	3.6228	3.0293	0.065	5	7.7680	7.7439	0.83	5
Power law + density model																
no BE	0.9640	0.1942	<0.0001	6	8.2572	3.3800	<0.0001	6	3.6536	3.0191	0.017	5	7.7301	7.7179	0.88	9
log Ind	0.9644	0.1927	<0.0001	8	8.2600	3.3798	<0.0001	7	3.6567	3.0169	0.014	8	7.7266	7.7154	0.88	6
Ind	0.9651	0.1922	<0.0001	9	8.2565	3.3791	<0.0001	5	3.6578	3.0166	0.014	9	7.7267	7.7156	0.88	7
Inv Ind	0.9640	0.1930	<0.0001	7	8.2620	3.3797	<0.0001	8	3.6558	3.0170	0.015	7	7.7265	7.7153	0.88	5
Bin Ind	0.9638	0.1934	<0.0001	5	8.2628	3.3801	<0.0001	9	3.6548	3.0176	0.016	6	7.7273	7.7158	0.88	8
NE log Ind	0.9615	0.1932	<0.0001	3	8.2485	3.3765	<0.0001	2	3.6494	3.0203	0.021	2	7.7255	7.7143	0.88	3
NE Ind	0.9621	0.1932	<0.0001	4	8.2478	3.3774	<0.0001	1	3.6493	3.0205	0.021	1	7.7259	7.7145	0.88	4
NE Inv Ind	0.9613	0.1931	<0.0001	1	8.2499	3.3752	<0.0001	3	3.6499	3.0198	0.02	3	7.7250	7.7139	0.88	2
NE Bin Ind	0.9613	0.1928	<0.0001	2	8.2532	3.3737	<0.0001	4	3.6521	3.0185	0.018	4	7.7248	7.7137	0.88	1
Power law + urbanicity model																
no BE	0.9705	0.1962	<0.0001	7	8.2417	3.3891	<0.0001	4	3.6574	3.0213	0.016	5	7.7464	7.7310	0.86	9
log Ind	0.9708	0.1944	<0.0001	8	8.2461	3.3888	<0.0001	7	3.6603	3.0187	0.013	8	7.7420	7.7278	0.87	5
Ind	0.9715	0.1938	<0.0001	9	8.2420	3.3885	<0.0001	5	3.6613	3.0185	0.013	9	7.7421	7.7280	0.87	6
Inv Ind	0.9705	0.1948	<0.0001	6	8.2480	3.3890	<0.0001	8	3.6596	3.0190	0.014	7	7.7424	7.7280	0.87	7
Bin Ind	0.9702	0.1954	<0.0001	5	8.2483	3.3892	<0.0001	9	3.6586	3.0197	0.015	6	7.7435	7.7287	0.86	8
NE log Ind	0.9670	0.1951	<0.0001	2	8.2320	3.3838	<0.0001	2	3.6514	3.0225	0.021	2	7.7402	7.7259	0.87	3
NE Ind	0.9675	0.1949	<0.0001	4	8.2280	3.3849	<0.0001	1	3.6505	3.0231	0.022	1	7.7403	7.7261	0.87	4
NE Inv Ind	0.9668	0.1951	<0.0001	1	8.2360	3.3830	<0.0001	3	3.6527	3.0217	0.019	3	7.7400	7.7258	0.87	2
NE Bin Ind	0.9671	0.1950	<0.0001	3	8.2422	3.3819	<0.0001	6	3.6556	3.0200	0.017	4	7.7400	7.7258	0.87	1

Table 9: Multivariate Dawid-Sebastiani scores for power model predictions (power, population, density, and urbanicity) accounting for effects of proximity to border in different compartments for the 2008 epidemic season at various levels of aggregation. “NE” denotes the covariate is in the neighborhood component, otherwise the covariate is in endemic component.

Although including distance to the border as a covariate in the neighborhood component performed better overall, no one form of the border covariate performed best for every model and aggregation level. Thus weekly and regional scores were determined for all combinations of model and border covariate. To summarize the weekly analysis (Table 10), the mean weekly score and rank is given for each power model and border covariate combination, followed by the mean weekly mDSS for the model without border effects. The best type of border effect covariate again changes from model to model. Again, the best ranked models by week better align with the rankings of the border effect models at the space-time level, not at the time level. All top models remain the same except for the density model, whose best-ranked space-time model includes the inverse border index covariate in the neighborhood component and whose best-ranked average weekly model includes the binary covariate in the neighborhood component. However, the scores of the top two border effects density models at the space-time level are indistinguishable to four decimal places.

The rank of the model without border effects against those with border effects is given along with that weekly mDSS score. The models without border effects rank between middle (5th) and last (9th) by average weekly scoring, as seen for overall mDSS at all levels of aggregation. The rankings of the top power models with border effects also remains unchanged, with the pure power model ranking best, followed by the density, urbanicity and population models. As illustrated in Figure 14, the differences between the various power models remain small. The range of scores over the weeks is much larger than the range of differences in scores for the various models and these differences cannot be seen in the plot. From the full results shown in Table 23, one can see that models with border effects in the neighborhood component perform better than those with border effects in the endemic component when counts are unexpectedly high, such as the early rise in counts in weeks four and five and the unexpected counts in week eighteen. The endemic component border effect models perform better when counts are low or expectedly high, such as during the expected peak around week nine and ten.

	Power		Population		Density		Urbanicity	
	mDSS	rank	mDSS	rank	mDSS	rank	mDSS	rank
log Index	1.0406	8	1.0539	1	1.0474	8	1.0536	8
Index	1.0413	9	1.0541	4	1.0483	9	1.0545	9
Inverse	1.0402	6	1.0540	2	1.0469	7	1.0531	6
Binary	1.0398	4	1.0544	6	1.0465	5	1.0528	5
NE log Index	1.0398	3	1.0545	7	1.0441	3	1.0494	2
NE Index	1.0405	7	1.0546	8	1.0447	4	1.0500	4
NE Inverse	1.0392	2	1.0544	5	1.0438	2	1.0492	1
NE Binary	1.0385	1	1.0541	3	1.0438	1	1.0495	3
No BE	1.0400	5	1.0549	9	1.0467	6	1.0531	7
Best overall	1.0400	1	1.0542	4	1.0457	2	1.0515	3

Table 10: Mean weekly multivariate Dawid-Sebastiani scores for models with border effects and overall means per model type and overall ranking.

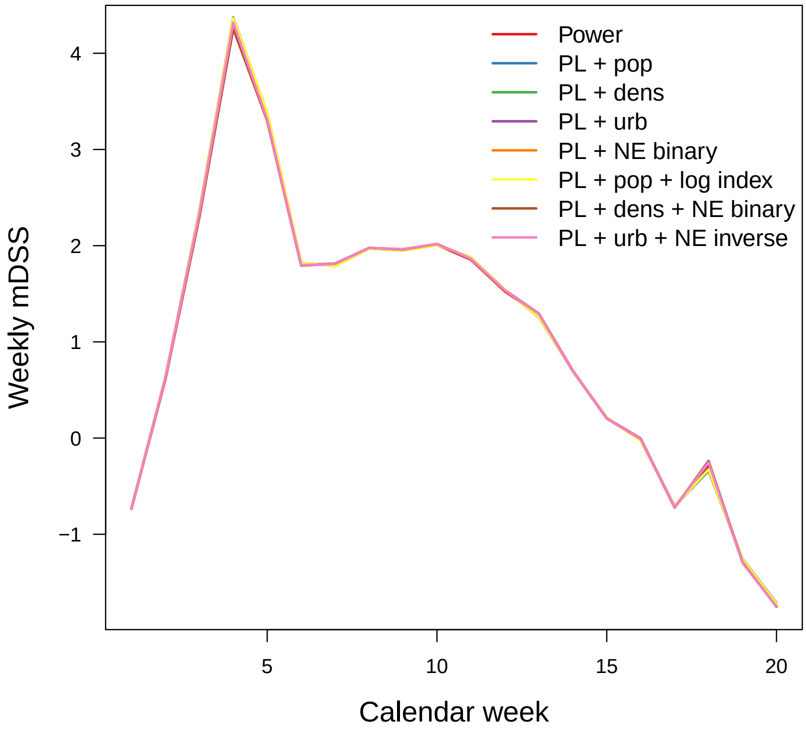

Figure 14: Weekly multivariate Dawid-Sebastiani scores for top-ranked power models with border effects compared to the same models without border effects. There seems to be one line plotted, but they actually lie on top of each other. The scale of the weekly mDSS is large than the scale of the differences in mDSS of the different models.

	Power		Population		Density		Urbanicity	
	mDSS	rank	mDSS	rank	mDSS	rank	mDSS	rank
log index	0.9709	8	0.9860	1	0.9778	8	0.9842	8
index	0.9714	9	0.9860	2	0.9785	9	0.9849	9
inverse	0.9706	6	0.9861	3	0.9774	7	0.9837	7
binary	0.9703	4	0.9865	7	0.9770	5	0.9834	5
NE log index	0.9701	3	0.9865	6	0.9746	2	0.9801	2
NE index	0.9708	7	0.9866	8	0.9751	4	0.9805	4
NE inverse	0.9696	2	0.9864	5	0.9744	1	0.9800	1
NE binary	0.9690	1	0.9862	4	0.9746	3	0.9805	3
no BE	0.9703	5	0.9870	9	0.9770	6	0.9835	6

Table 11: Mean regional multivariate Dawid-Sebastiani scores for power models with border effects.

Average regional scores for power models with border effects (Table 11) reveal similar patterns to those found for the general and weekly analyses. Overall, the power model still performs best, followed by density, urbanicity, and population models. Models without border effects perform worse than those with the covariate in the neighborhood component, but often perform better than those with the covariate in the endemic component. Models with the border effects in the neighborhood component are best for all power models except the population model.

We, however, observe differences between the 10 least and 10 most dense regions (Table 26 and Table 25). Border effect covariates in the endemic component perform better than the neighborhood component covariates more often for both the power and population models for the least dense regions. For the densest regions, the neighborhood covariates perform best more often. The inverse distance to the border covariate in the endemic component never ranks best by mDSS for either cities or rural areas. The distance covariate performs best for the rural regions, but whether to include it in the endemic or neighborhood component changes with the model. For the power and population models, the distance covariate performs best in the endemic component, and for density and urbanicity models it performs better in the neighborhood component. For the densest regions, no clear pattern emerges. The rank patterns for mDSS averaged over regions do not align with the patterns found for regions at extreme ends of the density scale.

5 Updated influenza surveillance data for 2001-2017

Due to the varying nature of seasonal influenza year to year, the results for each epidemic season from 2011 to 2017 are also quite varied. Few patterns between observed epidemic characteristics and model rankings are apparent. The only two years in which all models underestimate the final count are 2013 and 2015, which are also the only two years the epidemic season peaks “on time,” or within one week of the median. Only during these two years do the top ranked models indicated by the mDSS and the univariate Dawid-Sebastiani score match for every aggregation level (compare with Table 21 in appendix).

In general, the top-ranked models by multivariate and univariate Dawid-Sebastiani score are the same for at least 3 of the 4 aggregation levels each year. Across years, the top-ranking models by average weekly and average regional score (see appendix Table 20) also tend to agree with each other and be the same as indicated by mDSS at the space-time level. The results for the 2014 and 2017 seasons, however, are the exception to these findings. The 2014 season is marked by very low peak and final count size while 2017 has the largest counts. Unusually large p-values are observed for space-time, weekly, and regional count aggregation levels for both years.

The performance of power models against the baseline models varies by year and aggregation level and so we discuss specific results for each year.

5.1 Analysis by year

2011

The 2011 epidemic season starts and peaks earlier than most observed seasons, and all models fail to catch this, with predicted peaks at week 9 instead of week 6 (Figure 15).

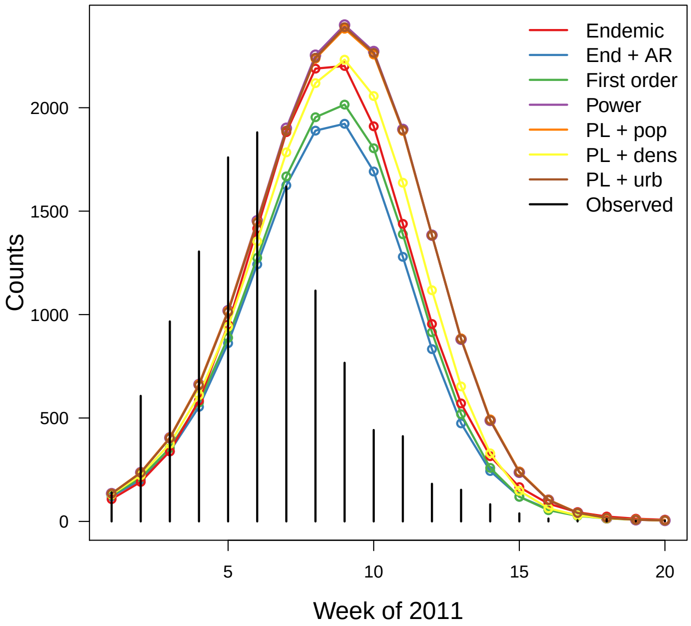

Figure 15: Time series of observed and predictive mean counts for the 2011 epidemic season.

First indications of issues with large variances are apparent in the p-values of the autoregressive and population models at the raw data level (1.00 and 0.0012, Table 12) and for the autoregressive and first order models for regional counts (p-value = 1.00, 0.0043).

Models including space outperform models without at the raw data level in terms of mDSS and DS, with the density model ranking best. The power models rank best at the weekly count level by mDSS. The density model has the lowest DS at both of these levels, partly due to high autocorrelations. The endemic model, ignoring all relationships in space and time, has the second lowest DS at the weekly count level. Weekly scores (Figure 17) are high for all models during the early rise in counts, with the endemic model performing worst during this time.

For regional counts, mDSS indicates the autoregressive and first order models to be best. These two models also have p-values suggesting model calibration but the largest DS scores. Plots (Figure 18 and Figure 16) of each model reveal large 99% prediction intervals for these two models at the weekly and final count levels. They are as large as or larger than that of more complex models for weekly and final counts. In contrast, the univariate score at the space level indicates that the power models perform best. At the final count level the autoregressive model ranks best, followed by the density model. This result is a clear indication that mDSS accounts for both calibration and sharpness because the predicted final count of the first order model is closer to the observed than that of the density model, but the 99% prediction interval for the density model is smaller (Figure 16) and so it ranks second and the first order model third.

For 2011, the power models perform better at the space-time and weekly count levels, but results are mixed for regional and final counts.

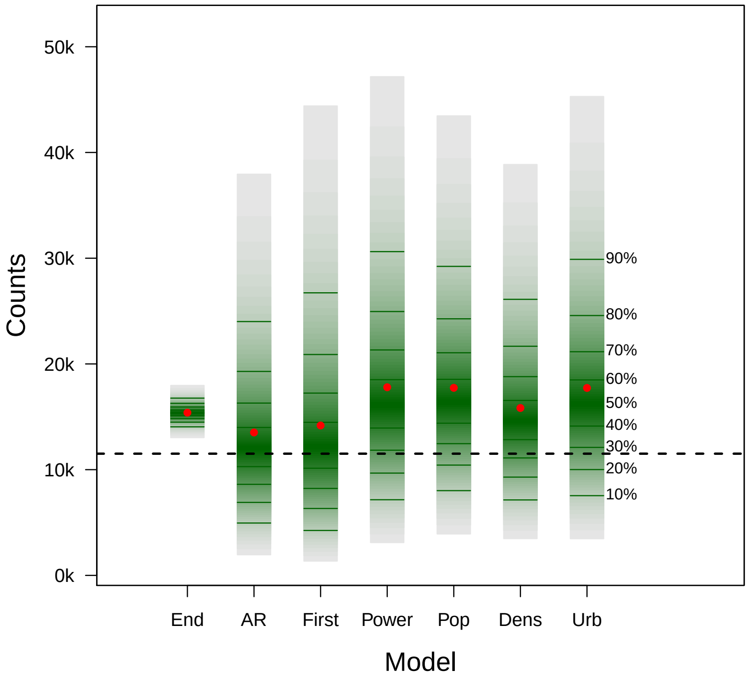

Figure 16: Fan box plots for final count predictions from each model predicting the 2011 epidemic season.

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	1.6696	0.9493	<0.0001	7	12.4613	4.2812	<0.0001	7
End + AR	1.5135	1.0519	1.00	6	6.9887	4.9975	<0.0001	6
First order	1.4779	0.8417	<0.0001	3	6.6552	4.8152	<0.0001	5
Power	1.4849	0.9247	<0.0001	5	6.1738	4.3859	<0.0001	1
PL + pop	1.4661	0.9245	0.0012	2	6.3110	4.3486	<0.0001	3
PL + dens	1.4241	0.7839	<0.0001	1	6.6126	4.1939	<0.0001	4
PL + urb	1.4811	0.9213	<0.0001	4	6.2532	4.3702	<0.0001	2
	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	5.2103	3.4522	<0.0001	7	13.6624	6.9662	0.0003	7
End + AR	4.7684	4.5221	1.00	2	8.9976	8.9647	0.80	1
First order	4.6705	3.9994	0.0043	1	9.1861	9.1456	0.78	3
Power	4.8576	3.8164	<0.0001	4	9.3816	9.1658	0.51	6
PL + pop	4.7970	3.8101	<0.0001	3	9.3203	9.0563	0.47	4
PL + dens	4.9202	3.7205	<0.0001	6	9.1039	8.9451	0.57	2
PL + urb	4.8593	3.8087	<0.0001	5	9.3479	9.1131	0.49	5

Table 12: Multivariate Dawid-Sebastiani score, determinant sharpness, p-value, and rank for 2011 predictions by each model at each data aggregation level.

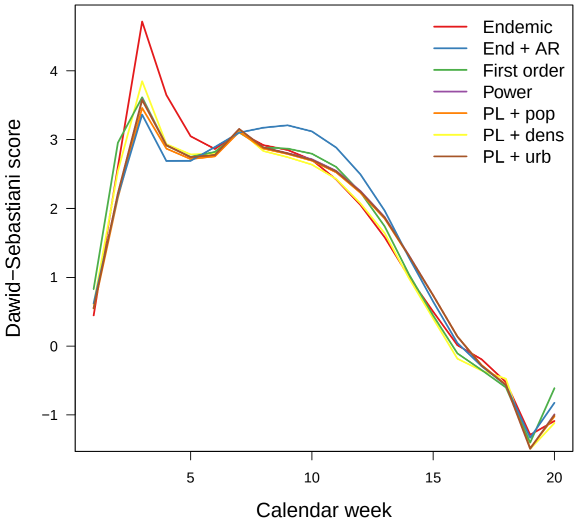

Figure 17: Weekly mDSS for 2011.

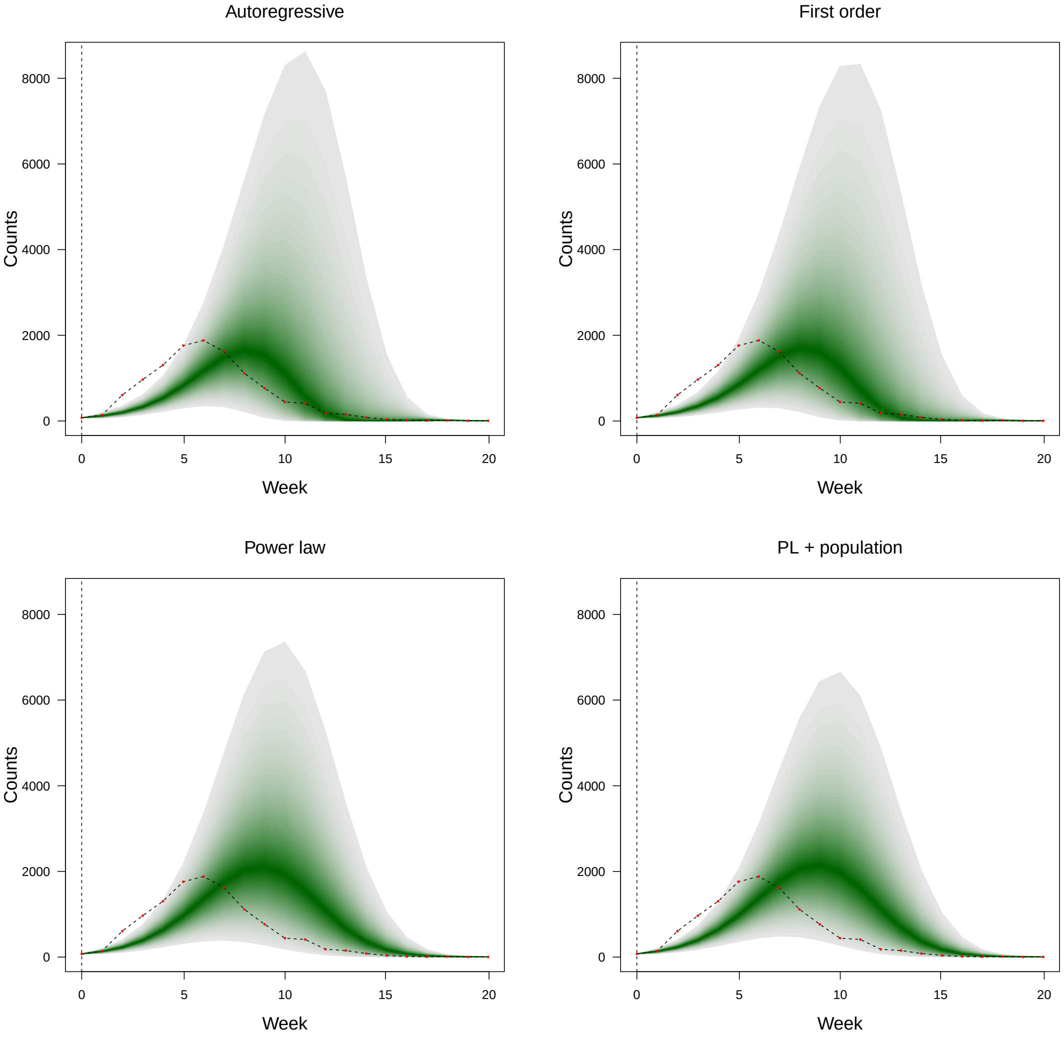

Figure 18: 2011 prediction fan plots for selected models: autoregressive, first order, power, and population models.

2012

The 2012 epidemic season is late (peak at week 11, median peak week = 9), with a relatively low final count. A power model ranks best and the baseline models tend to rank at the center or worst for all aggregation levels except the original space-time level (Table 13). Power models have the lowest DS at all levels except the final count level. Here, the endemic model has the lowest DS, but there is evidence against only the endemic model's calibration (p-value <0.0001) at this level and this model ranks last by mDSS.

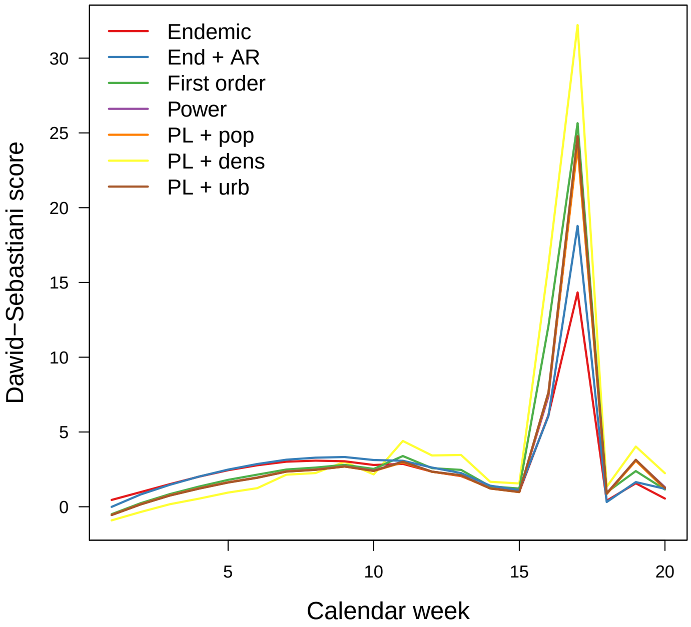

Figure 19: Plotted weekly mDSS scores for models predicting the 2012 epidemic season.

The weekly mDSS score is very sensitive to unexpected counts at the end of the epidemic season (Figure 19). At week 17 there is a large spike in the weekly score of all models, where more suburban/rural towns (LK Passau, LK Rems-Murr-Kreis, LK Rhön-Grabfeld) still have 2 or 3 counts of influenza at a time when they had previously had zero. The two models without a spatial component, the endemic and autoregressive models, perform best at this time and during a smaller peak at week 19, but are worst for the first 10 weeks of the prediction. These models also overestimate the counts the most (Figure 20, top and bottom).

Often, there is evidence against calibration of all models at the regional aggregation level, but in 2012 such evidence exists only against the density and endemic models (Table 13). In Figure 21 we see that the density model has the smallest 98% prediction interval and the autoregressive model the widest. This pattern seems to occur at the regional level as well.

Often, the power models perform alike, but in all plots a clear difference between the density model and other power models is obvious. The density model is the only model to underestimate the influenza counts, however, its final count prediction is only 172 counts below the observed 6172 counts, the closest to the observed of any model. At the final count level, this model is second sharpest (endemic first), and it is most sharp by DS at every other aggregation level. However, by mDSS, this model performs poorly at all levels, except the final count level where it is ranked best. Differences between the density model and other models is also apparent in the weekly fanplots (Figure 21), where the density model's underestimation and sharpness are depicted.

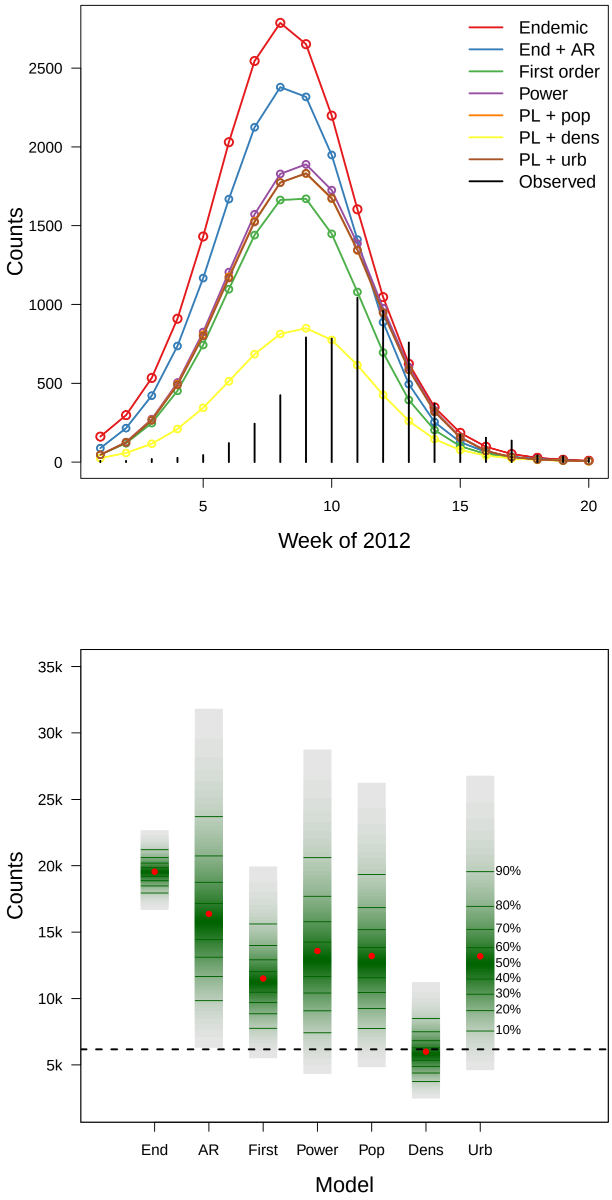

Figure 20: Time series of observed and mean predicted counts (top) and fan box plots for final count predictions (bottom) for models predicting the 2012 epidemic season.

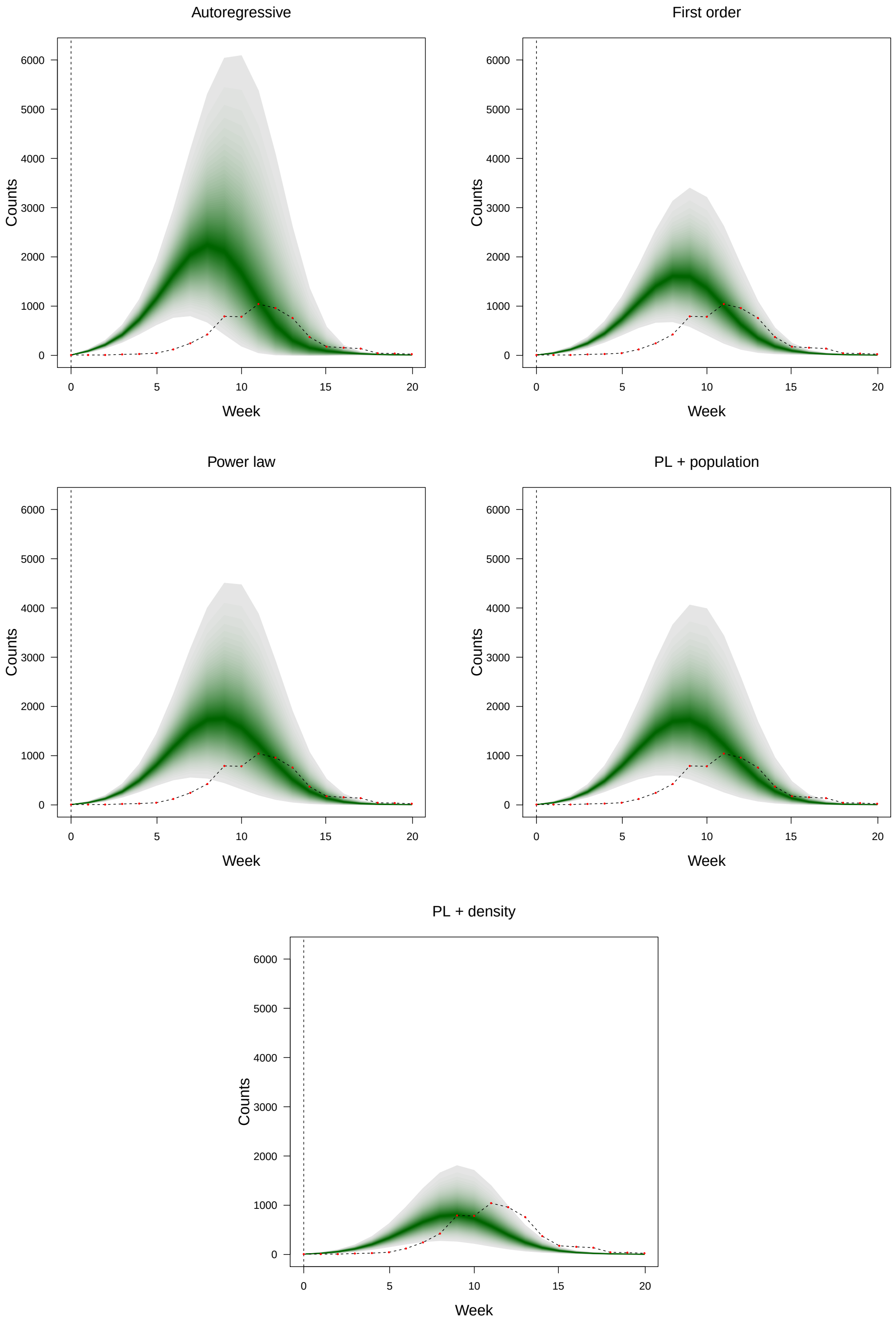

Figure 21: 2012 prediction fan plots for selected models: autoregressive, first order, power, population, and density models.

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	2.7323	1.1369	<0.0001	1	11.5200	4.4527	<0.0001	5
End + AR	2.8909	1.0826	<0.0001	2	9.9743	4.5900	<0.0001	4
First order	3.3581	0.6347	<0.0001	6	12.8068	3.8486	<0.0001	6
Power	3.0270	0.6471	<0.0001	4	9.6596	3.9194	<0.0001	2
PL + pop	3.0065	0.6313	<0.0001	3	9.6528	3.8573	<0.0001	1
PL + dens	3.9749	0.1136	<0.0001	7	15.0299	3.2344	<0.0001	7
PL + urb	3.0551	0.6268	<0.0001	5	9.7727	3.8704	<0.0001	3
	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	4.9483	3.6601	<0.0001	7	62.5745	7.1475	<0.0001	7
End + AR	4.7860	4.6020	1.00	6	10.3348	8.6121	0.063	6
First order	4.1232	3.7324	0.97	4	9.5221	8.0382	0.085	2
Power	4.0101	3.5212	0.56	3	9.5594	8.5688	0.16	4
PL + pop	3.9846	3.4993	0.58	1	9.6004	8.4361	0.13	5
PL + dens	4.6459	2.8073	<0.0001	5	7.5422	7.5380	0.93	1
PL + urb	4.0089	3.4938	0.39	2	9.5486	8.4715	0.14	3

Table 13: Multivariate Dawid-Sebastiani score, determinant sharpness, p-value, and rank for 2012 predictions by each model at each data aggregation level.

2013

The 2013 epidemic season is characterized by a tall peak of 2374 counts at week 8 (median week 8.5). All models except the endemic model underestimate the counts (Figure 22, left and right). The 98% prediction interval of the endemic model, however, barely supports the observed final count of 18325. At the final count and space-time level, the power models perform best in terms of mDSS and DS (Table 14). The power models give similar predictions to baseline models at the final count level, but have smaller 98% prediction intervals.

On the other hand, the baseline models perform best for weekly and regional counts. Unusually high p-values and larger prediction intervals (Table 14 and Figure 24) draw attention to issues with variance/overdispersion estimation of the baseline models. Except at the final count level, DS indicates that power models are sharper than baseline models.

The weekly scores fluctuate more in 2013 (Figure 23). At week 15 all models see a sharp rise in score. As seen in Figure 22 (left) there is a local peak at this time. Counts which are usually low at this time reach as high as 39 in Ludwigsburg and 16 in Rastatt. The power model scores fluctuate most between weeks 4 and 8. This is partly due to the dip in counts at week 7. Until then, the power models are underestimating the counts. Because of the dip at week 7 the predictions are relatively close to the observed, but the counts increase again at week 8.

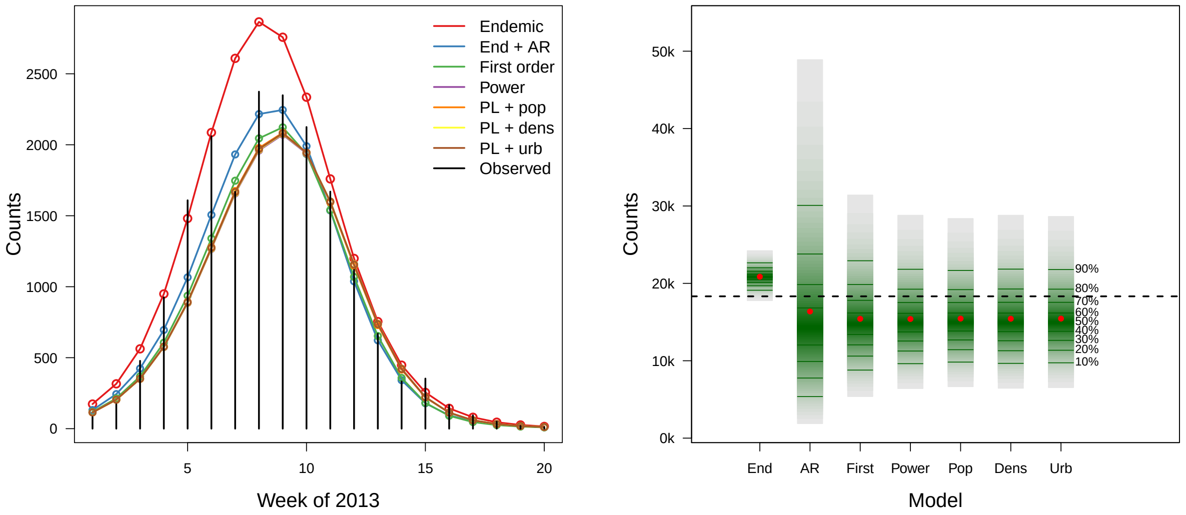

Figure 22: Time series of observed and predictive mean counts (left), fan box plots for final count predictions (right) for models predicting the 2013 epidemic season.

	Space-time				Time			
	mDSS	DS	*p*-value	rank	mDSS	DS	*p*-value	rank
Endemic	1.7369	1.3350	1.00	6	5.2021	4.6549	0.35	1
End + AR	1.7711	1.3441	1.00	7	5.5653	5.2432	0.88	3
First order	1.6462	0.9806	<0.0001	5	5.3747	4.4238	0.0088	2
Power	1.6037	0.8690	<0.0001	3	5.7451	4.0617	<0.0001	5
PL + pop	1.5935	0.8698	<0.0001	1	5.7643	4.0459	<0.0001	7
PL + dens	1.6044	0.8695	<0.0001	4	5.7430	4.0613	<0.0001	4
PL + urb	1.6025	0.8702	<0.0001	2	5.7491	4.0563	<0.0001	6
	Space				Final			
	mDSS	DS	*p*-value	rank	mDSS	DS	*p*-value	rank
Endemic	5.1736	3.7752	<0.0001	3	8.9141	7.2325	0.067	6
End + AR	5.1027	4.9091	1.00	1	9.2512	9.2329	0.85	7
First order	5.1628	4.0520	<0.0001	2	8.7681	8.6347	0.61	5
Power	5.5869	3.6385	<0.0001	6	8.6662	8.4823	0.54	4
PL + pop	5.4560	3.6438	<0.0001	4	8.6410	8.4521	0.54	1
PL + dens	5.5899	3.6405	<0.0001	7	8.6612	8.4807	0.55	3
PL + urb	5.5666	3.6428	<0.0001	5	8.6522	8.4697	0.55	2

Table 14: Multivariate Dawid-Sebastiani score, determinant sharpness, *p*-value, and rank for 2013 predictions by each model at each data aggregation level.

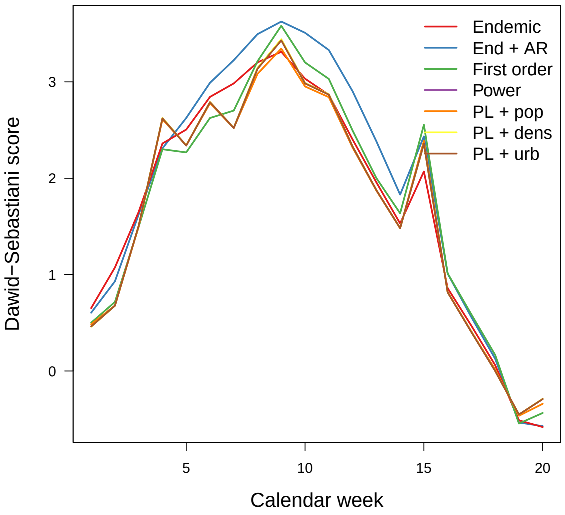

Figure 23: Plotted weekly mDSS scores for models predicting the 2013 epidemic season.

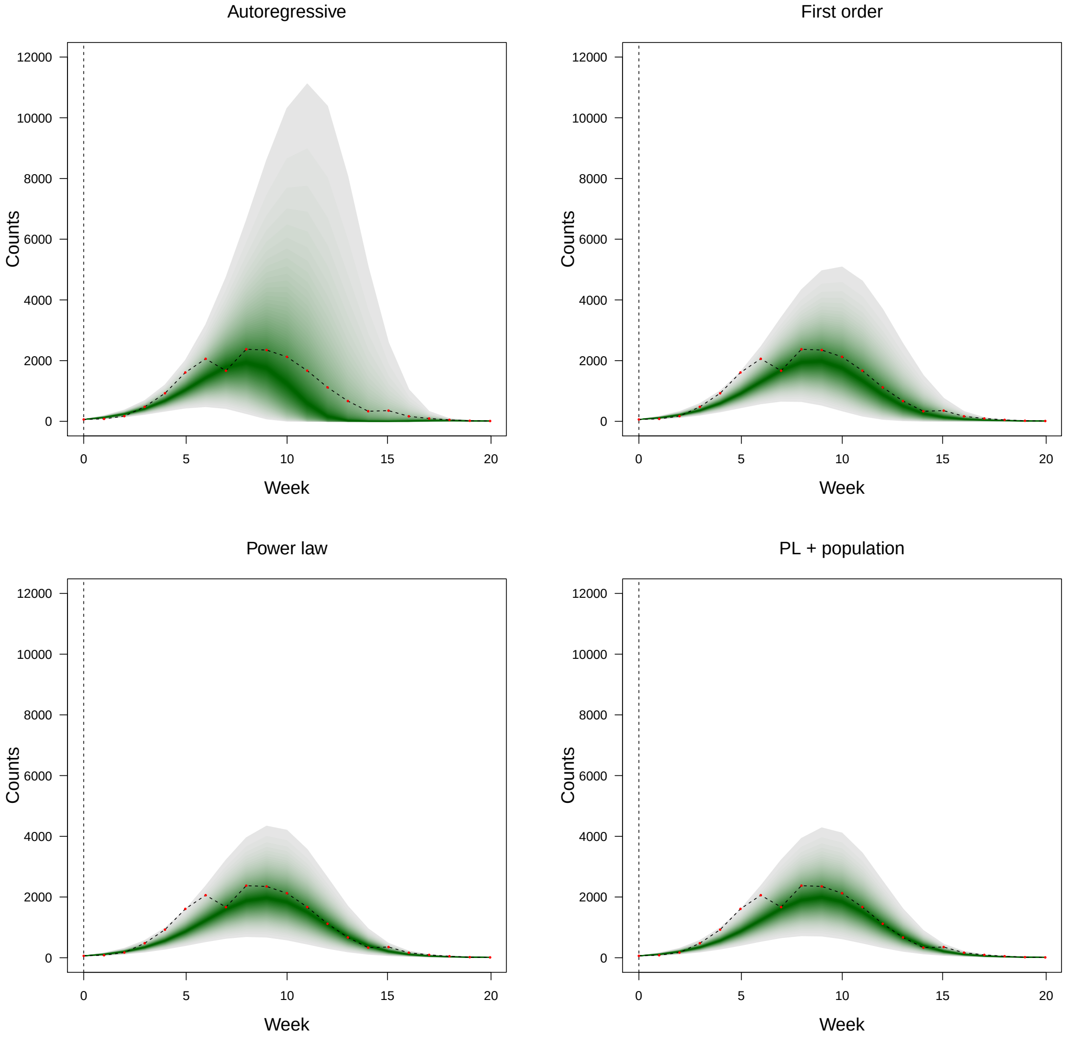

Figure 24: 2013 prediction fan plots for selected models: autoregressive, first order, power, and population models.

2014

The 2014 epidemic season is marked by the lowest peak and final count of the years since 2008. All models overestimate the incidence, but power models overestimate less than the baseline models (Figure 25, left and right). The time series predicted by every model expects a peak in counts that never really occurs. These predicted peaks lead to higher predicted final counts than actually observed.

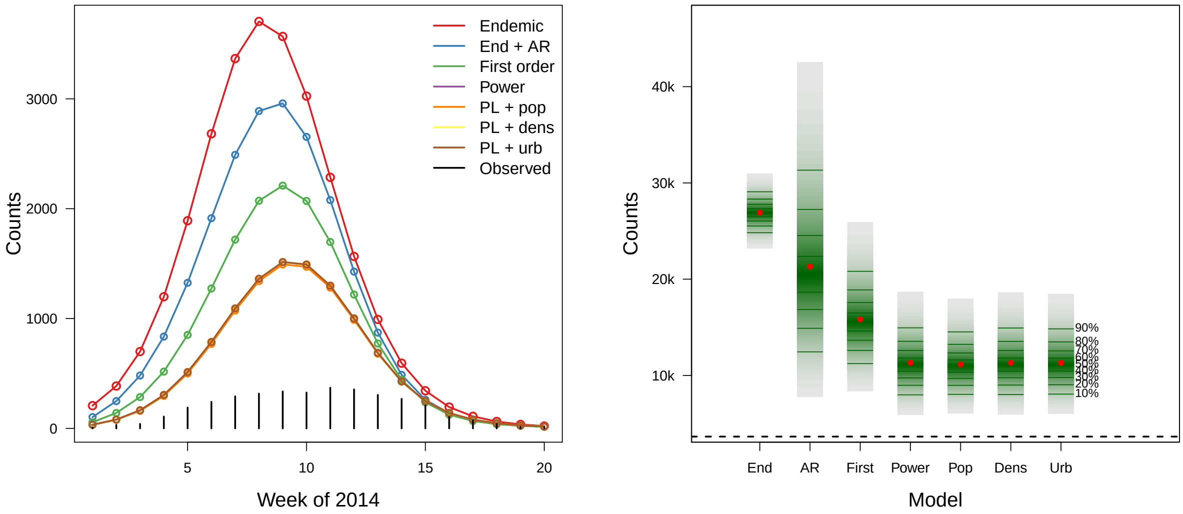

Figure 25: Time series of observed and predictive mean counts (left) and fan box plots for final count predictions (right) for models predicting the 2014 epidemic season.

The power models perform best by mDSS at all aggregation levels except the final count level. The endemic and first order models rank among the worst at every aggregation level. Power models are also among the sharpest at all levels except for final counts. The sharpest model at the final count level is the endemic model, but we observe more evidence against calibration of this model than for the others. At the final count level the mDSS indicates that the autoregressive model performs best, however, from the fan box plots of the final counts in Figure 15, one can see that the autoregressive model prediction overestimates much more than the other models and has an unusually large prediction interval. The difference between the observed and predicted count is twice as large for the autoregressive model as the second best model, the density model, but its variance is seven times as large, giving the autoregressive model the advantage in the second part of the mDSS equation $((\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{P}})^\top \boldsymbol{\Sigma}_{\boldsymbol{P}}{}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{P}}))$. The DS, however, correctly indicates that the autoregressive model is the least sharp.

As seen in Table 20 (appendix), mean weekly and mean regional scores resulting from "slicing" the data usually indicate the same top-ranking model. In 2014, however, the population model ranks best by the weekly score and the power model by regional score. The top models by these mean scores also usually correspond with the top model at the space-time aggregation level. Here, the model indicated at this aggregation level only agrees with the regional score. When the "slice" scores disagree, there is also usually more disagreement between the multivariate and univariate scores. While here they indicate different models at the raw and time levels, they both indicate some form of power model.

Unusual activity is detected in Figure 26. At week 5, 56 cases are reported in LK Miltenberg, which otherwise reports few cases. All models underestimate this unexpected count, but there is a larger penalty for this for the power models. The power models have smaller (better) weekly mDSS scores for almost the whole prediction period, but rise sharply above the baseline models at week 5.

As in 2008, the 98% prediction interval for the final count predicted by the endemic model does not even contain the observed final count of 3646. However, in 2014, no model does; all prediction intervals fall above the dashed line representing the observed final count. The p-values at the final count level of Table 15 are also lower than usual. There is evidence against calibration of all models at an aggregation level where there usually is not. This is corroborated by Figure 25, right.

Unusually high p-values are observed for all other aggregation levels in 2014. Unlike in other years, however, this observation is not limited to the autoregressive and first order models, but common to most models. In Figure 27, however, an unusually wide prediction interval is only observed for the autoregressive model.

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	1.8466	1.5174	1.00	7	11.7834	4.8402	<0.0001	7
End + AR	1.6496	1.4240	1.00	6	6.6284	4.9939	<0.0001	6
First order	1.2932	0.9641	1.00	5	5.2156	4.1593	0.0026	5
Power	1.1242	0.6495	0.97	1	4.3669	3.7746	0.26	4
PL + pop	1.1273	0.6420	0.86	4	4.3441	3.7436	0.24	1
PL + dens	1.1264	0.6493	0.96	2	4.3661	3.7717	0.25	3
PL + urb	1.1270	0.6491	0.95	3	4.3633	3.7641	0.24	2
	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	6.0705	3.9637	<0.0001	7	105.8109	7.4146	<0.0001	7
End + AR	5.1939	4.9333	1.00	6	11.6994	8.9229	0.018	1
First order	4.2117	3.9754	1.00	5	13.4501	8.2355	0.0012	6
Power	3.6191	3.3168	1.00	3	11.8387	7.9161	0.0051	2
PL + pop	3.6082	3.3043	1.00	1	12.1768	7.8444	0.0032	5
PL + dens	3.6200	3.3174	1.00	4	11.8957	7.9088	0.0047	3
PL + urb	3.6147	3.3171	1.00	2	12.0253	7.8886	0.004	4

Table 15: Multivariate Dawid-Sebastiani score, determinant sharpness, p-value, and rank for 2014 predictions by each model at each data aggregation level.

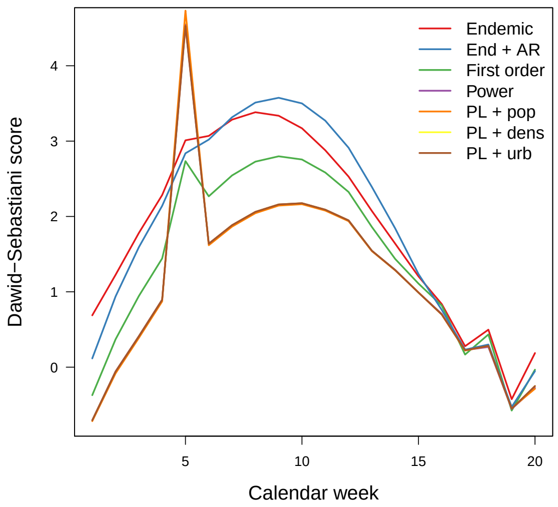

Figure 26: Plotted weekly mDSS scores for models predicting the 2014 epidemic season.

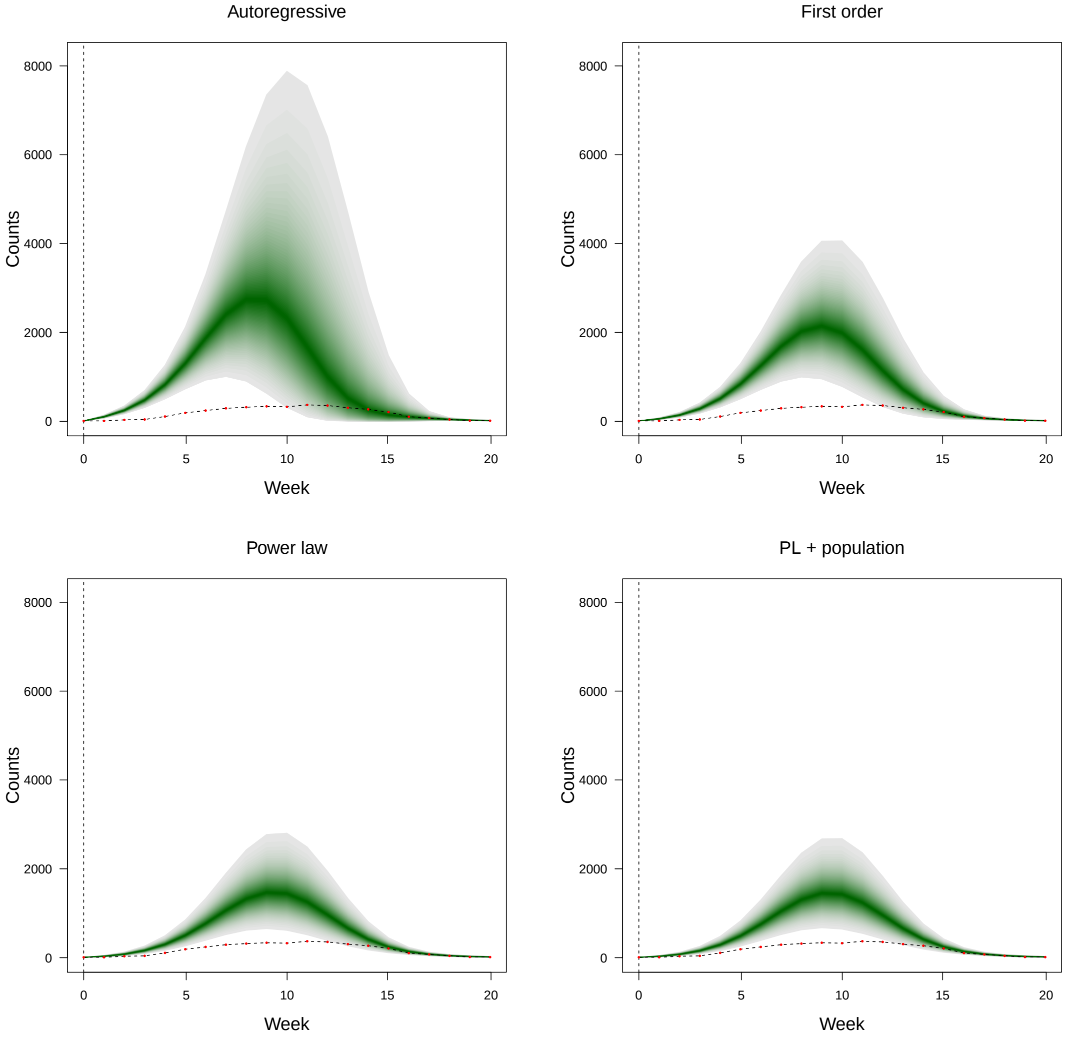

Figure 27: 2014 prediction fan plots for selected models: autoregressive, first order, power, and population models.

2015

The 2015 epidemic season peaks higher and ends with a larger final count than any observed previous year (except for the 2009 pandemic). An unexpected peak of 431 cases occurs at week 16 (Figure 28, left), with 62 cases in LK Straubing-Bogen, 58 cases in LK Passau, and 39 cases in LK Regensburg. This causes a dramatic peak in weekly scores for all models (Figure 28, bottom), with power models performing worse at this time. Again, the power model scores fluctuate more than baseline model scores, but in 2015 the power models also often perform worse according to weekly mDSS scores. The differences in penalties for the peak size predictions at week 8 are, however, much more subtle.

As seen on the map in Figure 38, cases in 2015 have a different spatial pattern: the county surrounding Munich has a higher final count than the city itself. This is also true for the counties surrounding Bamberg and Passau. Because of these peculiarities in time and space, the baseline models perform better than the power models at all aggregation levels (Table 16). According to DS, the power models are sharper, owing to both large spatial/temporal correlations and smaller variance.

All models underestimate the counts (Figure 28, left and right) but the 98% prediction intervals of all models support the observed final count of 25774. The endemic model ranks best by mDSS twice in 2015, at the space-time and final count levels, and the autoregressive model ranks best at the time and space aggregations. As observed in several years, calibration p-values for the autoregressive model are unusually high at the time and space aggregations. This is clearly depicted in Figure 29.

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	3.6607	1.5266	<0.0001	1	7.2460	4.8435	<0.0001	3
End + AR	4.0271	1.5029	<0.0001	2	5.9287	5.5011	0.65	1
First order	4.5179	1.0863	<0.0001	3	6.3937	4.5630	<0.0001	2
Power	4.6471	0.9125	<0.0001	7	8.3086	4.1140	<0.0001	4
PL + pop	4.5200	0.9117	<0.0001	4	8.4655	4.0916	<0.0001	7
PL + dens	4.6347	0.9125	<0.0001	6	8.3426	4.1115	<0.0001	5
PL + urb	4.6014	0.9123	<0.0001	5	8.3967	4.1053	<0.0001	6
	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	6.6291	3.9421	<0.0001	3	7.6621	7.3760	0.45	1
End + AR	5.3932	5.1237	1.00	1	9.6832	9.5535	0.61	3
First order	5.6816	4.1324	<0.0001	2	9.5678	8.8237	0.22	2
Power	7.9458	3.6111	<0.0001	7	11.0718	8.4281	0.021	4
PL + pop	7.8911	3.6120	<0.0001	4	11.2713	8.3883	0.016	7
PL + dens	7.9414	3.6122	<0.0001	6	11.0877	8.4226	0.021	5
PL + urb	7.9339	3.6125	<0.0001	5	11.1478	8.4089	0.019	6

Table 16: Multivariate Dawid-Sebastiani score, determinant sharpness, p-value, and rank for 2015 predictions by each model at each data aggregation level.

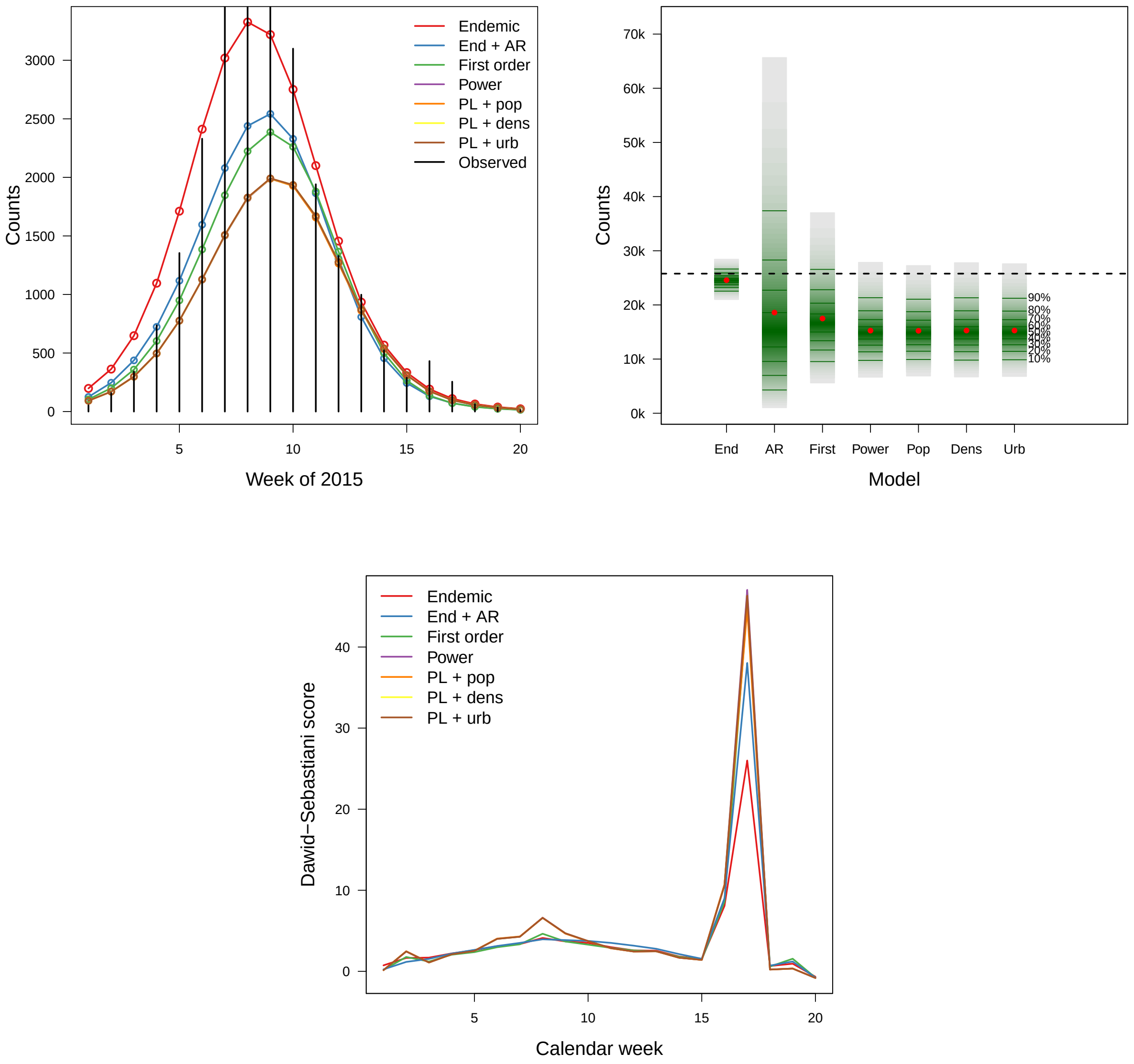

Figure 28: Time series of observed and mean predicted counts (left), fan box plots for final count predictions (right), and plotted weekly mDSS scores (bottom) for models predicting the 2015 epidemic season.

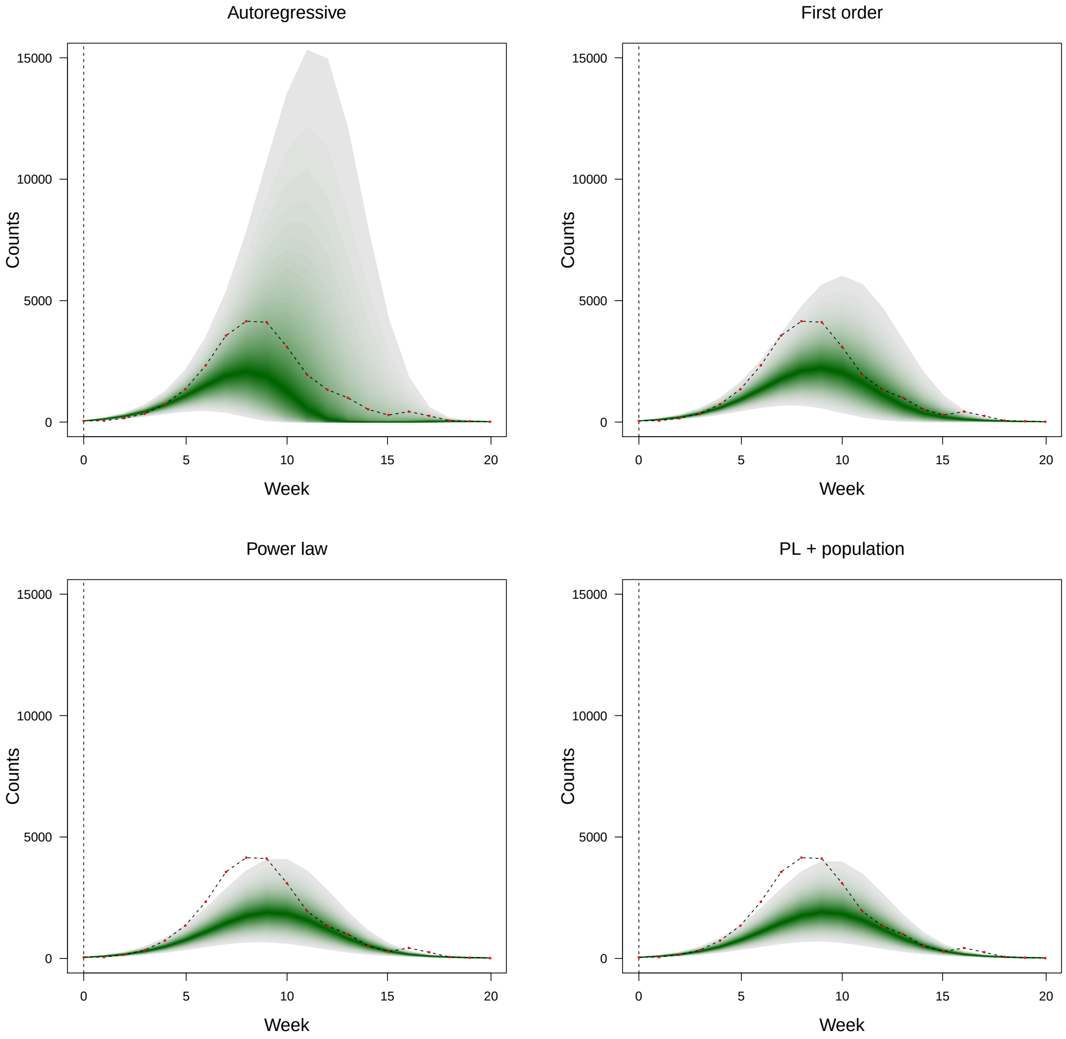

Figure 29: 2015 prediction fan plots for selected models: autoregressive, first order, power, and population models.

2016

The 2016 epidemic season peaks late at week 11 (median 9) and, although not a low count year (12880 final count), it is much smaller than 2015 and 2017, with 26078 and 27143 counts each. This pattern of high incidence year followed by a lower incidence year is common in seasonal infectious disease. A portion of the population may be protected from infection in 2016 by previous infection or vaccination in 2015. The time series also exhibits a small first peak in week 5 before really taking off later in the season. As in 2008, no model predicts this.

The baseline models overestimate more than the power models (Figure 30, left), leading the power models to perform better at almost all levels (Table 17). The 98% final count prediction interval of the endemic model does not even support the observed count (Figure 30, right). The first order model, however, does rank best for the weekly count level. Again, the DS of power models is lower than that of baseline models. This is readily apparent in Figure 32, in which the scale necessary to visualize the autoregressive and first order models makes it difficult to see and interpret the power models.

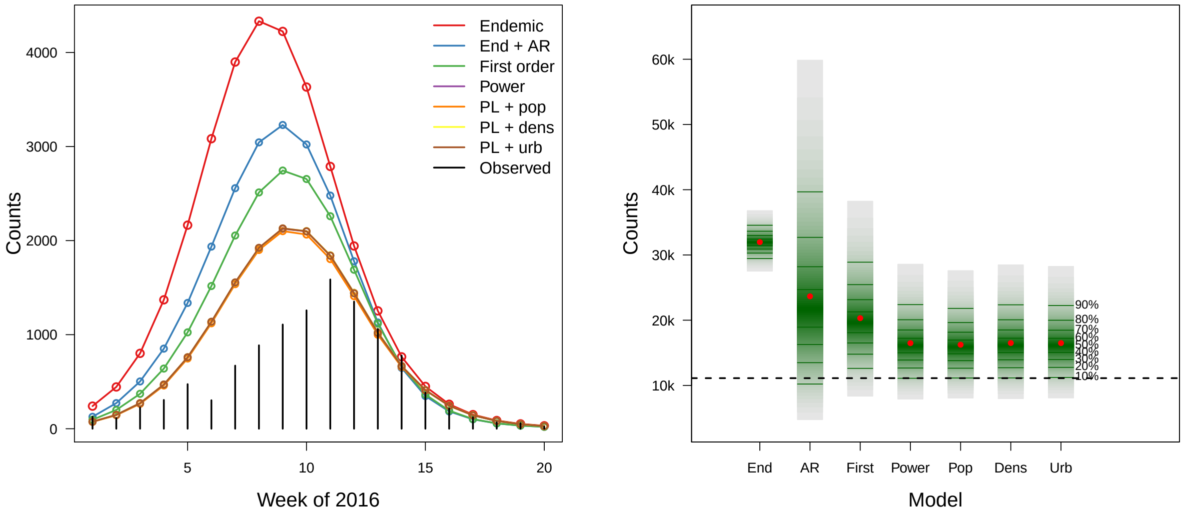

Figure 30: Time series of observed and mean predicted counts (left), fan box plots for final count predictions (right) for models predicting the 2016 epidemic season.

Regions with medium to high counts are not limited to the largest cities in 2016 (Figure 38); several cities and counties have counts as high as or larger than that of Munich. At the county level, we again observe unusually high p-values, this time for all models except the endemic model.

The differences in plots of weekly scores from baseline and power models are visually greater this year for the period from week 4 to 13 (Figure 31). The power models perform best here. However, the power models and the first order model are heavily penalized for underestimating in the first prediction week.

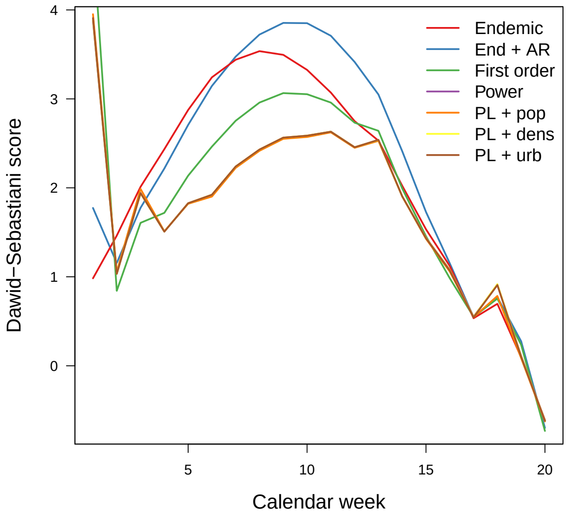

Figure 31: Plotted weekly mDSS scores for models predicting the 2016 epidemic season.

	Space-time				Time			
	mDSS	DS	*p*-value	rank	mDSS	DS	*p*-value	rank
Endemic	2.0276	1.7492	1.00	7	8.7184	5.0587	<0.0001	7
End + AR	2.0000	1.6931	1.00	6	6.1712	5.4554	0.095	5
First order	1.8276	1.2504	<0.0001	5	5.8221	4.6625	0.0007	1
Power	1.6331	1.0106	<0.0001	2	6.1543	4.2192	<0.0001	3
PL + pop	1.6246	0.9990	<0.0001	1	6.2361	4.1764	<0.0001	6
PL + dens	1.6340	1.0104	<0.0001	4	6.1529	4.2145	<0.0001	2
PL + urb	1.6335	1.0091	<0.0001	3	6.1595	4.2041	<0.0001	4
	Space				Final			
	mDSS	DS	*p*-value	rank	mDSS	DS	*p*-value	rank
Endemic	5.5338	4.1839	<0.0001	7	62.4449	7.5966	<0.0001	7
End + AR	5.4217	5.3252	1.00	6	9.9395	9.3882	0.29	6
First order	4.5616	4.3287	1.00	5	9.7897	8.7713	0.15	5
Power	4.3124	3.7349	0.10	3	9.1279	8.4015	0.23	2
PL + pop	4.3106	3.7218	0.074	1	9.0866	8.3441	0.22	1
PL + dens	4.3124	3.7361	0.10	4	9.1363	8.3936	0.22	3
PL + urb	4.3119	3.7355	0.10	2	9.1419	8.3764	0.22	4

Table 17: Multivariate Dawid-Sebastiani score, determinant sharpness, *p*-value, and rank for 2016 predictions by each model at each data aggregation level.

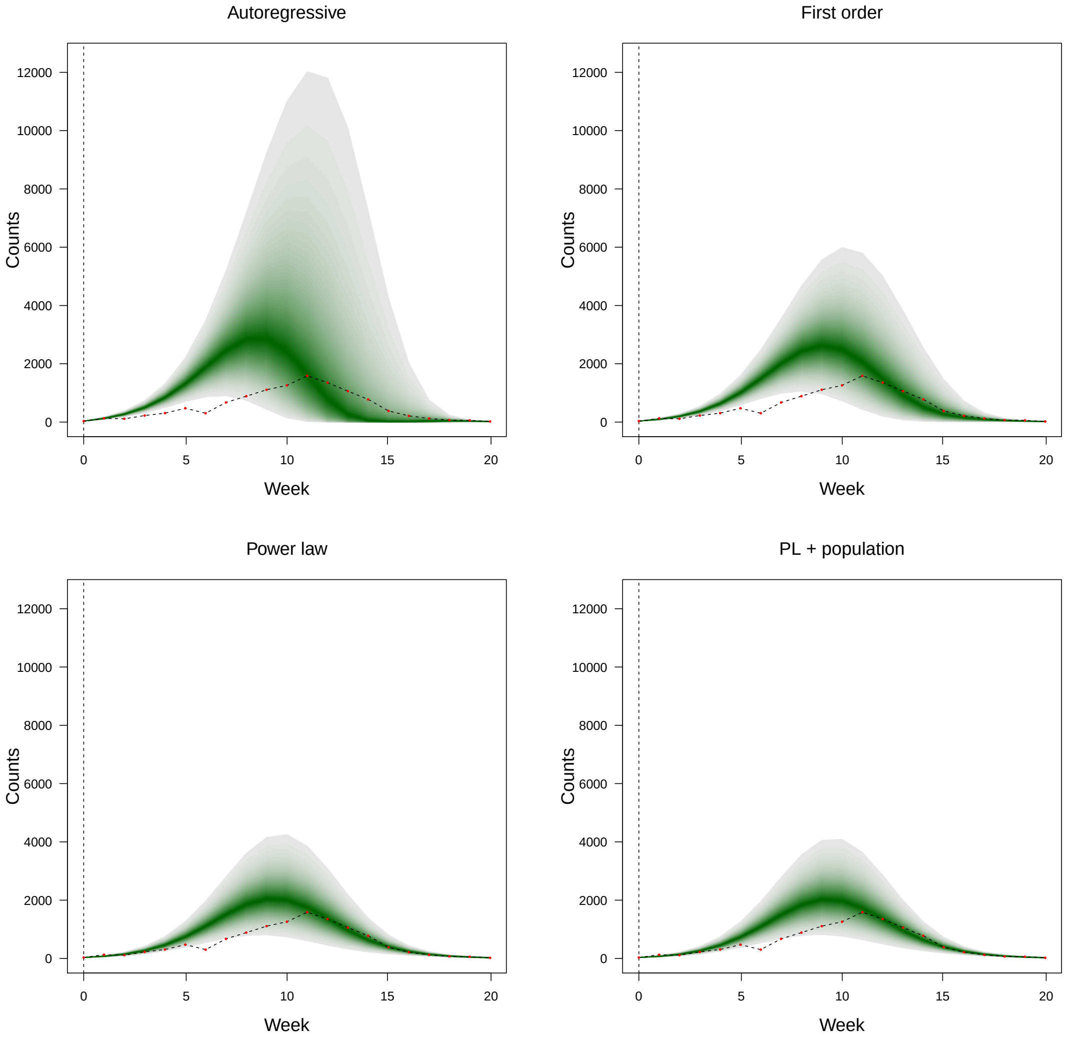

Figure 32: 2016 prediction fan plots for selected models: autoregressive, first order, power, and population models.

2017

The 2017 epidemic season is marked by the largest peak and final size, and peaks earlier (week 6, median 8.5) than most years. This combination leads the power models, but also the first order model, to severely overestimate the weekly and final counts (Figure 33, left and right). Conditioning on such high counts as the 671 counts in the last week of 2016 leads these models to predict excessively large peaks. Nonetheless, the first order model ranks best at the space-time and final count levels and the power models perform best for aggregations over time or space (Table 18). In 2017, DS indicates one model to be sharpest across all aggregation levels: the endemic model. Again we observe extremely wide 98% prediction intervals (Figure 34) and unusually large p-values at the raw, time, and space aggregation levels for almost all models.

Figure 33 (bottom) depicts high penalties for the endemic model underestimating the early counts. It is worth noting that the autoregressive model and models with spatial components perform poorly compared to the endemic model from week 6 on, but for different reasons. The spatial models extremely overestimate the counts while the variance of the autoregressive model is extremely large.

This is one of only two years in which the top model by average weekly, regional, and univariate scores do not match the top model according to mDSS at the space-time level. The space-time mDSS indicates the first order model performs best (Table 18), while the average weekly, regional, and univariate scores point to the endemic, autoregressive, and endemic models, respectively (Table 19).

	Space-time				Time			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	2.5498	1.9016	<0.0001	6	16.1832	5.1704	<0.0001	7
End + AR	2.5350	2.2979	1.00	2	7.8759	7.7054	1.00	6
First order	2.5143	2.3534	1.00	1	7.2352	7.1290	1.00	5
Power	2.5405	2.4311	1.00	3	6.3700	5.9867	0.76	3
PL + pop	2.5565	2.4533	1.00	7	6.3738	5.9690	0.70	4
PL + dens	2.5421	2.4327	1.00	4	6.3672	5.9830	0.76	2
PL + urb	2.5440	2.4354	1.00	5	6.3639	5.9771	0.75	1
	Space				Final			
	mDSS	DS	p-value	rank	mDSS	DS	p-value	rank
Endemic	5.7695	4.3010	<0.0001	4	15.6201	7.6578	<0.0001	7
End + AR	6.0994	6.0061	1.00	7	11.9177	11.9104	0.90	2
First order	5.8722	5.7266	1.00	6	11.8133	11.5659	0.48	1
Power	5.7418	5.6111	1.00	1	12.6563	10.5606	0.041	3
PL + pop	5.7836	5.6658	1.00	5	12.8786	10.5789	0.032	6
PL + dens	5.7464	5.6169	1.00	2	12.6834	10.5601	0.039	4
PL + urb	5.7534	5.6261	1.00	3	12.7298	10.5587	0.037	5

Table 18: Multivariate Dawid-Sebastiani score, determinant sharpness, p-value, and rank for 2017 predictions by each model at each data aggregation level.

	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	2.5498	2.8437	2.7224	2.6820	2.6974	2.6837	2.6857
Regional	2.5498	2.5350	2.5531	2.5952	2.6111	2.5967	2.5986
Univariate	5.0995	5.6875	5.6048	5.6683	5.7011	5.6713	5.6751

Table 19: Mean weekly and regional multivariate Dawid-Sebastiani scores followed by univariate Dawid-Sebastiani scores for 2017 predictions.

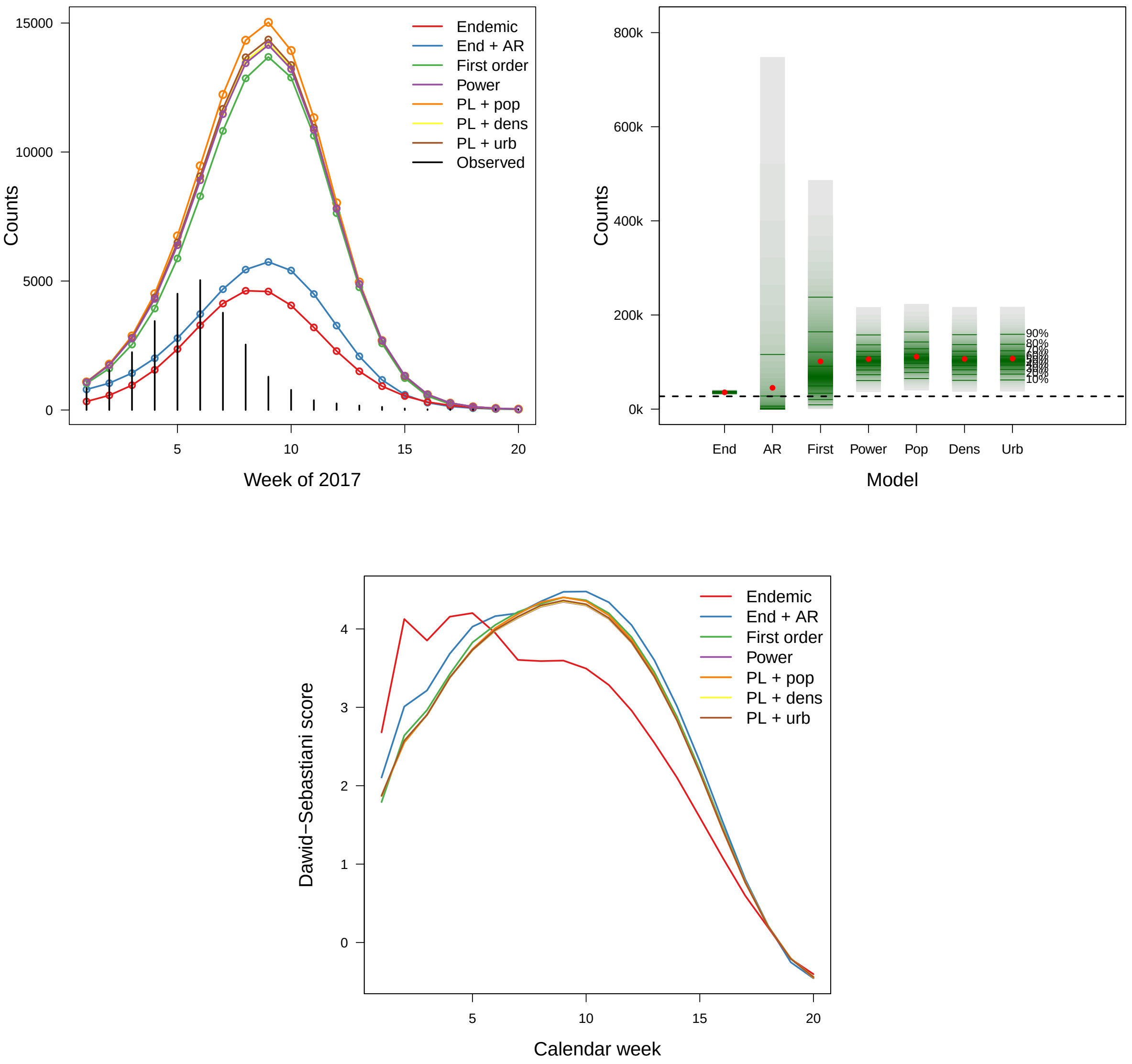

Figure 33: Time series of observed and mean predicted counts (left), fan box plots for final count predictions (right), and plotted weekly mDSS scores for models predicting the 2017 epidemic season.

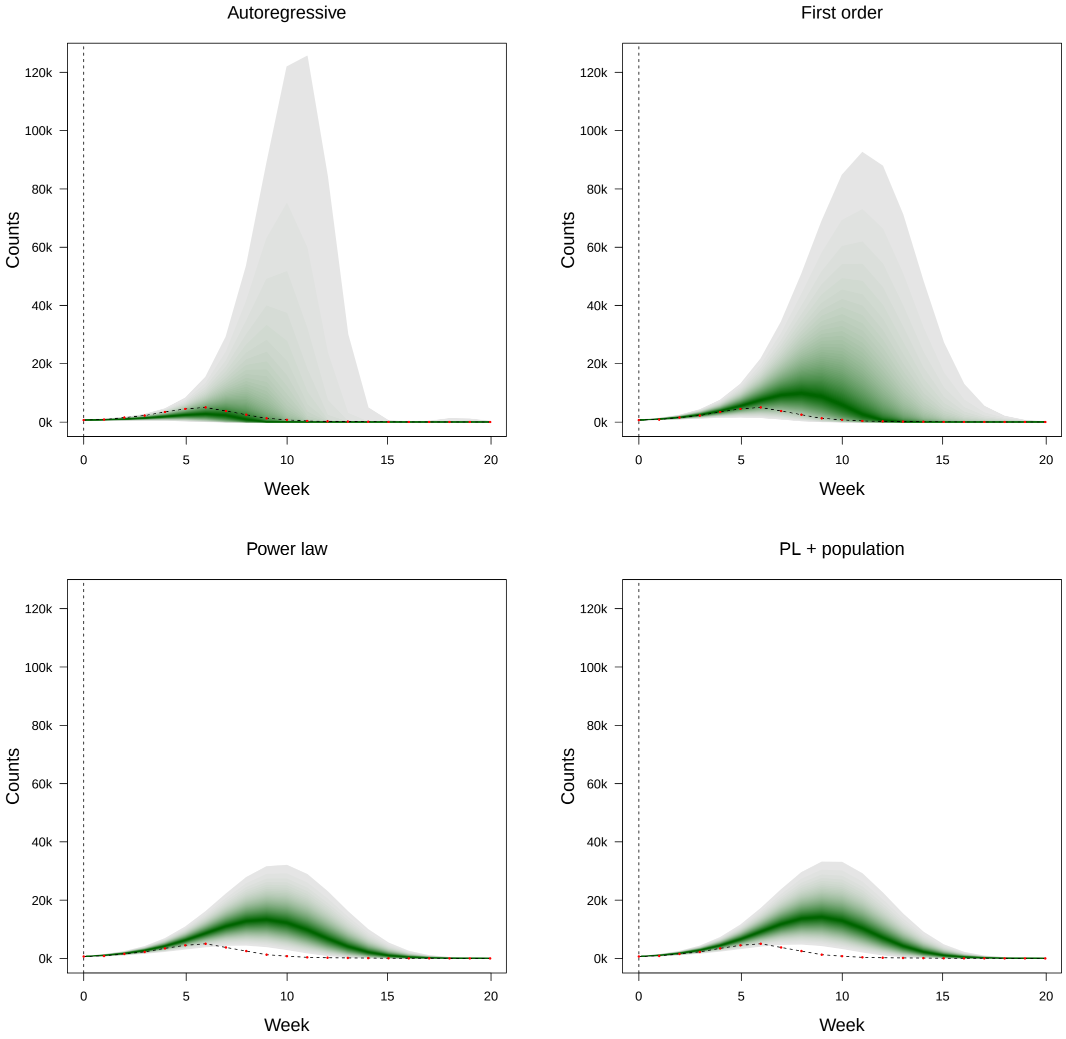

Figure 34: 2017 prediction fan plots for selected models: autoregressive, first order, power, and population models.

5.2 Effect of changes in conditioning week

Overestimation, as seen in the time series of counts in 2017, is also sometimes observed when the conditioning week is changed. Figure 35 shows the effect changing conditioning week has on predictions of the 2008 epidemic season. The black “week 0” time series plots the original prediction conditioning on the last week of 2007. The rest of the predictions condition on weeks 1-6 of 2008. Changing conditioning week has no effect on the endemic model prediction (not shown) and only a relatively small effect on the prediction from the autoregressive model. Conditioning on week 1 or 2 barely changes the autoregressive model prediction at all and the maximum change (434 counts) in the weekly predictions is a result of conditioning on week 5. For the models with a neighborhood component, conditioning on week 4 causes the biggest change in predictions, with peak differences of 1139, 1808, and 1935 counts for the first order, power, and population models. The density and urbanicity model yield similar results to the power and population models.

Changes in conditioning week can also cause predicted counts to decrease. The direction of the change is dependent on whether the conditioning week has higher or lower counts than expected at that particular time. As evident in Figure 35, conditioning on week 1 of 2008 causes the power and population models to predict slightly less than when conditioning on week 0.

As earlier in this section, the results vary by year. In 2011, conditioning on one of the first six weeks of the year results in predictions larger than or equal to those based on the last week of 2010 (Figure 42). The predictions are successively larger until week 5 or 6, when the predictions decrease, but not yet to the level of the original prediction. The increases are greatest for the power models.

In 2012 and 2014 the predictions based on weeks 1 through 6 are, however, always smaller than the original prediction and each prediction is smaller than the last (Figures 43 and 45). In 2012 the predicted peak of the autoregressive model is always larger than the observed peak, but for the other models the peak sinks below the observed. The 2014 season was extremely low, so even with changes in the conditioning week, no model’s predicted peak sinks to the level of the observed peak.

For 2013, 2015, and 2016, predictions from the new conditioning weeks can fall above or below the observed counts (Figures 44, 46, and 47). The long-term predictions based on week 1 are smaller or similar to the observed for 2013 and 2015, but the predictions rise with conditioning week, surpassing the observed. The opposite is observed in 2016, in which the predictions conditioning on week 1 are larger than the observed and predictions conditioning on the following weeks decrease and fall below the observed.

In 2017 the results are model-dependent (Figure 44). The autoregressive predictions increase until conditioning on week 6. The first order, power, and population model predictions for conditioning weeks 1 through 6 all successively decrease and all fall below their original predictions. The scale of the changes reflects the scale of overestimation in this year.

The long-term predictions are conditional on conditioning week $\boldsymbol{Y}_{\cdot,t-1}$ and so the dependence of the predictions on the values of the conditioning week is apparent. However, as seen in Figures 35 and 42 - 48, the conditioning value can have a very large effect on the resulting predictions. Single week changes in conditioning week sometimes lead to vastly different predictions due to differences between the observed and the expected for each week. When predicting on week three of 2008 (when counts start to approach peak

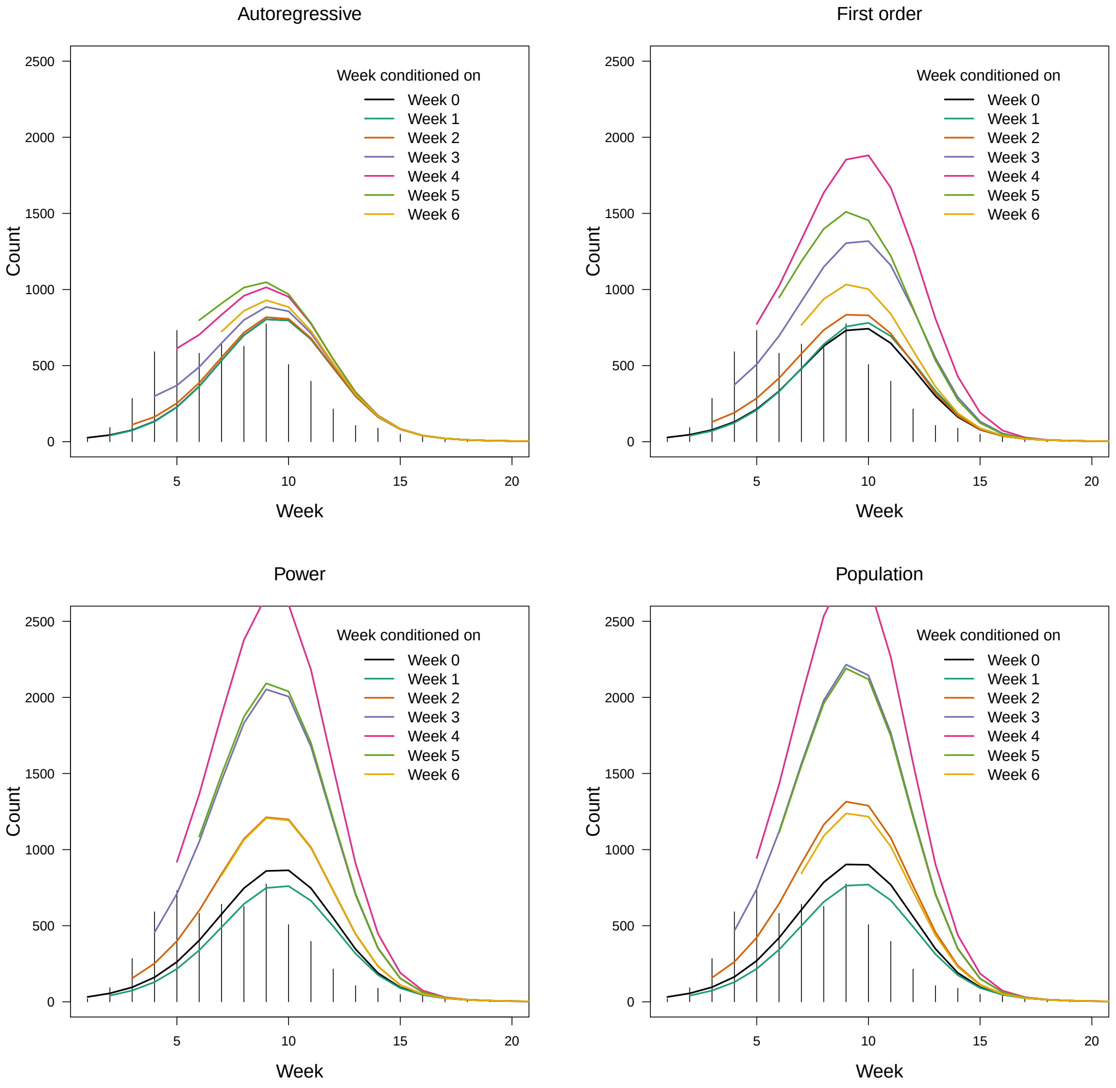

Figure 35: Time series of 2008 influenza season long-term predictions based on various conditioning weeks using the autoregressive (top-left), first order (top-right), power (bottom-left) and population (bottom-right) models.

height), for example, instead of week 52 of the previous year, final count predictions for all models increased, but excessively so for power models specifically.

6 Conclusion

In this analysis, long-term predictions were made using the analytical method outlined in Held et al. (2017). While spatial components have been shown to lead to better fit, their effects on out-of-sample predictions were little explored. Using proper scoring rule methods, performance of various spatio-temporal model configurations were compared to a model without spatial components and to models with simplistic spatial assumptions.

The proposed long-term prediction method within the HHH4 framework was shown to be an easy-to-implement alternative to simulation for obtaining probabilistic predictions of infection incidence in time and space.

For 2008, power law models outperform baseline models when making predictions at the finest data granularity, for the epidemic curve, and in space. Comparing performance between the power law models, and to some extent comparing the results in general, however, was difficult. Some way of deciding whether differences in score are actually relevant could be helpful. In the case of 2008, the novel application of the multivariate Dawid-Sebastiani score to data "slices" corroborated the positive results, yielding the same model rankings as the score applied to the complete raw data and indicating that all power models outperformed the baseline models.

Predictions for the 2008 epidemic season were based only on seven previous epidemic seasons so similar analyses were run on an updated data set with more recent years. Across the years, the power models generally outperform the baseline models at at least 2 of the 4 aggregation levels per year, however, they never perform better at all aggregations in any given year. Only at the final count level do the baseline models outperform the power models more often. The results from one year to the next were quite different. There were no obvious patterns between characteristics of the epidemic season in relation to seasons trained on, e.g., peak timing and peak height, and aggregation levels at which power models performed better.

The extended analysis revealed two problems. First, incomplete reporting became even more evident, as even counts from as early as 2001 were updated since data retrieval in 2008. Reporting seems to have improved since the law was enacted in 2000, but regions with zero counts for entire epidemic seasons still exist in the more recent years. Underreporting can occur at the community level, in which not all infected individuals seek care, or the healthcare-level, in which there is a failure to adequately report cases. Methods to deal with underreporting, such as multiplication factors, are often disease-, country-, and age-specific (Gibbons et al., 2014). Within this dataset, rates of underreporting may even be county-specific. Ongoing work, including inclusion of additional lags (Bracher and Held, 2017), may help address difficulties due to underreporting.

Second, the inflexibility of the model regarding epidemic offset and peak time is evident: all models consistently predict season peaks at week 8 or 9 even though the peak week varies from week 6 to week 11 and only two of the peaks from the 2011 through 2017 epidemic seasons occur within a week of the predicted peak. This inflexibility concerning peak time combined with dependence of predictions on characteristics of the conditioning week can lead all models considering space (including the baseline first order model) to grossly overestimate the epidemic curve, as observed in 2017. Greater flexibility concerning season offset and peak time could reduce the chance of excessively large predictions such as found in 2017 by allowing the model to register that the high conditioning values are a result of an early season, as opposed to early extremely high values indicative of very large peak values. The models considered here only accounted for incidence from the immediately preceding week. Additional lags in the model may also prevent excessive peak counts.

The differences in scores for the various gravity models were difficult to interpret, however, estimated weights for these models were found to be similar, suggesting that the differences in scores may not be of great importance. The proposed gravity models should be applied to different data sets to explore their properties further. Additionally, there may be ways to improve the urbanicity measure for use in southern Germany. We mention how population fraction does not adequately distinguish between the urban character of Fürth and Kelheim, however, urbanicity might also underestimate the urban character of Fürth, which is now considered a part of the Nuremberg metropolitan area. The urban character of Nuremberg may need to be considered when measuring urbanicity of Fürth, as well as Erlangen and Schwabach. Goodall et al. (1998) discuss how measuring urbanicity can be adapted for such large metropolitan areas. Inclusion of other easy-to-obtain data on social structure, however, may result in larger improvements against the power law model alone.

References

Andersson, E., Kühlmann-Berenzon, S., Linde, A., Schióler, L., Rubinova, S. and Frisén, M. (2008). Predictions by early indicators of the time and height of the peaks of yearly influenza outbreaks in Sweden. *Scandinavian Journal of Public Health* **36** 475–482.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2016). *Time series analysis: forecasting and control.* 5th ed. Wiley series in probability and statistics, John Wiley & Sons, Inc.

Bracher, J. and Held, L. (2017). Periodically stationary multivariate non-gaussian autoregressive models. *Preprint* .

Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science* **342** 1337–1342.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006). The scaling laws of human travel. *Nature* **439** 462–465.

Chretien, J.-P., George, D., Shaman, J., Chitale, R. A. and McKenzie, E. (2014). Influenza forecasting in human populations: a scoping review. *PLoS ONE* **9** e94130.

Cowling, B. J., Fang, V. J., Riley, S., Peiris, J. S. M. and Leung, G. M. (2009). Estimation of the serial interval of influenza. *Epidemiology* **20** 344–347.

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

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics* **27** 65–81.

Funk, S., Camacho, A., Kucharski, A. J., Eggo, R. M. and Edmunds, W. J. (2016). Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics* **in press.**

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M. and Kretzschmar, M. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health* **14**.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American St* **106** 746–762.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17**.

Goldstein, E., Cobey, S., Takahashi, S., Miller, J. C. and Lipsitch, M. (2011). Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: A statistical method. *PLoS Medicine* **8** e1001051.

Goodall, C. R., Kafadar, K. and Tukey, J. W. (1998). Computing and using rural versus urban measures in statistical applications. *The American Statistician* **52** 101–111.

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

Held, L., Meyer, S. and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: The thirteenth Armitage lecture. *Statistics in Medicine* .

Höhle, M., Meyer, S. and Paul, M. (2016). *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena.* R package version 1.13.0. URL `https://CRAN.R-project.org/package=surveillance`

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* **22** 679–688.

Johansson, M. A., Reich, N. G., Hota, A., Brownstein, J. S. and Santillana, M. (2016). Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and eason dengue forecasts for Mexico. *Scientific Reports* **6**.

Kafadar, K. and Tukey, J. W. (1993). U.S. cance death rates: a simple adjustment for urbanization. *International Statistical Review* **61** 257–281.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London* **115** 700–721. Series A.

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

Mooney, J. D., Holmes, E. and Christie, P. (2002). Real-time modelling of influenza outbreaks – a linear regression analysis. *Eurosurveillance* **7**.

Nsoesie, E. O., Beckman, R., Marathe, M. and Lewis, B. (2011). Prediction of an epidemic curve: A supervised classification approach. *Statistical Communications in Infectious Diseases* **3**.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* **30** 1118–1136.

Paul, M., Held, L. and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* **27** 6250–6267.

Reich, N. G., Lessler, J., Sakrejda, K., Lauer, S. A., Iamsirithaworn, S. and Cummings, D. A. T. (2016). Case study in evaluating time series prediction vb model using relative mean absolute error. *The American Statistician* **70** 285–292.

Ross, R. (1911). *The prevention of malaria.* John Murray, London.

Sebastiani, P., Mandl, K. D., Szolivits, P., Kohane, I. S. and Ramoni, M. F. (2006). A bayesian dynamic model for influenza surveillance. *Statistics in Medicine* **25** 1803–1816.

Siettos, C. I. and Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence* **4** 295–306.

Sočan, M., Erčulj, V. and Lajovic, J. (2012). Early detection of influenza-like illness through medication sales. *Central European Journal of Public Health* **20** 156–162.

Statistical Office of the European Union (2016). Statistics on commuting patterns at regional level.
URL `http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_commuting_patterns_at_regional_level`

Wei, W. and Held, L. (2014). Calibration tests for count data. *Test* **23** 787–805.

Winkler, R. L., noz, J. M., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M. and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test* **5** 1–60.

World Health Organization (2016). Influenza (seasonal) fact sheet. http://www.who.int/mediacentre/factsheets/fs211/en/.

Xia, Y., Bjørnstad, O. N. and Grenfell, B. T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist* **164**.

Yang, W., Olson, D. R. and Shaman, J. (2016). Foforecasts influenza outbreaks in boroughs and neighborhoods of New York City. *PLoS Computational Biology* **12** e1005201.

7 Appendix

7.1 Fitting the models

The following code was used for fitting the models using the `fluBYBW` data. It draws largely from the code of Michaela Paul and Sebastian Meyer, found in Paul and Held (2011) and Meyer and Held (2014). The first lines set up the basic formulas for the endemic, epidemic, and neighborhood components. This includes random effects for the different regions and seasonality represented by the harmonic waves with frequency $2\pi/52$. The start values are taken by the model fits of Paul and Held (2011).

```
### basic formulas for the most complex model used in Held & Paul (2012)
f.S0 <- ~ -1 + ri(type="iid", corr="all")
f.S1 <- addSeason2formula(f = f.S0, S=1, period=52)
f.end.S3 <- addSeason2formula(
  f = ~ -1 + ri(type="iid", corr="all") + I((t-208)/100),
  S=3, period=52)

## start values for fixed and variance parameters taken from Paul's
## model fit (normalized first-order weights)
START <- list(fixed = c(0.8, 1.7, -2.4,                          # ar
                        0.8, 2.1, -3.1,                          # ne
                        0.5, 2.1, 2.2, 0.4, -0.6, 0.1, -0.1, 0.3,# endemic
                        0.1),                             # -log(overdisp)
              sd.corr = c(-1, 0, -0.3, 0, 0.1, 0.6))
START$random <- {
  set.seed(1)
  rnorm(3*ncol(fluBYBW), sd=rep(exp(START$sd.corr[1:3]), each=ncol(fluBYBW)))
}
init.logd <- log(1.6)  # as found by Brockmann et al (2006)
start.fixed.pl <- append(START$fixed, init.logd, after=length(START$fixed)-1)
```

This next section defines the models. First, a general control object is created which defines the formulas to be used for each model component, the type of model to fit (here, negative-binomial with a universal overdispersion parameter), and the data to use for fitting. The other parts help to ensure convergence. Following this, each individual model is specified. At the end, all models are fit together in parallel.

```
## basic control object
CONTROL <- list(ar = list(f=f.S1),
                ne = list(f=f.S1),
                end = list(f=f.end.S3, offset=population(fluBYBW)),
                family = "NegBin1",
                subset = which(year(fluBYBW) %in% 2001:2007)[-1],
                optimizer = list(stop = list(tol = 1e-05, niter = 100),
                                 regression = list(method = "nlminb"),
                                 variance = list(method = "Nelder-Mead")),
                start = START, verbose = FALSE )

### control objects for the models: norm/raw x first-order/power-law
controls <- list(
# normalized first order model
  norm1 = modifyList(CONTROL,list(
    ne = list(normalize=TRUE)
  )),
# normalized power law model
  normPL = modifyList(CONTROL, list(
    ne=list(weights = W_powerlaw(maxlag=MAXLAG, normalize=TRUE, log=TRUE
                                 , initial=init.logd)) ,
    start = list(fixed=start.fixed.pl)
  ))
)

# power law + population model
controls$PL.nePop <- modifyList(controls$normPL, list(
  ne = list(f=update(controls$normPL$ne$f, ~ . + log(pop))),
  data = list(t=epoch(fluBYBW)-1, pop=population(fluBYBW))
  , start = list(fixed=append(controls$normPL$start$fixed, 1, after=5))
))
# power law + density model
controls$PL.neDens <- modifyList(controls$normPL, list(
  ne = list(f=update(controls$normPL$ne$f, ~ . + log(PopDens))),
  data = list(t=epoch(fluBYBW)-1,
    PopDens= population(fluBYBW)/do.call(rbind,
      replicate(416, t(as.matrix(fluBYBW@map@data$Gebietsflaeche)),
                simplify=FALSE)))
  , start = list(fixed=append(controls$normPL$start$fixed, 1, after=5))
))
# power law + urbanicity model
controls$PL.neUrb <- modifyList(controls$normPL, list(
  ne = list(f=update(controls$normPL$ne$f, ~ . + log(Urb))),
  data = list(t=epoch(fluBYBW)-1,
              Urb=do.call(rbind,
                          replicate(416, t(as.matrix(fluBYBW@map@data$max)),
                                           simplify=FALSE)))
  , start = list(fixed=append(controls$normPL$start$fixed, 1, after=5))
))
# endemic model
```

```r
controls$endemic <- modifyList(CONTROL, list(
  ar = NULL, ne = NULL
  , start = list(fixed=START$fixed[-(1:6)],
                 random=matrix(START$random,ncol=3)[,3],
                 sd.corr=START$sd.corr[3])
))
# endemic+autoregressive model
controls$noNE <- modifyList(CONTROL, list(
  ne = NULL
  , start = list(fixed=START$fixed[-(4:6)],
                 random=c(matrix(START$random,ncol=3)[,c(1,3)]),
                 sd.corr=START$sd.corr[c(1,3,5)])
))

### Do all the fits in parallel
  fit <- function (control) hhh4(fluBYBW, control)
  flufits.RE <- mclapply(controls, fit, mc.preschedule=FALSE, mc.cores=4)
```

7.2 Incidence maps for each year from 2001-2017

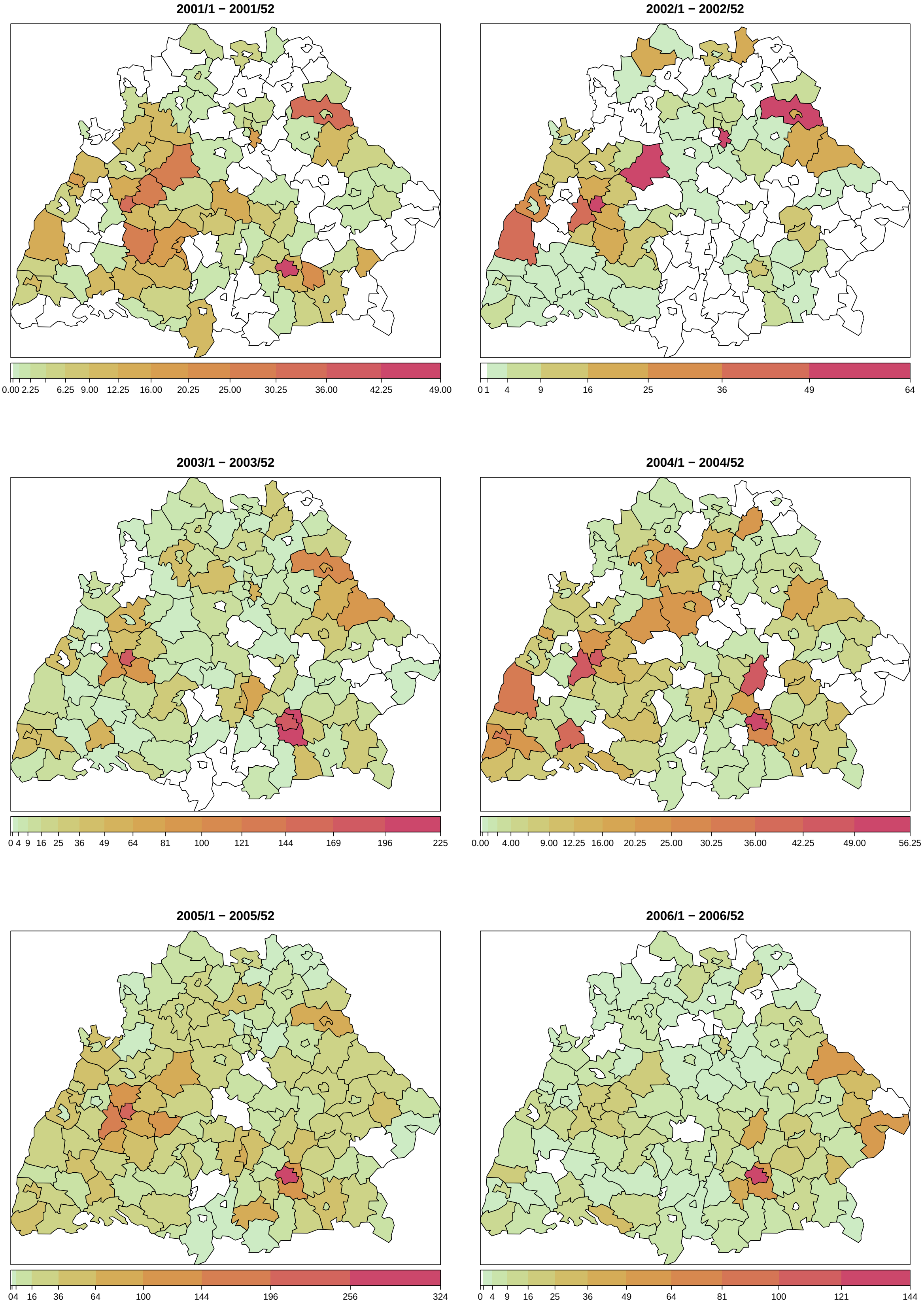

Figure 36: Total incidence per county per year (2001-2006)

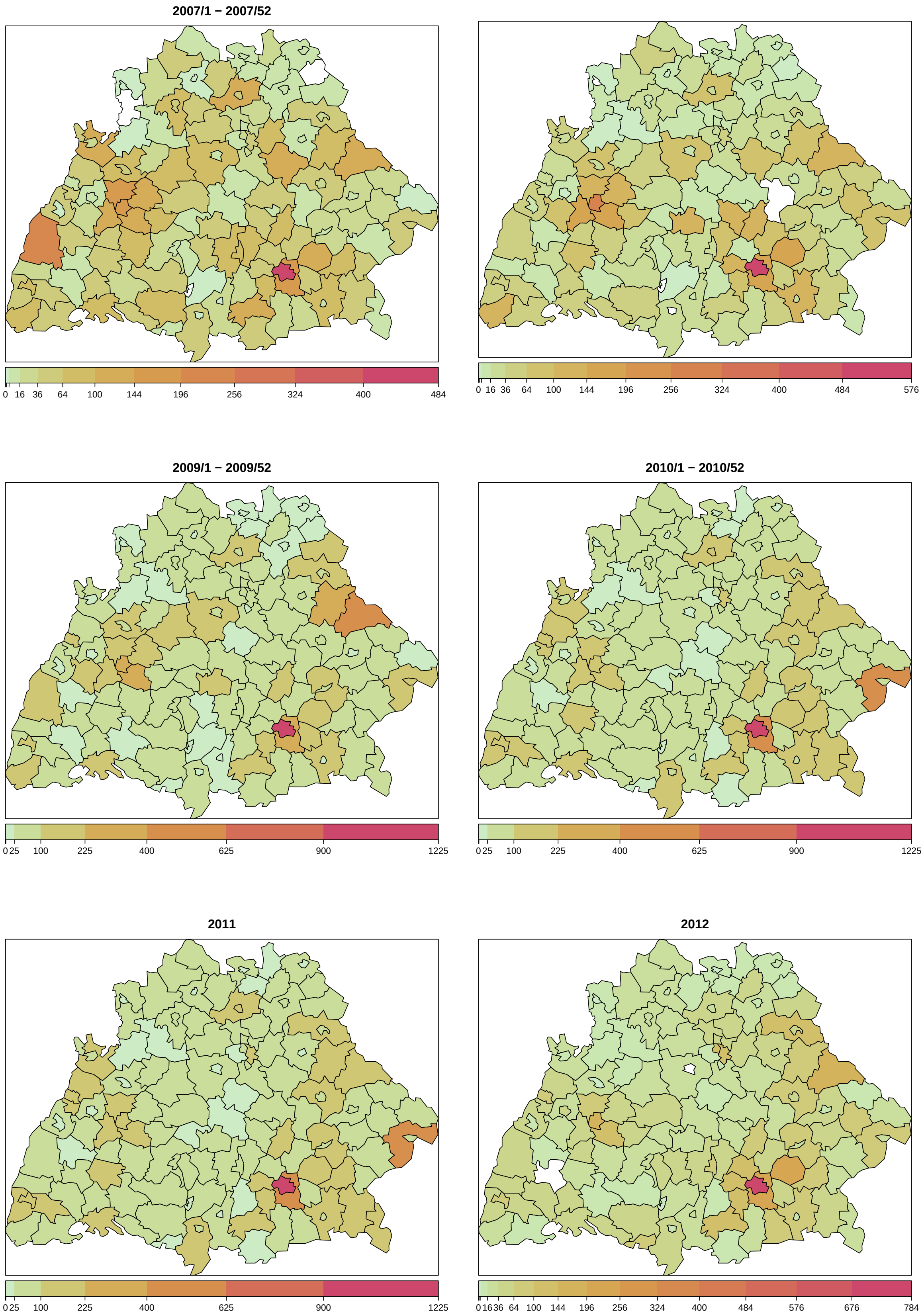

Figure 37: Total incidence per county per year (2007-2012)

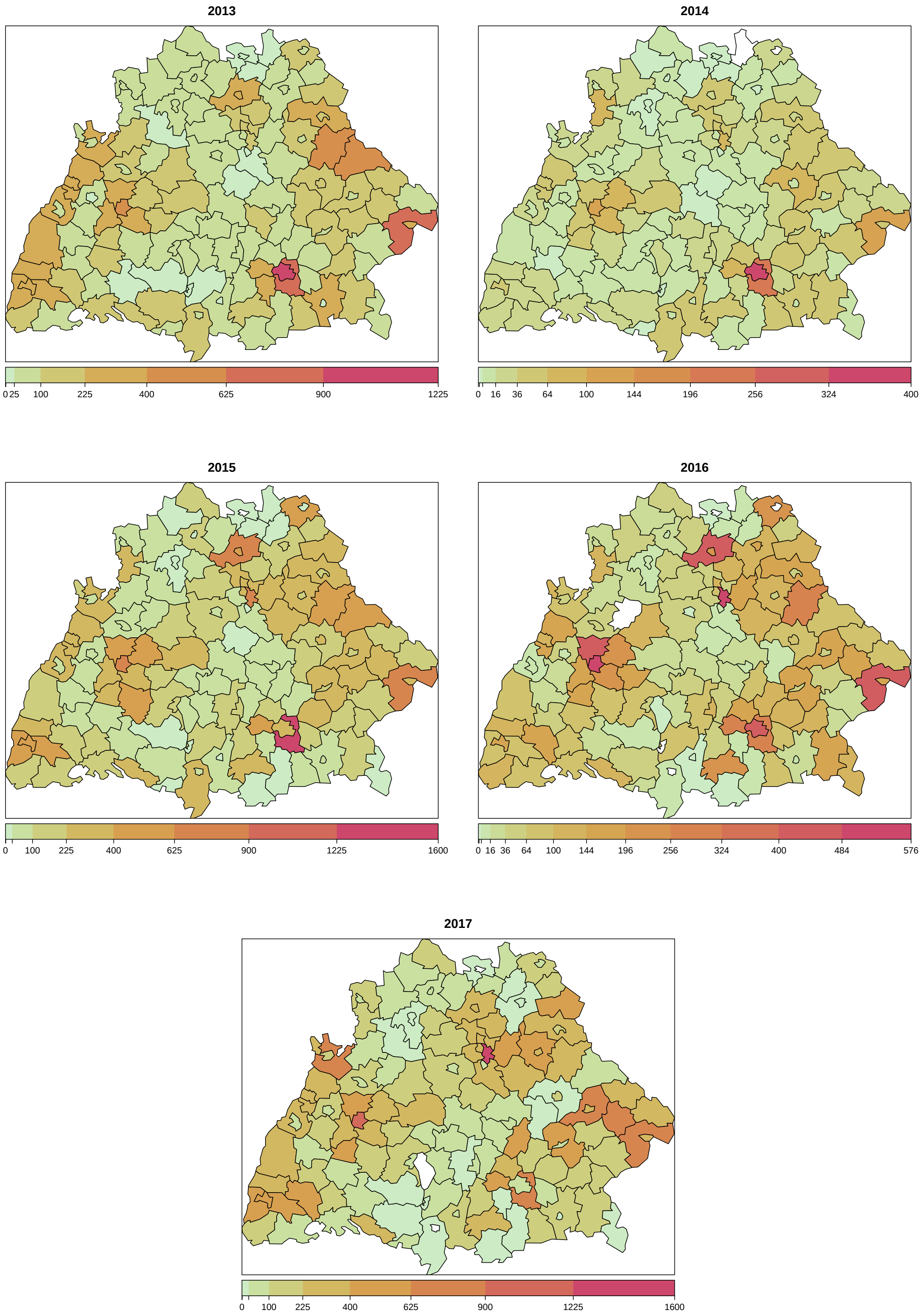

Figure 38: Total incidence per county per year (2013-2017)

7.3 Fanplots for models predicting 2008 incidence

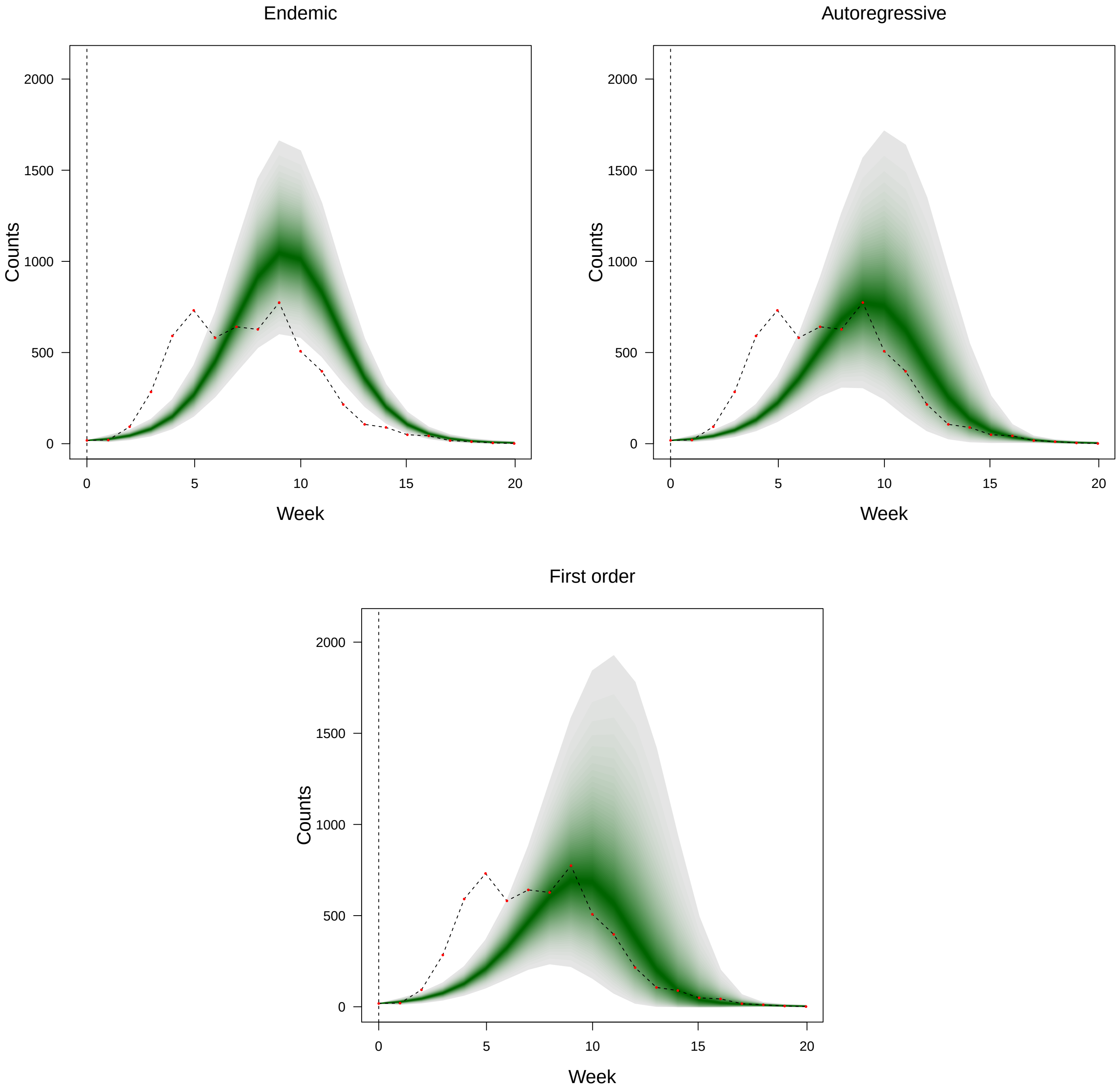

Figure 39: Prediction fan plots for baseline models: endemic, autoregressive, and first order models.

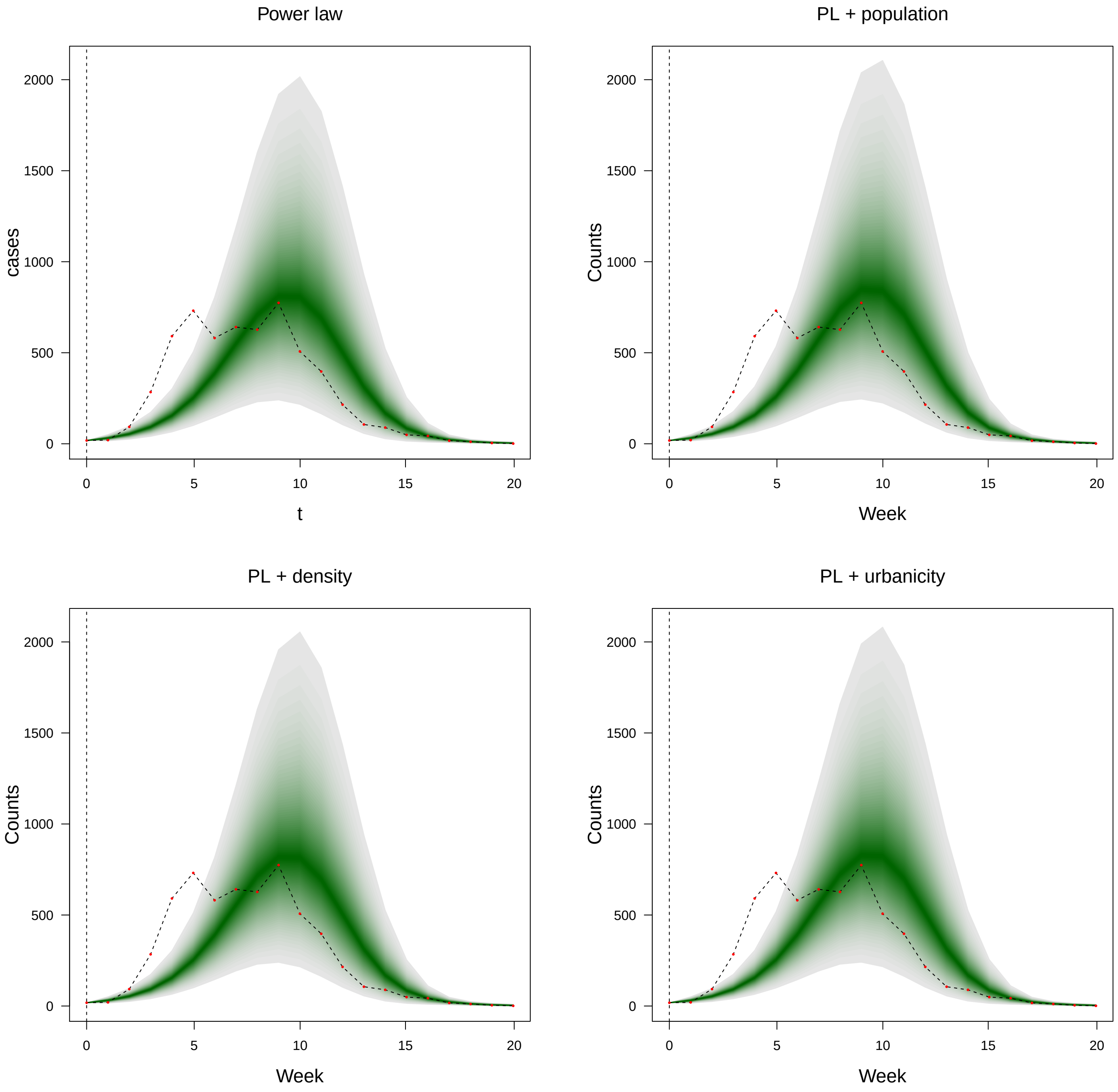

Figure 40: Prediction fan plots for power models: power, population, density, and urbanicity models.

7.4 Maps of predicted incidence in 2008 for each model

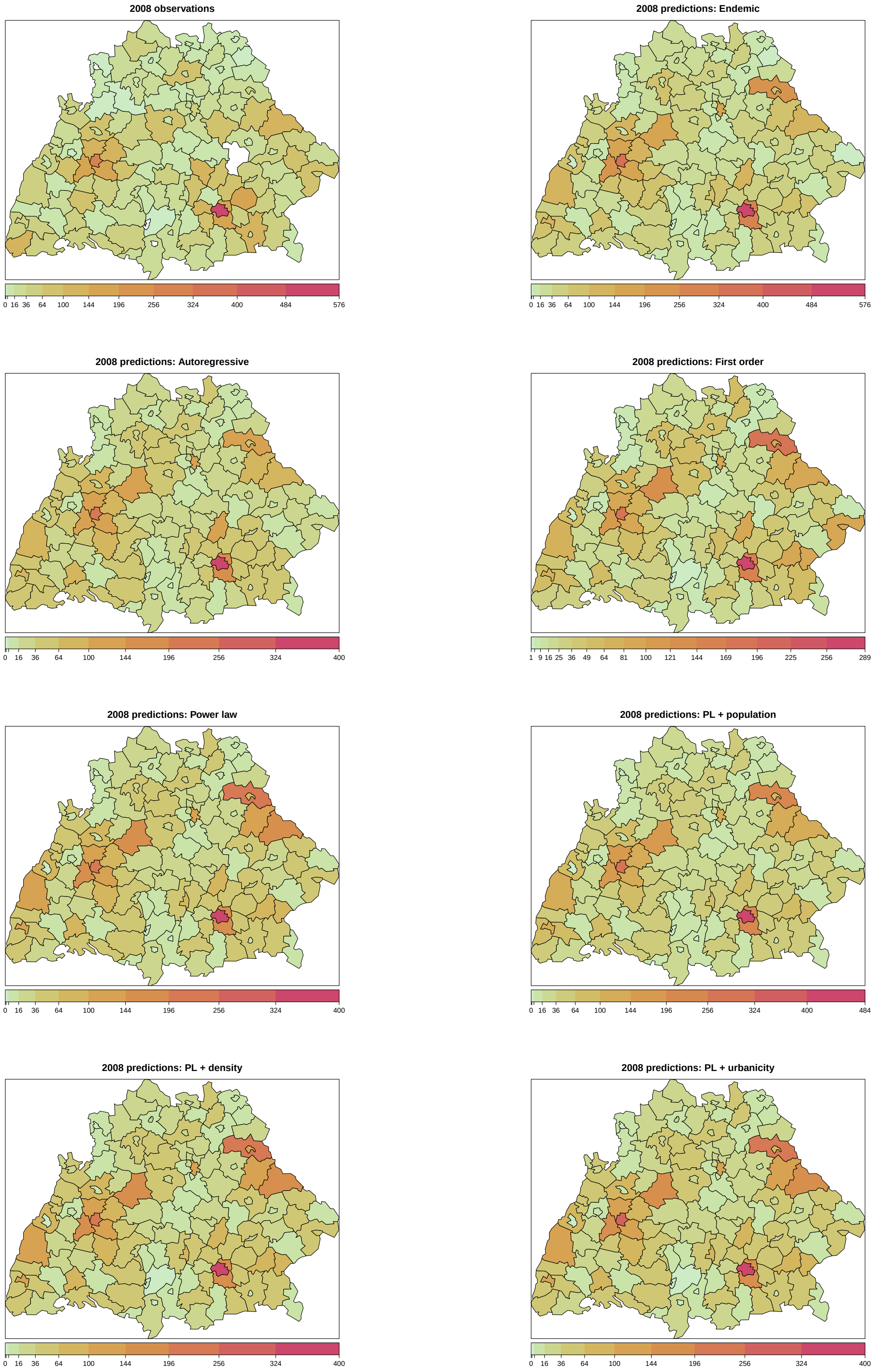

Figure 41: Observed counts (top left) and predicted counts by model for 2008 epidemic season.

7.5 Long-term predictions based on various conditioning weeks for selected models.

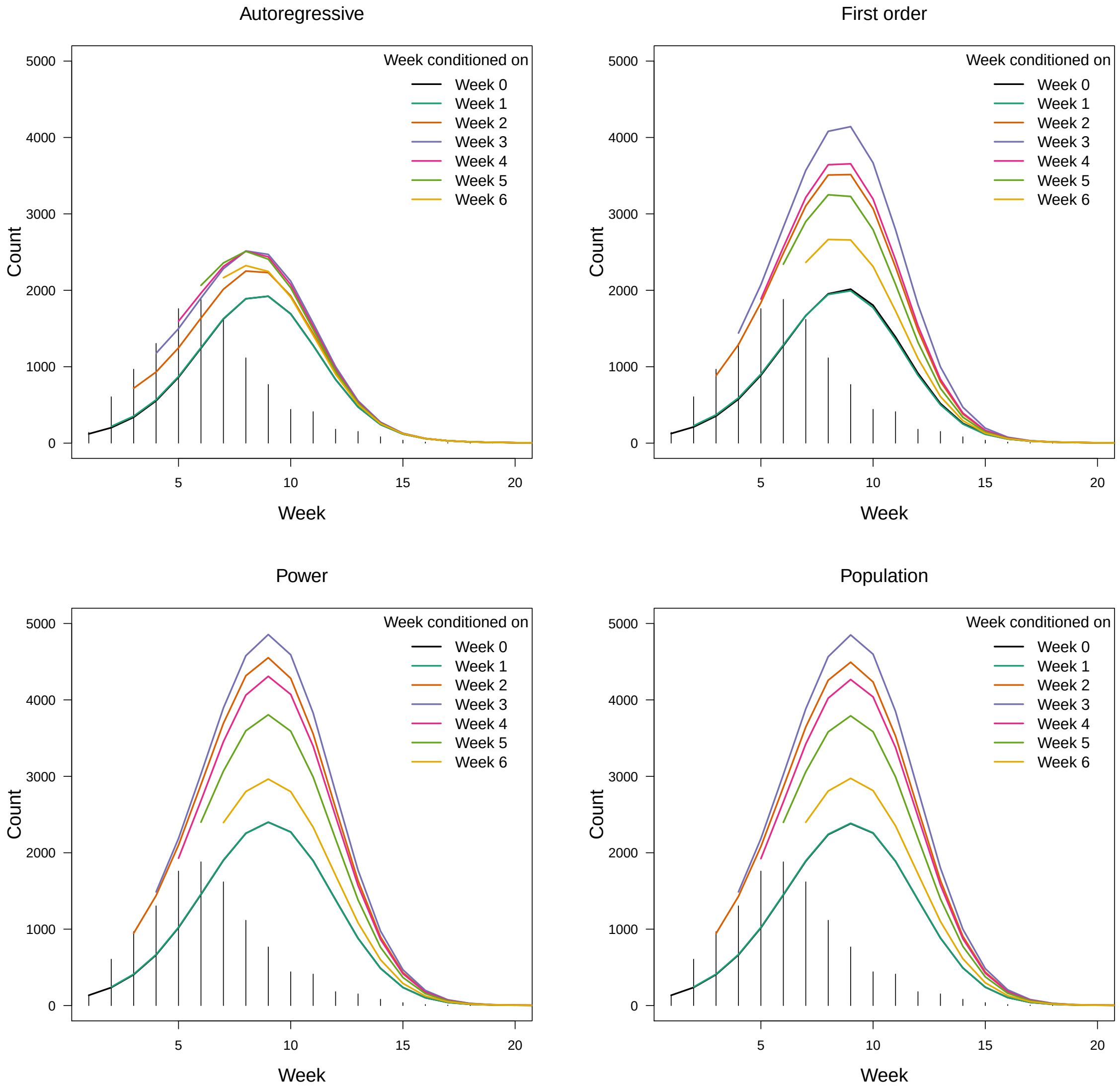

Figure 42: 2011 long-term predictions based on various conditioning weeks for selected models.

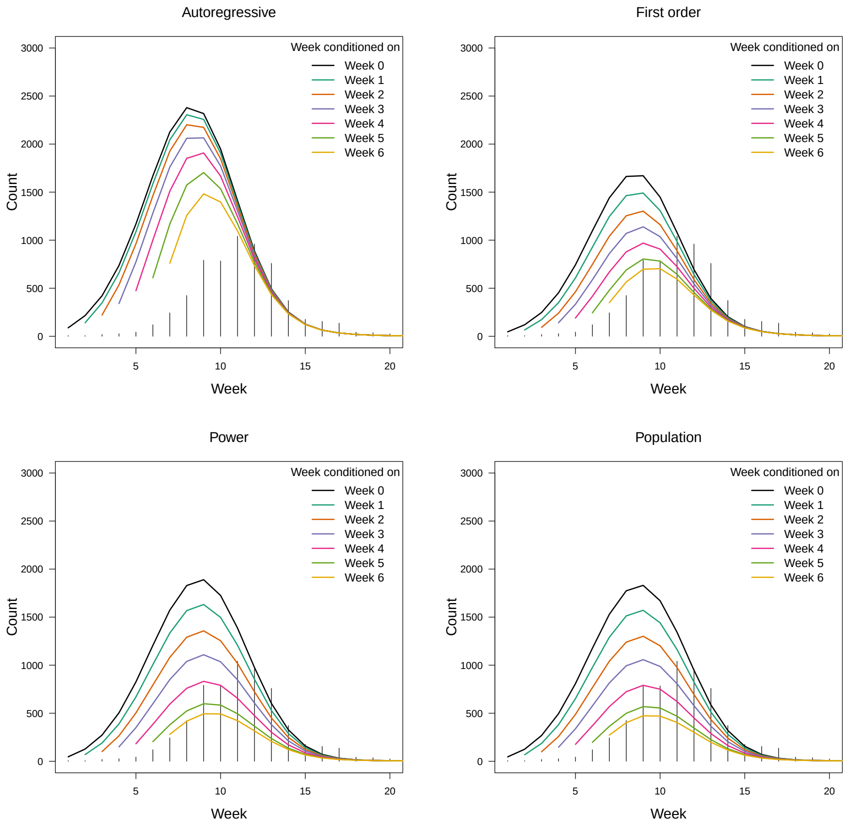

Figure 43: 2012 long-term predictions based on various conditioning weeks for selected models.

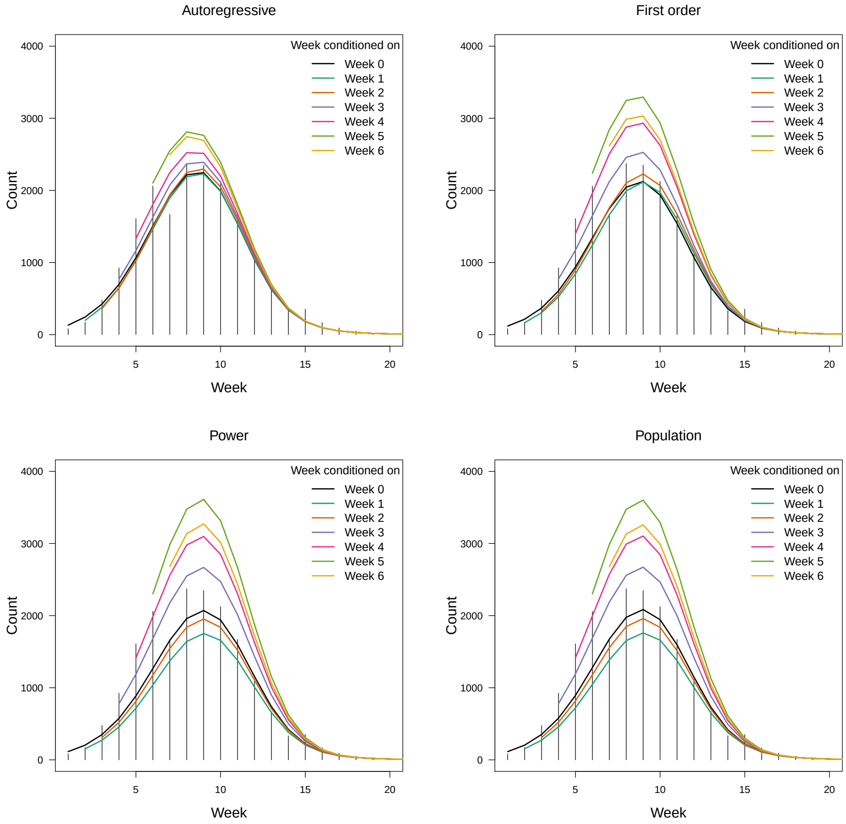

Figure 44: 2013 long-term predictions based on various conditioning weeks for selected models.

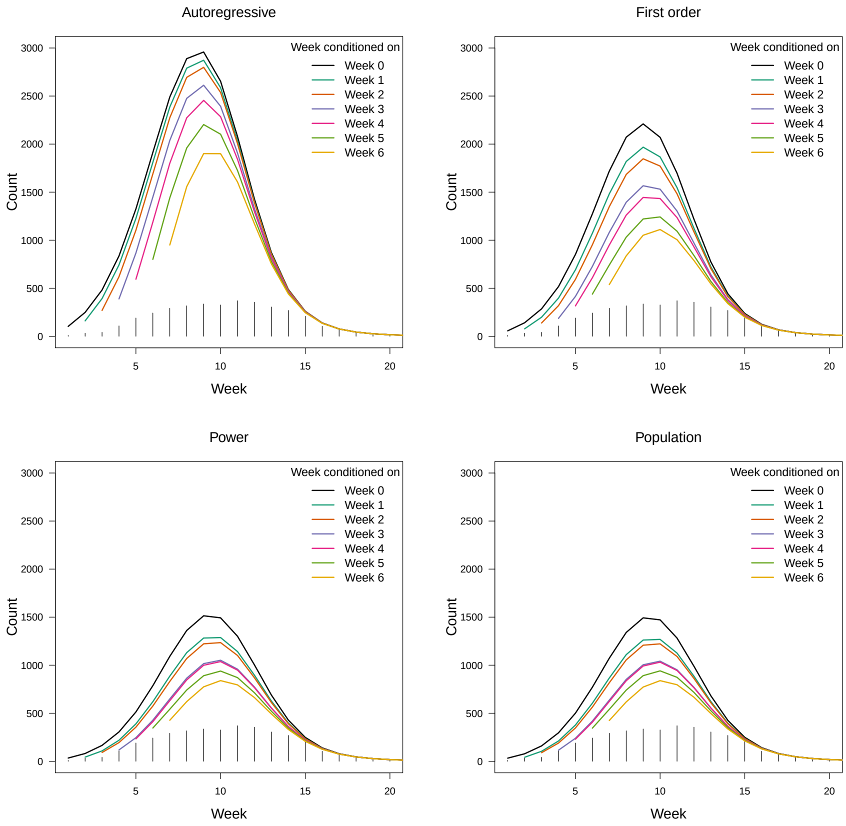

Figure 45: 2014 long-term predictions based on various conditioning weeks for selected models.

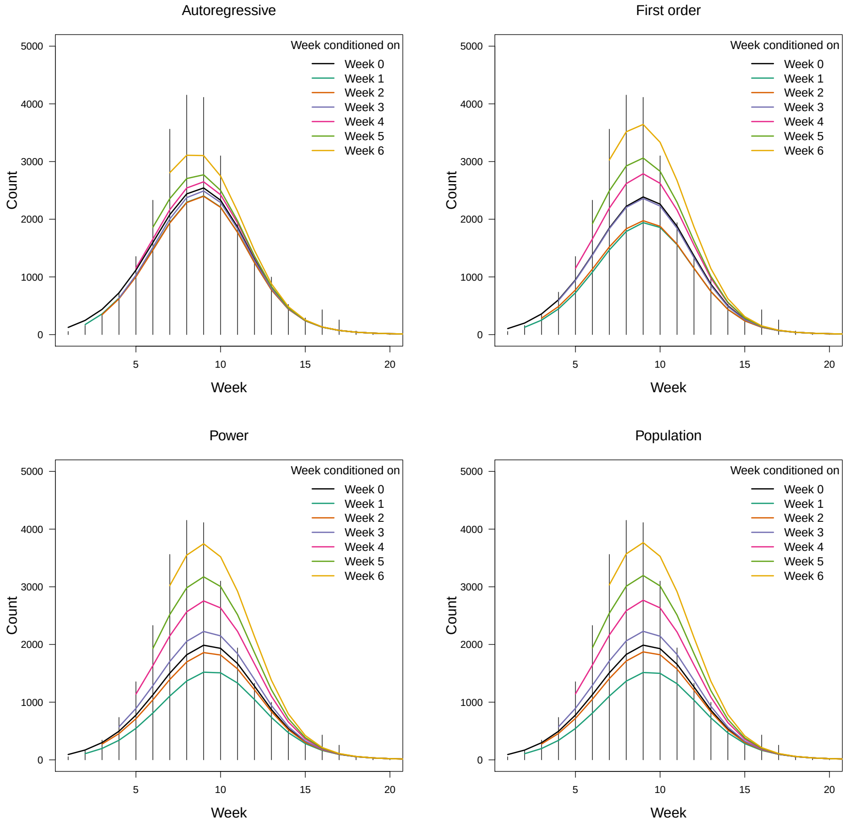

Figure 46: 2015 long-term predictions based on various conditioning weeks for selected models.

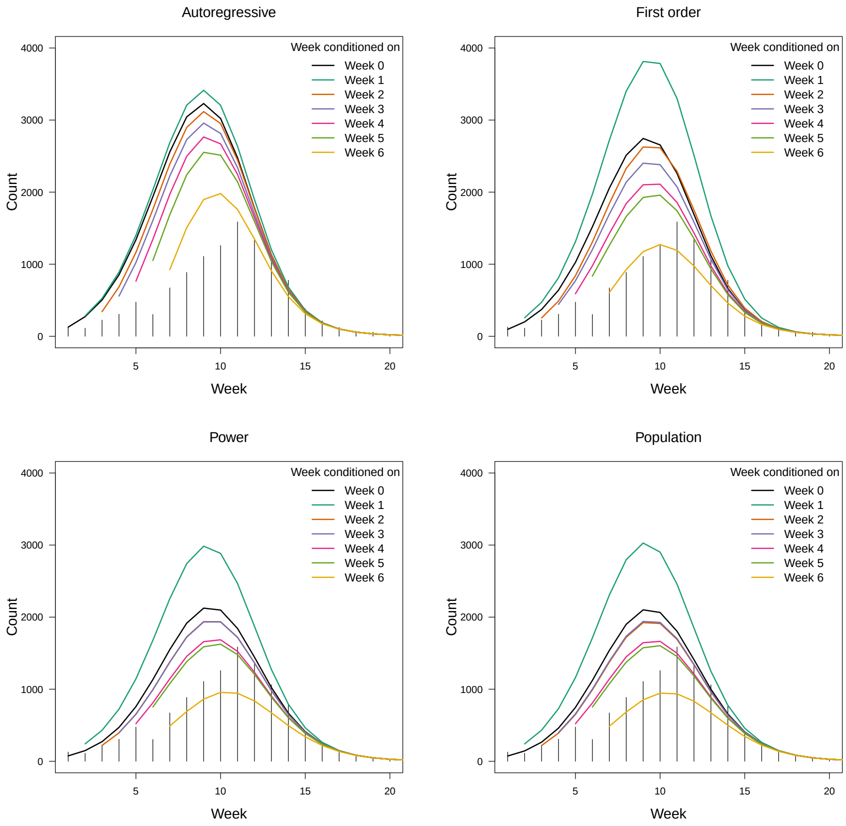

Figure 47: 2016 long-term predictions based on various conditioning weeks for selected models.

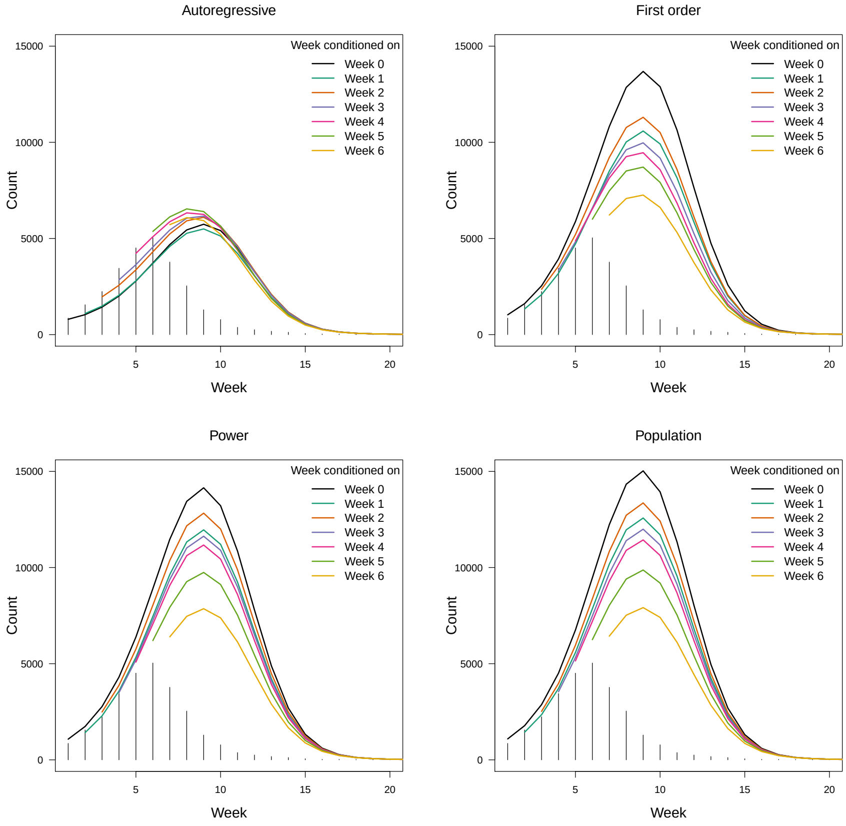

Figure 48: 2017 long-term predictions based on various conditioning weeks for selected models.

7.6 Additional tables

2011							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	1.6696	1.6677	1.6271	1.5941	1.5708	1.5384	1.5913
Regional	1.6696	1.5135	1.5102	1.5419	1.5233	1.4796	1.5391
Univariate	3.3392	3.3355	3.3925	3.4297	3.3798	3.3163	3.4241
2012							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	2.7323	3.0511	3.4802	3.0936	3.0723	4.0871	3.1218
Regional	2.7323	2.8909	3.3946	3.0513	3.0343	4.0475	3.0819
Univariate	5.4647	6.1022	7.0718	6.3460	6.3144	8.4918	6.4131
2013							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	1.7369	1.9205	1.7555	1.6945	1.6817	1.6953	1.6928
Regional	1.7369	1.7711	1.6481	1.5940	1.5849	1.5946	1.5928
Univariate	3.4738	3.8410	3.5795	3.4564	3.4300	3.4575	3.4517
2014							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	1.8466	1.8458	1.4169	1.1623	1.1614	1.1646	1.1624
Regional	1.8466	1.6496	1.3315	1.1691	1.1716	1.1713	1.1721
Univariate	3.6932	3.6916	2.9547	2.4934	2.4874	2.4978	2.4934
2015							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	3.6607	4.3675	4.6005	5.0029	4.8855	4.9910	4.9591
Regional	3.6607	4.0271	4.3684	4.6833	4.5665	4.6714	4.6399
Univariate	7.3214	8.7350	9.1391	10.3854	10.1553	10.3625	10.2997
2016							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	2.0276	2.2017	1.9490	1.7466	1.7389	1.7472	1.7464
Regional	2.0276	2.0000	1.8329	1.6348	1.6263	1.6358	1.6354
Univariate	4.0551	4.4034	3.9780	3.5696	3.5522	3.5709	3.5694
2017							
	Endemic	End + AR	First order	Power	PL + pop	PL + dens	PL + urb
Weekly	2.5498	2.8437	2.7224	2.6820	2.6974	2.6837	2.6857
Regional	2.5498	2.5350	2.5531	2.5952	2.6111	2.5967	2.5986
Univariate	5.0995	5.6875	5.6048	5.6683	5.7011	5.6713	5.6751

Table 20: Mean weekly and mean regional mDSS and mean univariate DSS for the epidemic seasons of 2011 to 2017.

	Space-time	rank	Time	rank	Space	rank	Final	rank
				2011				
Endemic	3.3392	3	24.9225	7	10.4206	7	27.3248	7
End + AR	3.3355	2	16.9487	6	9.5367	6	17.9952	1
First order	3.3925	5	15.5331	5	9.4539	5	18.3721	3
Power	3.4297	7	14.2307	1	9.3566	2	18.7633	6
PL + pop	3.3798	4	14.6004	3	9.3217	1	18.6406	4
PL + dens	3.3163	1	14.6592	4	9.4207	4	18.2077	2
PL + urb	3.4241	6	14.4436	2	9.3661	3	18.6957	5
				2012				
	Space-time	rank	Time	rank	Space	rank	Final	rank
Endemic	5.4647	1	23.0400	4	9.8966	7	125.1489	7
End + AR	6.1022	2	26.0770	5	9.5720	6	20.6697	6
First order	7.0718	6	35.1202	6	8.5685	3	19.0442	2
Power	6.3460	4	21.8731	1	8.6060	4	19.1189	4
PL + pop	6.3144	3	22.3579	2	8.5551	1	19.2009	5
PL + dens	8.4918	7	43.0619	7	9.0187	5	15.0844	1
PL + urb	6.4131	5	22.4614	3	8.5586	2	19.0972	3
				2013				
Endemic	3.4738	5	10.4043	1	10.3472	2	17.8281	6
End + AR	3.8410	7	11.6970	7	10.2053	1	18.5024	7
First order	3.5795	6	11.4783	6	10.5375	3	17.5361	5
Power	3.4564	3	11.1969	5	11.1192	6	17.3324	4
PL + pop	3.4300	1	11.1730	2	10.8829	4	17.2820	1
PL + dens	3.4575	4	11.1933	4	11.1224	7	17.3223	3
PL + urb	3.4517	2	11.1879	3	11.0803	5	17.3044	2
				2014				
	Space-time	rank	Time	rank	Space	rank	Final	rank
Endemic	3.6932	7	23.5667	7	12.1411	7	211.6217	7
End + AR	3.6916	6	18.5190	6	10.3878	6	23.3989	1
First order	2.9547	5	17.0628	5	9.4207	5	26.9003	6
Power	2.4934	2	13.3782	1	8.7764	3	23.6774	2
PL + pop	2.4874	1	13.4456	3	8.7468	1	24.3537	5
PL + dens	2.4978	4	13.4050	2	8.7779	4	23.7915	3
PL + urb	2.4934	3	13.4635	4	8.7756	2	24.0506	4

Table 21: Univariate Dawid-Sebastiani scores and ranks for each combination of model, aggregation level, and year from 2011 to 2014.

	Space-time	rank	Time	rank	Space	rank	Final	rank
2015								
	Space-time	rank	Time	rank	Space	rank	Final	rank
Endemic	7.3214	1	14.4921	3	13.2583	3	15.3243	1
End + AR	8.7350	2	12.7117	1	10.7864	1	19.3663	3
First order	9.1391	3	13.5178	2	11.2525	2	19.1356	2
Power	10.3854	7	16.6709	4	17.3160	5	22.1436	4
PL + pop	10.1553	4	17.2113	7	17.3339	7	22.5425	7
PL + dens	10.3625	6	16.7483	5	17.3149	4	22.1753	5
PL + urb	10.2997	5	16.9074	6	17.3223	6	22.2956	6
2016								
Endemic	4.0551	6	17.4369	7	11.0676	7	124.8898	7
End + AR	4.4034	7	14.4324	6	10.8434	6	19.8791	6
First order	3.9780	5	13.4617	5	9.5897	5	19.5793	5
Power	3.5696	3	11.9537	2	8.9672	2	18.2558	2
PL + pop	3.5522	1	11.9267	1	8.9415	1	18.1731	1
PL + dens	3.5709	4	11.9548	4	8.9685	4	18.2726	3
PL + urb	3.5694	2	11.9538	3	8.9674	3	18.2839	4
2017								
Endemic	5.0995	1	32.3664	7	11.5390	1	31.2401	7
End + AR	5.6875	6	16.7076	6	12.1987	2	23.8355	2
First order	5.6048	2	15.7911	1	12.3464	3	23.6265	1
Power	5.6683	3	15.8920	2	12.9879	4	25.3125	3
PL + pop	5.7011	7	16.2045	5	13.1240	7	25.7572	6
PL + dens	5.6713	4	15.9217	3	13.0002	5	25.3669	4
PL + urb	5.6751	5	15.9753	4	13.0218	6	25.4596	5

Table 22: Univariate Dawid-Sebastiani scores and ranks for each combination of model, aggregation level, and year from 2015 to 2017.

	Prediction week										
	1	2	3	4	5	6	7	8	9	10	11
					Power						
log Index	-0.7321	0.6039	2.2961	4.2597	3.3166	1.8050	1.8075	1.9749	1.9514	2.0063	1.8539
Index	-0.7319	0.6039	2.2979	4.2628	3.3227	1.8056	1.8078	1.9751	1.9516	2.0065	1.8543
Inverse	-0.7322	0.6046	2.2950	4.2576	3.3123	1.8046	1.8075	1.9748	1.9514	2.0062	1.8537
Binary	-0.7324	0.6063	2.2943	4.2562	3.3079	1.8044	1.8075	1.9744	1.9514	2.0061	1.8536
NE log Index	-0.7324	0.6090	2.2933	4.2549	3.3052	1.8051	1.8081	1.9745	1.9525	2.0067	1.8543
NE Index	-0.7322	0.6091	2.2951	4.2586	3.3069	1.8066	1.8082	1.9745	1.9525	2.0070	1.8549
NE Inverse	-0.7325	0.6088	2.2914	4.2516	3.3032	1.8047	1.8079	1.9747	1.9525	2.0065	1.8537
NE Binary	-0.7327	0.6087	2.2883	4.2467	3.2995	1.8062	1.8073	1.9753	1.9524	2.0066	1.8528
No BE	-0.7322	0.6090	2.2947	4.2574	3.3055	1.8048	1.8080	1.9741	1.9521	2.0064	1.8542
					Power + population						
log Index	-0.7303	0.6246	2.3557	4.3649	3.3853	1.8210	1.7883	1.9697	1.9490	2.0046	1.8765
Index	-0.7303	0.6245	2.3556	4.3639	3.3902	1.8213	1.7885	1.9696	1.9489	2.0043	1.8766
Inverse	-0.7304	0.6255	2.3567	4.3664	3.3823	1.8208	1.7884	1.9697	1.9494	2.0048	1.8767
Binary	-0.7305	0.6275	2.3598	4.3704	3.3802	1.8207	1.7886	1.9695	1.9500	2.0051	1.8770
NE log Index1	-0.7306	0.6292	2.3597	4.3685	3.3789	1.8216	1.7891	1.9699	1.9510	2.0059	1.8775
NE Index	-0.7306	0.6292	2.3603	4.3692	3.3798	1.8212	1.7892	1.9698	1.9511	2.0058	1.8776
NE Inverse	-0.7306	0.6291	2.3590	4.3677	3.3782	1.8220	1.7890	1.9700	1.9510	2.0059	1.8774
NE Binary	-0.7306	0.6289	2.3575	4.3658	3.3765	1.8230	1.7888	1.9702	1.9508	2.0059	1.8770
No BE	-0.7304	0.6293	2.3629	4.3734	3.3793	1.8209	1.7888	1.9695	1.9508	2.0057	1.8778
					Power + density						
log Index	-0.7358	0.6093	2.3148	4.2905	3.3236	1.7947	1.8125	1.9758	1.9582	2.0158	1.8611
Index	-0.7358	0.6095	2.3171	4.2944	3.3299	1.7951	1.8128	1.9758	1.9584	2.0159	1.8614
Inverse	-0.7359	0.6099	2.3138	4.2891	3.3196	1.7945	1.8125	1.9755	1.9581	2.0157	1.8609
Binary	-0.7359	0.6114	2.3127	4.2873	3.3153	1.7943	1.8127	1.9752	1.9582	2.0156	1.8608
NE log Index	-0.7372	0.6135	2.3042	4.2706	3.3052	1.7852	1.8139	1.9758	1.9609	2.0166	1.8596
NE Index	-0.7370	0.6141	2.3069	4.2747	3.3089	1.7844	1.8144	1.9751	1.9608	2.0161	1.8600
NE Inverse	-0.7373	0.6132	2.3023	4.2685	3.3022	1.7872	1.8132	1.9763	1.9608	2.0169	1.8592
NE Binary	-0.7373	0.6134	2.3006	4.2677	3.2987	1.7932	1.8123	1.9770	1.9601	2.0174	1.8587
No BE	-0.7356	0.6141	2.3134	4.2891	3.3133	1.7944	1.8134	1.9750	1.9590	2.0159	1.8617
					Power + urbanicity						
log Index	-0.7347	0.6197	2.3452	4.3439	3.3366	1.8041	1.8103	1.9760	1.9606	2.0171	1.8667
Index	-0.7348	0.6202	2.3478	4.3470	3.3424	1.8043	1.8104	1.9761	1.9608	2.0172	1.8671
Inverse	-0.7347	0.6201	2.3438	4.3421	3.3326	1.8040	1.8104	1.9759	1.9606	2.0170	1.8666
Binary	-0.7346	0.6216	2.3430	4.3412	3.3290	1.8038	1.8107	1.9757	1.9608	2.0170	1.8665
NE log Index	-0.7359	0.6242	2.3331	4.3217	3.3172	1.7933	1.8111	1.9765	1.9639	2.0173	1.8643
NE Index	-0.7359	0.6251	2.3368	4.3258	3.3220	1.7898	1.8120	1.9757	1.9643	2.0168	1.8648
NE Inverse	-0.7358	0.6236	2.3308	4.3195	3.3138	1.7973	1.8103	1.9771	1.9634	2.0177	1.8640
NE Binary	-0.7356	0.6236	2.3291	4.3195	3.3110	1.8052	1.8094	1.9778	1.9623	2.0182	1.8636
No BE	-0.7343	0.6240	2.3443	4.3438	3.3279	1.8037	1.8114	1.9755	1.9617	2.0174	1.8674

Table 23: Weekly mDSS, mean and rank for power models with border effects.

	Prediction week										
	12	13	14	15	16	17	18	19	20	Mean	rank
					Power						
log Index	1.5234	1.2885	0.6992	0.2053	-0.0170	-0.7165	-0.2815	-1.2762	-1.7570	1.0406	8
Index	1.5236	1.2892	0.6992	0.2053	-0.0194	-0.7188	-0.2751	-1.2750	-1.7592	1.0413	9
Inverse	1.5234	1.2877	0.6991	0.2053	-0.0149	-0.7153	-0.2862	-1.2763	-1.7551	1.0402	6
Binary	1.5234	1.2862	0.6988	0.2051	-0.0124	-0.7147	-0.2920	-1.2754	-1.7525	1.0398	4
NE log Index	1.5240	1.2840	0.6984	0.2050	-0.0118	-0.7150	-0.2980	-1.2710	-1.7509	1.0398	3
NE Index	1.5249	1.2850	0.6990	0.2055	-0.0113	-0.7147	-0.2994	-1.2704	-1.7503	1.0405	7
NE Inverse	1.5233	1.2833	0.6981	0.2047	-0.0120	-0.7151	-0.2973	-1.2718	-1.7513	1.0392	2
NE Binary	1.5222	1.2827	0.6976	0.2043	-0.0121	-0.7149	-0.2961	-1.2733	-1.7516	1.0385	1
No BE	1.5240	1.2840	0.6982	0.2048	-0.0114	-0.7151	-0.2977	-1.2707	-1.7504	1.0400	5
					Power + population						
log Index	1.5410	1.2591	0.7007	0.2114	-0.0251	-0.7074	-0.3258	-1.2635	-1.7214	1.0539	1
Index	1.5408	1.2596	0.7006	0.2111	-0.0280	-0.7096	-0.3177	-1.2635	-1.7239	1.0541	4
Inverse	1.5411	1.2582	0.7006	0.2115	-0.0228	-0.7064	-0.3315	-1.2626	-1.7190	1.0540	2
Binary	1.5414	1.2566	0.6999	0.2114	-0.0202	-0.7061	-0.3374	-1.2603	-1.7160	1.0544	6
NE log Index	1.5417	1.2557	0.6998	0.2115	-0.0199	-0.7061	-0.3408	-1.2582	-1.7150	1.0545	7
NE Index	1.5417	1.2556	0.6998	0.2115	-0.0200	-0.7062	-0.3407	-1.2580	-1.7149	1.0546	8
NE Inverse	1.5416	1.2558	0.6998	0.2116	-0.0199	-0.7060	-0.3409	-1.2584	-1.7151	1.0544	5
NE Binary	1.5413	1.2559	0.6998	0.2115	-0.0197	-0.7059	-0.3408	-1.2588	-1.7152	1.0541	3
No BE	1.5419	1.2555	0.6995	0.2114	-0.0194	-0.7061	-0.3406	-1.2576	-1.7141	1.0549	9
					Power + density						
log Index	1.5324	1.2984	0.7013	0.2058	-0.0130	-0.7230	-0.2297	-1.2876	-1.7573	1.0474	8
Index	1.5324	1.2988	0.7010	0.2055	-0.0157	-0.7255	-0.2201	-1.2866	-1.7593	1.0483	9
Inverse	1.5324	1.2975	0.7012	0.2058	-0.0108	-0.7216	-0.2370	-1.2876	-1.7553	1.0469	7
Binary	1.5325	1.2962	0.7010	0.2058	-0.0079	-0.7207	-0.2457	-1.2865	-1.7526	1.0465	5
NE log Index	1.5307	1.2912	0.6989	0.2035	-0.0094	-0.7241	-0.2347	-1.2879	-1.7542	1.0441	3
NE Index	1.5310	1.2909	0.6987	0.2035	-0.0093	-0.7248	-0.2349	-1.2862	-1.7537	1.0447	4
NE Inverse	1.5305	1.2915	0.6990	0.2036	-0.0092	-0.7233	-0.2357	-1.2892	-1.7543	1.0438	2
NE Binary	1.5303	1.2926	0.6992	0.2039	-0.0081	-0.7219	-0.2382	-1.2908	-1.7534	1.0438	1
No BE	1.5334	1.2937	0.7005	0.2056	-0.0067	-0.7211	-0.2530	-1.2812	-1.7504	1.0467	6
					Power + urbanicity						
log Index	1.5376	1.2995	0.7009	0.2065	-0.0108	-0.7197	-0.2396	-1.2947	-1.7531	1.0536	8
Index	1.5375	1.2999	0.7005	0.2061	-0.0137	-0.7225	-0.2276	-1.2943	-1.7552	1.0545	9
Inverse	1.5377	1.2987	0.7011	0.2067	-0.0082	-0.7181	-0.2485	-1.2943	-1.7511	1.0531	6
Binary	1.5379	1.2970	0.7010	0.2068	-0.0050	-0.7171	-0.2586	-1.2925	-1.7484	1.0528	5
NE log Index	1.5349	1.2912	0.6981	0.2039	-0.0082	-0.7201	-0.2493	-1.2973	-1.7509	1.0494	2
NE Index	1.5349	1.2902	0.6975	0.2034	-0.0085	-0.7219	-0.2452	-1.2963	-1.7508	1.0500	4
NE Inverse	1.5349	1.2921	0.6987	0.2044	-0.0076	-0.7187	-0.2531	-1.2978	-1.7506	1.0492	1
NE Binary	1.5349	1.2937	0.6993	0.2051	-0.0059	-0.7169	-0.2579	-1.2979	-1.7492	1.0495	3
No BE	1.5388	1.2944	0.7004	0.2065	-0.0036	-0.7175	-0.2654	-1.2876	-1.7463	1.0531	7

Table 24: Weekly mDSS, mean and rank for power models with border effects, continued.

	Endemic				Neighborhood			
	log index	index	inverse	binary	log index	index	inverse	binary
Power								
SK München	3.3281	3.3280	3.3281	3.3282	3.3280	3.3278	3.3284	3.3291
SK Stuttgart	2.2080	2.2081	2.2080	2.2079	2.2079	2.2083	2.2076	2.2074
SK Nürnberg	1.7266	1.7233	1.7289	1.7310	1.7324	1.7331	1.7320	1.7318
SK Mannheim	1.0185	1.0186	1.0190	1.0205	1.0227	1.0249	1.0203	1.0165
SK Augsburg	0.9585	0.9584	0.9584	0.9579	0.9577	0.9584	0.9572	0.9567
SK Fürth	0.2952	0.2994	0.2920	0.2885	0.2870	0.2854	0.2880	0.2888
SK Karlsruhe	0.6445	0.6435	0.6455	0.6470	0.6501	0.6504	0.6499	0.6497
SK Rosenheim	0.6461	0.6422	0.6481	0.6488	0.6423	0.6430	0.6434	0.6494
SK Regensburg	0.3242	0.3244	0.3238	0.3228	0.3224	0.3229	0.3221	0.3217
SK Schweinfurt	0.0169	0.0172	0.0163	0.0152	0.0146	0.0153	0.0140	0.0132
Power + population								
SK München	3.3010	3.3002	3.3014	3.3016	3.3011	3.3009	3.3012	3.3013
SK Stuttgart	2.2214	2.2215	2.2213	2.2212	2.2212	2.2212	2.2212	2.2212
SK Nürnberg	1.7301	1.7262	1.7328	1.7355	1.7367	1.7367	1.7367	1.7365
SK Mannheim	1.0105	1.0090	1.0124	1.0158	1.0166	1.0168	1.0162	1.0155
SK Augsburg	0.9716	0.9715	0.9713	0.9707	0.9707	0.9707	0.9707	0.9707
SK Fürth	0.2973	0.3030	0.2931	0.2885	0.2885	0.2887	0.2884	0.2883
SK Karlsruhe	0.6804	0.6791	0.6818	0.6840	0.6862	0.6862	0.6861	0.6860
SK Rosenheim	0.7172	0.7126	0.7189	0.7178	0.7128	0.7119	0.7139	0.7167
SK Regensburg	0.3117	0.3117	0.3114	0.3106	0.3109	0.3109	0.3109	0.3108
SK Schweinfurt	0.0193	0.0194	0.0190	0.0184	0.0179	0.0180	0.0179	0.0177
Power + density								
SK München	3.2976	3.2974	3.2976	3.2975	3.2918	3.2914	3.2924	3.2936
SK Stuttgart	2.2067	2.2067	2.2068	2.2069	2.2047	2.2048	2.2048	2.2055
SK Nürnberg	1.7260	1.7219	1.7286	1.7315	1.7317	1.7317	1.7317	1.7320
SK Mannheim	1.0139	1.0140	1.0144	1.0159	1.0031	1.0063	1.0012	1.0006
SK Augsburg	0.9636	0.9634	0.9635	0.9631	0.9611	0.9612	0.9612	0.9615
SK Fürth	0.2677	0.2720	0.2645	0.2608	0.2613	0.2619	0.2606	0.2594
SK Karlsruhe	0.6478	0.6468	0.6488	0.6505	0.6523	0.6522	0.6525	0.6528
SK Rosenheim	0.5653	0.5604	0.5681	0.5696	0.5438	0.5412	0.5490	0.5624
SK Regensburg	0.3311	0.3312	0.3306	0.3297	0.3289	0.3287	0.3289	0.3291
SK Schweinfurt	0.0239	0.0242	0.0233	0.0222	0.0198	0.0201	0.0196	0.0193
Power + urbanicity								
SK München	3.2721	3.2714	3.2723	3.2723	3.2624	3.2602	3.2645	3.2677
SK Stuttgart	2.2046	2.2044	2.2047	2.2048	2.2020	2.2016	2.2025	2.2037
SK Nürnberg	1.7231	1.7190	1.7260	1.7290	1.7292	1.7292	1.7292	1.7291
SK Mannheim	1.0092	1.0087	1.0101	1.0119	0.9883	0.9895	0.9889	0.9926
SK Augsburg	0.9662	0.9660	0.9661	0.9657	0.9636	0.9633	0.9639	0.9643
SK Fürth	0.2605	0.2651	0.2570	0.2530	0.2578	0.2592	0.2565	0.2547
SK Karlsruhe	0.6518	0.6506	0.6531	0.6552	0.6573	0.6569	0.6576	0.6579
SK Rosenheim	0.5795	0.5735	0.5827	0.5843	0.5621	0.5526	0.5720	0.5895
SK Regensburg	0.3307	0.3308	0.3303	0.3293	0.3282	0.3279	0.3284	0.3286
SK Schweinfurt	0.0244	0.0248	0.0238	0.0226	0.0189	0.0191	0.0189	0.0191

Table 25: Regional mDSS for 10 densest regions

	Endemic				Neighborhood			
	log index	index	inverse	binary	log index	index	inverse	binary
Power								
LK Amberg-Sulzbach	0.5796	0.5796	0.5797	0.5800	0.5802	0.5805	0.5800	0.5800
LK Garmisch-Partenkirchen	0.1179	0.1169	0.1191	0.1213	0.1253	0.1257	0.1250	0.1245
LK Bayreuth	0.1731	0.1723	0.1735	0.1736	0.1729	0.1733	0.1734	0.1761
LK Rhön-Grabfeld	0.2137	0.2140	0.2134	0.2130	0.2134	0.2128	0.2140	0.2150
LK Regen	0.8664	0.8662	0.8665	0.8665	0.8667	0.8671	0.8665	0.8663
LK Freyung-Grafenau	1.7805	1.7952	1.7610	1.7257	1.6577	1.6546	1.6602	1.6655
SK Straubing	0.1005	0.1007	0.0997	0.0980	0.0965	0.0958	0.0974	0.0985
LK Neustadt/Aisch	0.5028	0.5027	0.5025	0.5019	0.5017	0.5023	0.5012	0.5005
LK Tirschenreuth	0.2523	0.2518	0.2531	0.2542	0.2569	0.2566	0.2574	0.2586
LK Neustadt a.d.Waldnaab	1.9328	1.9326	1.9339	1.9357	1.9382	1.9407	1.9362	1.9335
Power + population								
LK Amberg-Sulzbach	0.5661	0.5650	0.5666	0.5667	0.5659	0.5657	0.5661	0.5665
LK Garmisch-Partenkirchen	0.1403	0.1386	0.1422	0.1456	0.1482	0.1483	0.1481	0.1478
LK Bayreuth	0.2054	0.2047	0.2057	0.2056	0.2054	0.2050	0.2058	0.2069
LK Rhön-Grabfeld	0.2079	0.2084	0.2075	0.2070	0.2075	0.2074	0.2076	0.2079
LK Regen	0.8702	0.8692	0.8715	0.8736	0.8748	0.8749	0.8746	0.8741
LK Freyung-Grafenau	1.7719	1.7964	1.7450	1.7019	1.6606	1.6604	1.6607	1.6614
SK Straubing	0.1757	0.1757	0.1752	0.1742	0.1745	0.1745	0.1745	0.1742
LK Neustadt/Aisch	0.5069	0.5067	0.5068	0.5065	0.5068	0.5069	0.5068	0.5066
LK Tirschenreuth	0.1896	0.1887	0.1903	0.1914	0.1937	0.1935	0.1939	0.1943
LK Neustadt a.d.Waldnaab	1.9534	1.9524	1.9549	1.9569	1.9570	1.9571	1.9568	1.9563
Power + density								
LK Amberg-Sulzbach	0.5904	0.5903	0.5904	0.5906	0.5905	0.5900	0.5910	0.5920
LK Garmisch-Partenkirchen	0.1211	0.1202	0.1224	0.1247	0.1272	0.1277	0.1269	0.1264
LK Bayreuth	0.1953	0.1946	0.1956	0.1955	0.1975	0.1949	0.2007	0.2071
LK Rhön-Grabfeld	0.2079	0.2082	0.2075	0.2070	0.2087	0.2079	0.2093	0.2098
LK Regen	0.8741	0.8739	0.8742	0.8744	0.8733	0.8733	0.8733	0.8736
LK Freyung-Grafenau	1.8336	1.8522	1.8109	1.7706	1.7205	1.7164	1.7236	1.7279
SK Straubing	0.0856	0.0857	0.0848	0.0829	0.0822	0.0807	0.0832	0.0838
LK Neustadt/Aisch	0.5042	0.5041	0.5041	0.5035	0.5007	0.5009	0.5006	0.5009
LK Tirschenreuth	0.2598	0.2593	0.2604	0.2614	0.2663	0.2650	0.2672	0.2684
LK Neustadt a.d.Waldnaab	1.9803	1.9803	1.9806	1.9823	1.9776	1.9788	1.9765	1.9762
Power + urbanicity								
LK Amberg-Sulzbach	0.5927	0.5923	0.5929	0.5931	0.5915	0.5903	0.5925	0.5942
LK Garmisch-Partenkirchen	0.1220	0.1209	0.1235	0.1261	0.1273	0.1277	0.1270	0.1267
LK Bayreuth	0.2066	0.2062	0.2066	0.2063	0.2095	0.2059	0.2129	0.2189
LK Rhön-Grabfeld	0.2041	0.2045	0.2037	0.2032	0.2056	0.2047	0.2062	0.2066
LK Regen	0.8776	0.8773	0.8779	0.8783	0.8758	0.8759	0.8758	0.8761
LK Freyung-Grafenau	1.8083	1.8305	1.7820	1.7364	1.6896	1.6886	1.6901	1.6916
SK Straubing	0.0896	0.0896	0.0887	0.0866	0.0901	0.0879	0.0912	0.0910
LK Neustadt/Aisch	0.5052	0.5050	0.5050	0.5044	0.5006	0.5006	0.5009	0.5015
LK Tirschenreuth	0.2525	0.2519	0.2531	0.2541	0.2573	0.2557	0.2584	0.2598
LK Neustadt a.d.Waldnaab	2.0176	2.0181	2.0179	2.0191	2.0083	2.0095	2.0077	2.0082

Table 26: Regional mDSS for 10 least dense regions