6

Forecasting Based on Surveillance Data

Leonhard Held

University of Zurich

Sebastian Meyer

Friedrich-Alexander-Universität Erlangen-Nürnberg

CONTENTS

6.1 Introduction

Epidemic modelling has at least three distinct aims: Understanding the spread of infectious diseases, identifying suitable measures to control the spread of an epidemic, for example through isolation or vaccination (Daley and Gani, 1999, Section 1.5), and predicting the future course of an epidemic. Mathematical models are often used to better understand the dynamics of infectious disease spread and the effects of control measures (Keeling and Rohani, 2008, Section 1.5), but are less oriented towards predictions. In recent years, more emphasis has been placed on the development of predictive models and methods. The goal of this chapter is to review the literature in this area and to describe how general statistical principles from the forecasting literature can be applied to evaluate the quality of epidemic forecasts. The described methods will also be illustrated in two case studies, where we assess competing approaches to forecast time series of infectious disease counts.

The history of forecasting epidemics

Predicting the future course of epidemics has been a desire of mankind for a long time. Scientific forecasting based on mathematical models dates back to the pioneering work by Baroyan et al. (1971), who predicted the course of an influenza epidemic for the main cities in the Soviet Union. The rise of new infectious diseases has been a major trigger of novel forecasting methods, such as for AIDS (Daley and Gani, 1999, Section 6.2) and

SARS (Hufnagel et al., 2004). Furthermore, meteorologic forecasting methods have been adopted in epidemiological research, including modelling (Viboud et al., 2003; Shaman and Karspeck, 2012) and assessment (Paul and Held, 2011; Held et al., 2017) techniques. Recent developments in influenza forecasting have focussed on the integration of search logs from Google (Dukic et al., 2012; Shaman and Karspeck, 2012; Yang et al., 2015), social media data from Twitter (Paul et al., 2014), combinations thereof (Santillana et al., 2015), Wikipedia article views (Generous et al., 2014; Hickmann et al., 2015), or human mobility data (Pei et al., 2018). Compared to weather forecasting, epidemic forecasting is still in its infancy, and the human component makes it particularly challenging (World Health Organization, 2014; Moran et al., 2016).

Forecasting competitions

In order to develop better epidemic forecasting methods, the World Health Organization has joined forces in an informal consultation with more than 130 global experts (World Health Organization, 2016), and the Centers for Disease Control and Prevention in the USA have organized several seasonal forecasting competitions (Biggerstaff et al., 2016, 2018). Real-time influenza forecasts are now provided by Nicholas Reich and co-workers at the ReichLab (Tushar et al., 2017), see `http://reichlab.io/flusight/`. Other recent competitions include the DAPRA challenge on forecasting chikungunya (Del Valle et al., 2018), the White House/NOAA challenge on forecasting dengue (Buczak et al., 2018), and the RAPIDD ebola forecasting challenge (Viboud et al., 2018).

Forecasting targets

Several quantities are of interest in epidemic forecasting, such as timing of and incidence in the peak week (Ray et al., 2017), onset week (Pei et al., 2018), cumulative incidence (Lega and Brown, 2016), weekly incidence (Paul and Held, 2011; Reich et al., 2016a), outbreak size and duration (Farrington et al., 2003), and the epidemic curve (Jiang et al., 2009). Also of public health relevance are stratified forecasts of the above quantities, for example by region or by age group (Held et al., 2017). Forecasting targets for seasonal influenza epidemics in particular have been reviewed previously (Chretien et al., 2014; Nsoesie et al., 2014). More recently, the weekly proportion of doctor visits due to influenza-like illness is becoming a popular forecasting target (Biggerstaff et al., 2018).

6.2 Evaluating point forecasts

A point forecast is usually made for a continuous or integer (often count) outcome, for example the disease incidence in the next week, the number of newly confirmed cases in a certain time interval, or the timing of the peak week of an epidemic. Several measures have been proposed and used to evaluate the quality of a point forecast. Gneiting (2011) gives a comprehensive discussion of suitable measures for point predictions.

To introduce some notation, let $\hat{y}_i$ denote the point forecasts of the observations y_i, $i = 1, \ldots, n$. Given a suitable non-negative scoring function $S(\hat{y}_i, y_i)$, the overall predictive performance can be assessed with the mean score

$$\bar{S} = \frac{1}{n} \sum_{i=1}^{n} S(\hat{y}_i, y_i).$$

Scoring functions are usually negatively oriented, so the smaller a score, the better the

forecast. Commonly used scoring functions are the *absolute error* (AE) $S(\hat{y}, y) = |\hat{y} - y|$, the *squared error* (SE) $S(\hat{y}, y) = (\hat{y} - y)^2$, the *absolute percentage error* $S(\hat{y}, y) = |\hat{y} - y|/y$ and the *relative error* $S(\hat{y}, y) = |\hat{y} - y|/\hat{y}$. Note that the latter two scoring functions require y and $\hat{y}$ to be positive, respectively. Other summary measures used in the infectious disease literature may be based on one of the scoring functions mentioned above, for example the *root mean squared error* (RMSE) or the *relative mean absolute error* (Hyndman and Koehler, 2006; Reich et al., 2016b). The latter is defined as the ratio of the *mean absolute errors* of two competing forecasts and is not to be confused with the *mean relative absolute error* of one forecasting method.

The small simulation study reported by Gneiting (2011) reveals the key differences of the four scoring functions listed above. It is based on a time series model commonly used in econometrics, but the results equally apply to other fields. A simple conditionally heteroscedastic Gaussian time series model is used to generate a non-negative time series $y_1, \ldots, y_n$. Model-based one-step-ahead point forecasts (based on the mean of the forecast distribution) are then compared to naive approaches, such as the so-called "fence-sitter forecast", a forecast with constant predictions. Quite surprisingly, one of the naive methods outperforms the model-based one-step-ahead forecasts both under the *absolute error* and the *absolute percentage error* scoring functions.

Gneiting (2011) points out that the model-based forecast is only optimal under squared error loss but not necessarily for other loss functions, where other functions of the predictive distribution will be optimal. A scoring function is called consistent for a loss function, if the expected score is minimized when following this loss function. The *squared error* scoring function is consistent for the mean, and the *absolute error* scoring function for the median, both standard loss functions. However, the *mean absolute percentage error* commonly used in influenza forecasting (Chretien et al., 2014, Table 4), is consistent for a nonstandard and rather exotic functional. To quote Gneiting (2011), "it thus seems prudent that authors using this functional consider the intended or unintended consequences and reassess its suitability as a scoring function." Some scoring functions may be problematic *per se*, such as the commonly used correlation coefficient between predictions and observations. For example, point predictions always twice as large as the observations are obviously inappropriate, but the correlation coefficient between predictions and observations will be one.

6.3 Evaluating probabilistic forecasts

In many areas of science, researchers argue that forecasts should be probabilistic (Gneiting and Katzfuss, 2014) and this plea has recently been taken up in the infectious disease literature (Held et al., 2017; Funk et al., 2018).

Evaluating probabilistic forecasts requires suitable scores to quantify the "distance" between a cumulative distribution function F, the forecast, and the observation y that later realizes. Proper scoring rules play a key role, as they allow for a fair comparison of different probabilistic forecasting methods. In addition, the notions of calibration and sharpness are important. Calibration is a property of both the forecast and the observations, whereas sharpness is a property of the forecasts only. The paradigm of probabilistic forecasting is to "maximize sharpness subject to calibration" (Gneiting and Raftery, 2007).

Much of the literature on probabilistic forecasting focusses on continuous forecast distributions. However, in infectious disease epidemiology, the quantities to be predicted are often (incidence or prevalence) counts, which require suitable extensions of the assessment techniques (Czado et al., 2009; Held et al., 2017).

6.3.1 Calibration and sharpness

Calibration is defined as the statistical consistency of probabilistic forecasts and the observations (Gneiting et al., 2007). For continuous outcomes, calibration is usually assessed with the probability integral transform (PIT). Specifically, the PIT is $F(y)$, so will be uniformly distributed if the observation y is a realisation from the forecast F. In practice we will compute PIT values for a series of forecasts and visually assess their distribution in a histogram. Modifications for count data exist (Czado et al., 2009).

For binary data, the calibration slope studies the association between the (logit-transformed) predicted probabilities with the binary outcomes in a logistic regression (Cox, 1958; Steyerberg, 2009). Calibration can also be visually assessed in a calibration curve (Gneiting and Katzfuss, 2014).

Statistical calibration tests are commonly employed to assess the evidence for miscalibration. Different methods exist for continuous outcomes (Mason et al., 2007; Held et al., 2010), count outcomes (Wei and Held, 2014) and binary outcomes (Spiegelhalter, 1986), but they may require certain distributional assumptions on the forecasts.

For continuous forecasts, sharpness is usually defined as the width of the associated prediction intervals. For multivariate forecasts, sharpness is defined based on the predictive covariance matrix, for example its determinant (Gneiting et al., 2008).

6.3.2 Proper scoring rules

Proper scoring rules are summary measures of the predictive performance of probabilistic predictions $Y_i \sim F_i$, $i = 1, \ldots, n$, allowing for a joint assessment of calibration and sharpness (Gneiting and Katzfuss, 2014). As in Section 6.2 for point forecasts, the mean score across forecasts is usually reported and compared across different forecasting methods. For ease of presentation we drop the index i in this section and compare the probabilistic forecast $Y \sim F$ with the actual observation y.

The definition of propriety is mathematically somewhat demanding and we refer the interested reader to the review paper by Gneiting and Raftery (2007). Put simply, a proper scoring rule ensures that a forecaster reports his actual forecast and does not obtain a better score in expectation by digressing from his belief. Here we take scores to be negatively oriented penalties that forecasters wish to minimize, so the smaller a score, the better the forecast.

For forecasts of binary data $y \in \{0, 1\}$, commonly used scores are the *logarithmic score* (LS), *i. e.* the negative log-likelihood of the observation under the forecast distribution, and the *Brier score* (BS), also known as probability score. Specifically, let π denote the predicted probability of the observation $y = 1$, then

$$\begin{aligned} \mathrm{LS}(F, y) &= -y \log(\pi) - (1 - y) \log(1 - \pi), \text{ and} \\ \mathrm{BS}(F, y) &= (\pi - y)^2. \end{aligned}$$

Both scores are proper, whereas the *absolute score* $\mathrm{AS}(F, y) = |\pi - y|$ can be shown to be improper (Held and Sabanés Bové, 2014, Chapter 9), so should not be used.

A probabilistic forecast for count data $y \in \{0, 1, 2, \ldots\}$ is represented by probabilities $\pi_k = \mathsf{Pr}(y = k)$, $k \in \{0, 1, 2, \ldots\}$, where $\sum_{k=1}^{\infty} \pi_k = 1$. The logarithmic score for the observation $y = j$ now simply reads $\mathrm{LS} = \log(\pi_j)$ where π_j is the probability of the observation $y = j$. An extension of the Brier score to forecasts of count data is the *ranked probability score*

$$\mathrm{RPS}(F, y) = \sum_{j=1}^{\infty} (P_j - \mathbf{1}\{y \leq j\})^2.$$

where $P_j = \sum_{k=0}^{j} \pi_k$ and $\mathbf{1}$ denotes the indicator function. Probabilistic count forecasts are often summarized with the mean μ and the variance σ^2. A proper score that only uses these first two moments is the *Dawid-Sebastiani score*

$$\mathrm{DSS}(F, y) = 2 \log \sigma - (y - \mu)^2 / \sigma^2.$$

This score can easily be generalized to multivariate predictions,

$$\mathrm{mDSS}(F, \boldsymbol{y}) = \log|\boldsymbol{\Sigma}| + (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \tag{6.1}$$

which depends only on the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the predictive distribution. The first term $\log|\boldsymbol{\Sigma}|$ in (6.1) is called the log determinant sharpness (logDS) and is recommended as a (multivariate) measure of sharpness (Gneiting et al., 2008). To avoid unnecessarily large numbers it is common practice to report scaled versions of logDS and mDSS, obtained through division by $2d$ where d is the dimension of the observation $\boldsymbol{y}$. A possible alternative to mDSS is the *energy score*, a multivariate extension of the ranked probability score (Gneiting et al., 2008; Held et al., 2017).

The question arises which scoring rule should be used in practice. The logarithmic score is known to be sensitive to outliers if $\pi_j = 0$ is close to zero. The ranked probability score is reported to be less sensitive to extreme observations (Gneiting and Raftery, 2007) and provides an attractive alternative. The DSS is particularly useful if the first two predictive moments are available but not the whole predictive distribution. It has also been argued that the choice of a scoring rule should take into account the costs of bad forecasts (Merkle and Steyvers, 2013). In practice we recommend to evaluate several scores to obtain a more robust comparison of predictive performance.

The computation of the different scores depends on whether the forecast distributions are known analytically or derived from simulations. Care has to be taken with simulation-based forecasts as scores can become numerically instable, for example the logarithmic score for count forecasts will be infinite if an observation $y = j$ has empirical frequency equal to zero in the forecast samples. A possible remedy is to apply a Rao-Blackwell/importance sampling scheme (Gelfand and Smith, 1990) and to average the conditional predictive distributions, see also Ray et al. (2017, Supp Mat Section 3.2). Alternatively one may use kernel density estimation as implemented in the R package `scoringRules` (Jordan et al., 2018). The Dawid-Sebastiani score for multivariate forecasts can be numerically unstable if the predictive covariance matrix $\boldsymbol{\Sigma}$ is estimated from the empirical covariance matrix of Monte Carlo samples (Scheuerer and Hamill, 2015).

For independent forecasts, a simple paired t-test or a permutation test can then be used to assess the evidence for differences in predictive performance between two competing forecasting methods. If the forecasts are dependent as in a series of sequential k-step ahead forecasts, the Diebold-Mariano test can be used to account for the correlation between scores (Diebold and Mariano, 1995; Gneiting and Katzfuss, 2014).

6.4 Applications

We now describe two applications, in which we compare the quality of forecasts provided by different models and methods. The data and code to reproduce these analyses are available in the R package `HIDDA.forecasting` (Meyer, 2018) at `https://HIDDA.github.io/forecasting`. The code is presented in several package vignettes, which also give some additional results.

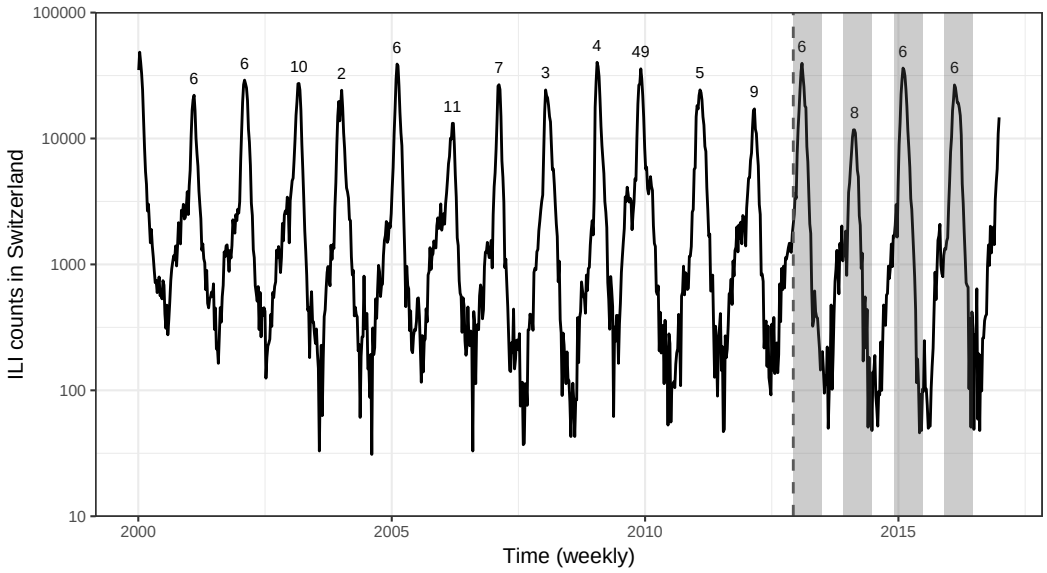


FIGURE 6.1: Surveillance of influenza-like illness (ILI) in Switzerland, 2000–2016. The last 213 weeks (to the right of the vertical dashed line) are used to assess one-week-ahead forecasts. Long-term forecasts are assessed for the last four seasons starting in the first week of December (shaded time periods). The seasonal peaks are labelled with the corresponding ISO week. Note that the y-axis employs a log-scale for the ILI counts.

6.4.1 Univariate forecasting of influenza incidence

We consider a time series of weekly incidence counts on influenza-like illness (ILI) in Switzerland from 2000 to 2016 (Figure 6.1). The last 213 weeks (starting from 2012-12-04) are used to assess one-week-ahead forecasts. Four sets of 30-weeks-ahead long-term forecasts (shaded time periods in Figure 6.1) are computed for each of the last four seasons, conditional upon all data prior to the respective first week of December.

We have used five different methods to produce forecasts of influenza activity in Switzerland: an ARMA(2,2) time series model as estimated by `auto.arima()` from the `forecast` package (Hyndman and Khandakar, 2008), an observation-driven ARMA model for negative-binomial counts as implemented in the `glarma` package (Dunsmuir and Scott, 2015), the endemic-epidemic negative-binomial time-series model `hhh4` from the `surveillance` package (Meyer et al., 2017), Facebook's forecasting tool `prophet` (Taylor and Letham, 2018), as well as a recently proposed method based on kernel conditional density estimation (Ray et al., 2017), which we implemented following the code provided at `https://github.com/reichlab/article-disease-pred-with-kcde`. Note that forecasts from `arima` and `prophet` are generated on a log-scale. All models were configured to account for yearly seasonality. In the first three models, seasonality was represented by parametric sine-cosine regressors (Held and Paul, 2012) with frequency $2\pi/52.1775$ (derived from the average number of calendar weeks per year). For `prophet` and `kcde`, we followed the documented examples (see the supplementary vignettes for the exact model formulations). We also included a separate Christmas effect for calendar week 52 (not with `kcde`). For reference, we computed naive historical forecasts for the 213 weeks of the test period, based on fitting log-normal distributions to the observed counts in the same calendar week of previous years.

We computed the RMSE and the mean DSS and LS for the one-week-ahead forecasts and

TABLE 6.1: Mean scores of the 213 one-week-ahead predictions (2012-W49 to 2016-W52) and of the long-term forecasts for the last four seasons of the Swiss ILI surveillance data. Computing these long-term forecasts with the experimental `kcde` implementation was too cumbersome. The runtime for computing a single one-week-ahead forecast is given in seconds. For `hhh4` and `glarma`, the long-term results are based on 1000 simulations.

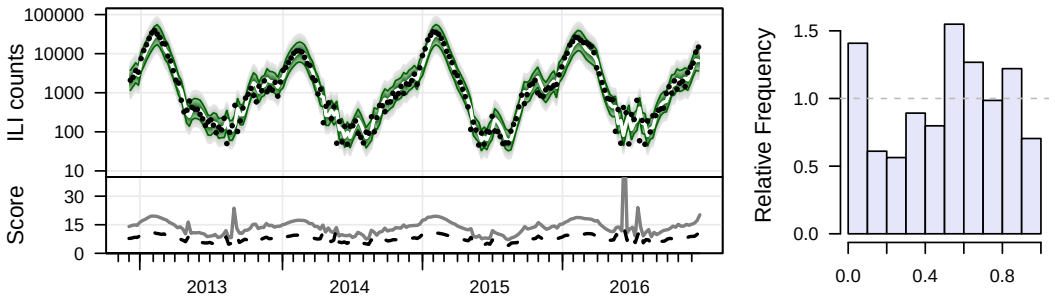
	One-week-ahead				Long-term		
Method	RMSE	DSS	LS	runtime	RMSE	DSS	LS
`arima`	2287	13.78	7.73	0.51	8471	16.43	8.88
`glarma`	2450	13.59	7.71	1.49	5558	19.61	9.12
`hhh4`	1769	13.58	7.71	0.02	8749	16.13	9.25
`prophet`	5614	15.00	8.03	3.01	7627	16.44	8.91
`kcde`	1963	13.79	7.80	1.30	—	—	—
naive	5010	14.90	8.06	0.00	6527	15.99	8.86

the long-term forecasts. The results are summarized in Table 6.1. As expected, the one-week-ahead foreacasts have smaller (*i. e.* better) mean scores than the long-term forecasts for all methods considered. Note that we have not attempted to compute long-term forecasts with `kcde` due to excessively long runtimes. The time taken to compute a single one-week-ahead forecast varied between 0.001 (naive) and 3 (`prophet`) seconds.

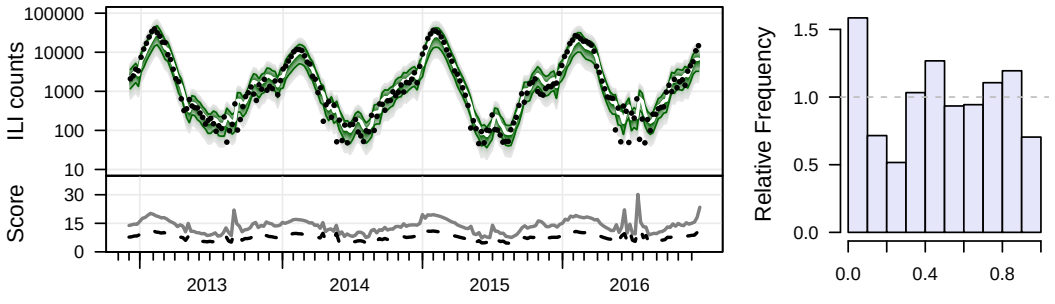
The best one-week-ahead forecasts in terms of all scores are obtained with `hhh4` followed by `glarma`, whose point forecasts are slightly worse with an average error of 2450 cases. The `arima` and `kcde` methods come next and the worst one-week-ahead predictions are obtained from `prophet`, which achieves similar scores as the naive approach. Figure 6.2 shows the weekly forecasts and associated scores as well as the overall PIT histograms, which are computed based on the method for count data (Czado et al., 2009). There is no clear evidence for miscalibration of any of the one-week-ahead forecasts, but the first three and the naive (not shown) forecasting methods have a distinct peak of the histogram in the first bin, indicating some evidence for biased or underdispersed predictions. The `prophet` and `kcde` (not shown) forecasts don't have this problem. PIT histograms for the `kcde` and naive forecasts can be found in the corresponding supplementary vignettes.

An interesting result is obtained from the long-term forecasts, where the best scores are achieved by taking simple historical averages, except for RMSE where `glarma` is better. For the more sophisticated forecasts shown in Figure 6.3, the ranking is less clear. While `hhh4` is the best model in terms of DSS, `arima` is the best model in terms of LS. Quite surprising is the large DSS value of the `glarma` method. Closer inspection of Figure 6.3 suggests that this is caused by the excessively large uncertainty of the `glarma` long-term predictions, compared to the other three methods. From Figure 6.4 we can see that the `prophet` method has a particularly large DSS score in the 2015/2016 season, where the DSS score is roughly twice as large as for the other methods. Closer inspection of Figure 6.3 reveals that this is due to under-prediction of the incidence combined with an insufficient amount of uncertainty at the end of the epidemic season.

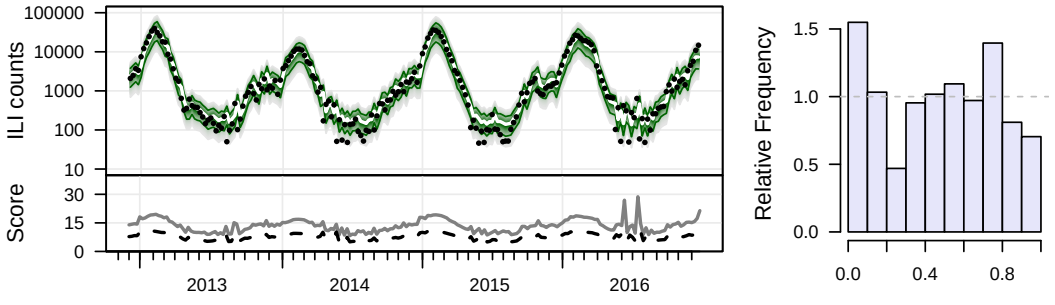
It may also be of interest to predict the peak week in each of the four years based on the long-term forecasts. For illustration, we have computed a probabilistic prediction of the peak week based on the samples from the `hhh4` model only. The median peak week forecasts is always calendar week 5. This is quite close to the observed peak weeks 6, 8, 6, and 6, respectively, shown in Figure 6.1. However, considerable uncertainty is attached to these predictions with 2.5% quantile the second week and 97.5% quantile week 16 of each season.



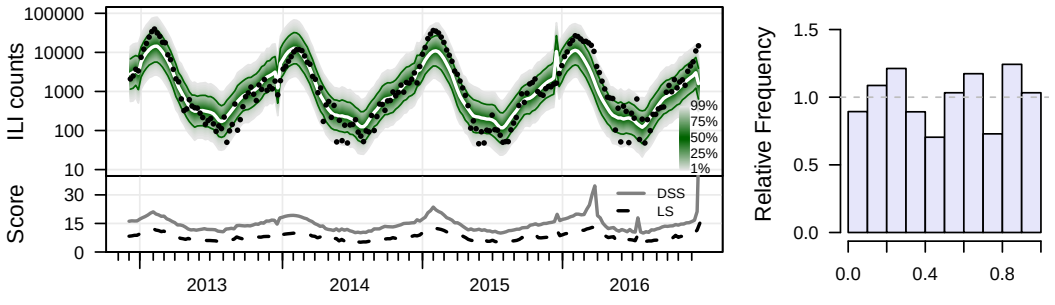
(a) `arima`



(b) `glarma`

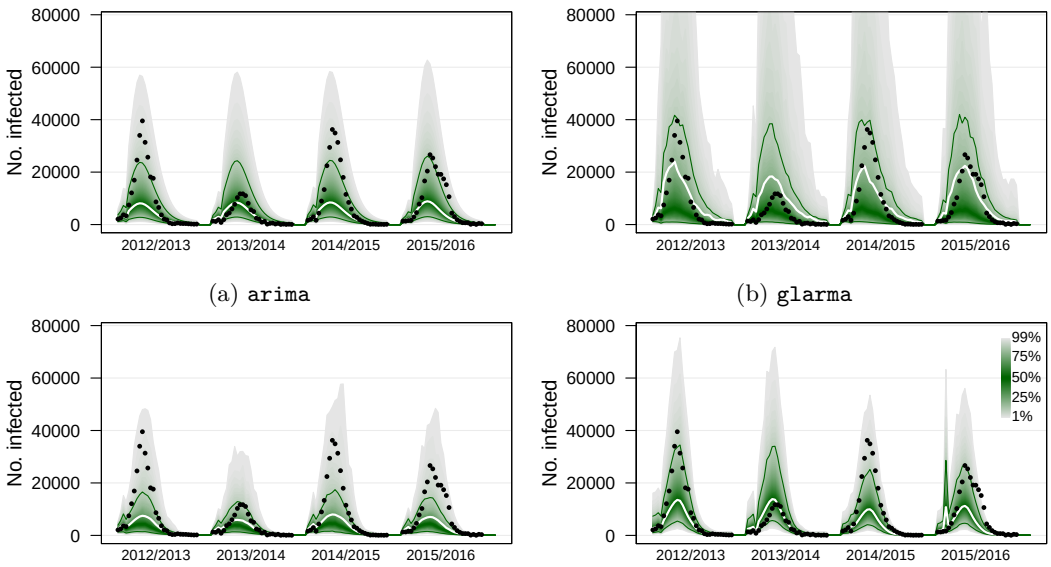


(c) `hhh4`



(d) `prophet`

FIGURE 6.2: One-week-ahead forecasts of ILI counts in Switzerland for the last 213 weeks (2012-W49 to 2016-W52) of the available surveillance data. Predictive distributions are displayed as fan charts on a log-scale. The 10% and 90% quantiles and the mean are highlighted. The dots correspond to the observed counts and the lower panels show the associated weekly scores. The right-hand column shows the corresponding PIT histograms. Plots for `kcde` and naive forecasts can be found in the supplementary vignettes.



(a) `arima` (b) `glarma`

(c) `hhh4` (d) `prophet`

FIGURE 6.3: Long-term forecasts (30-weeks-ahead) of ILI counts in Switzerland for the last four seasons of the available surveillance data. Predictive distributions are displayed as fan charts, where the 10% and 90% quantiles and the mean are highlighted. The dots represent observed counts. In (b), quantiles above 95% are truncated around the peak weeks with the 99% quantile reaching seasonal maxima of between 251000 and 325000 cases.

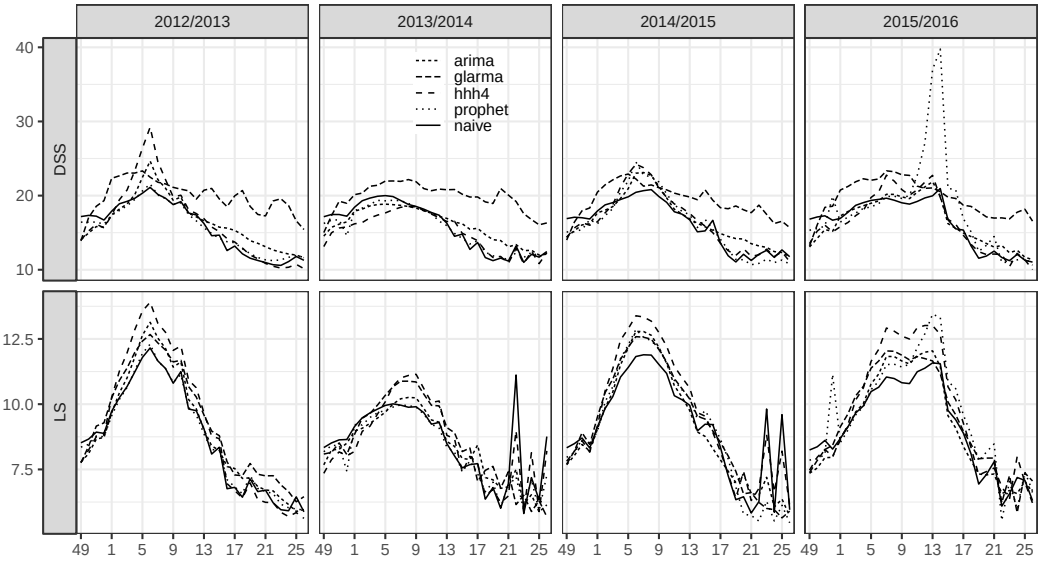


FIGURE 6.4: Weekly DSS (top) and LS (bottom) values of the different long-term forecasts of Swiss ILI counts displayed in Figure 6.3 and the naive forecasts.

6.4.2 Multivariate forecasting of norovirus incidence

To illustrate statistical methods for the assessment of multivariate forecasts we use age-stratified surveillance data on norovirus gastroenteritis from Berlin, Germany. We use the well-established `hhh4` modelling framework from the `surveillance` package to generate forecasts, building on our previous analyses of these data (Meyer and Held, 2017; Held et al., 2017). Here we only consider models for spatially aggregated counts Y_{gt}, $g = 1, \ldots, G$, *i. e.* without additional stratification by city district. The surveillance data are available from the package `hhh4contacts` (Meyer, 2017), cover five norovirus seasons from 2011-W27 to 2016-W26, and are stratified into $G = 6$ age groups: 0–4, 5–14, 15–24, 25–44, 45–64, and 65+ years of age. These groups reflect distinct social mixing of pre-school vs. school children, and intergenerational mixing. Figure 6.5 shows the age-specific incidence time series. Over the five years, the reported numbers of cases by age group were 2783, 326, 553, 1909, 2530, and 8335, respectively, corresponding to average yearly incidences of 346 (0–4), 24 (5–14), 30 (15–24), 36 (25–44), 52 (45–64), and 251 (65+) cases per 100 000 inhabitants.

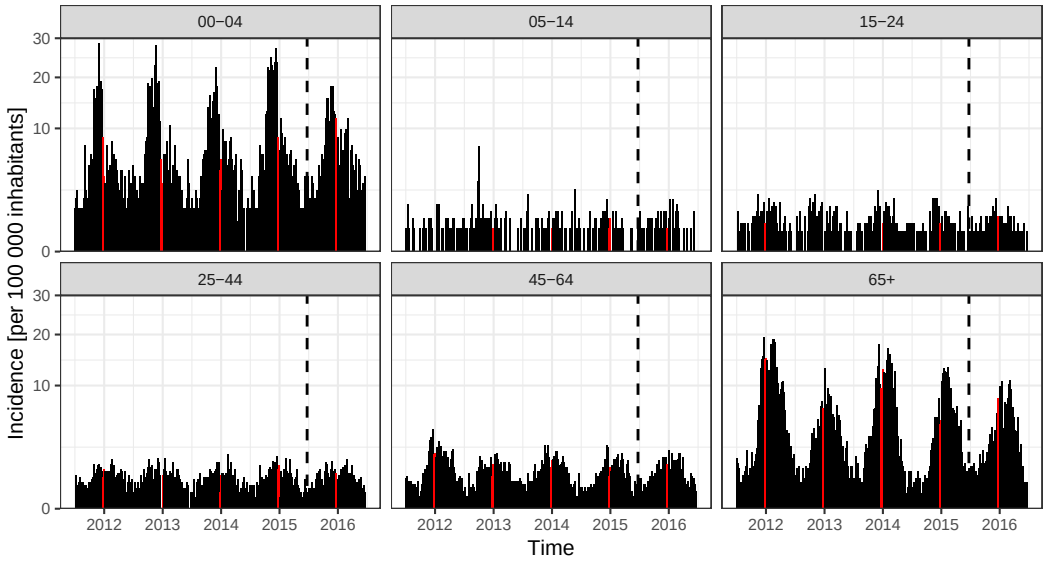


FIGURE 6.5: Age-stratified incidence of norovirus gastroenteritis in Berlin, Germany. The calendar weeks 52 and 1 (Christmas break period) are highlighted. The last 52 weeks (to the right of the vertical dashed line) are used to assess forecasts. Note that the y-axis employs a square-root scale.

Given the counts from the previous week, $Y_{g',t-1}$, $g' = 1, \ldots, G$, we assume Y_{gt} to follow a negative binomial distribution with a group-specific overdispersion parameter and mean

$$\mu_{gt} = \nu_{gt} + \phi_{gt} \sum_{g'} c_{g'g} Y_{g',t-1}.$$

The *endemic* log-linear predictor ν_{gt} contains group-specific intercepts, a Christmas break effect (via a simple indicator for the calendar weeks 52 and 1), and group-specific seasonal effects via $\sin(\omega t)$ and $\cos(\omega t)$ terms, $\omega = 2\pi/52$. The *epidemic* log-linear predictor ϕ_{gt} also contains group-specific intercepts, but shared seasonality and no Christmas break effect. The contact weights $c_{g'g}$ are estimated from the German subset of the POLYMOD survey (Mossong et al., 2008), taking into account the reciprocal nature of contacts (Wallinga

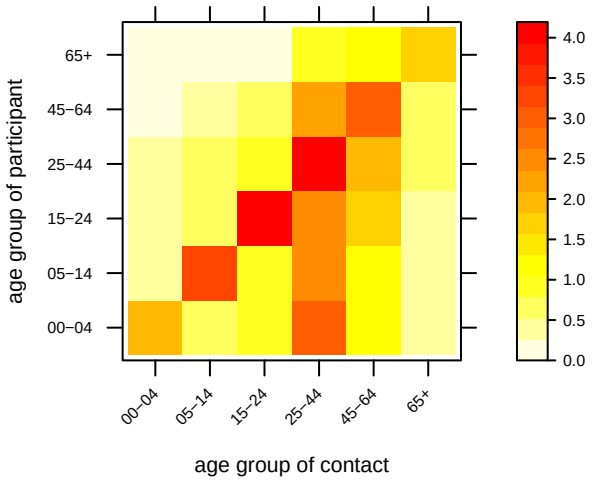


FIGURE 6.6: Age-structured social contact matrix estimated from the German POLYMOD sample. The entries refer to the mean number of contact persons per participant per day.

et al., 2006). The resulting contact matrix (Figure 6.6) is subsequently row-normalized to a transition matrix, removing differences in group-specific overall contact rates.

The model described above is the same as reference model 6 in Held et al. (2017, Section 3.3). As alternative models we consider homogeneous mixing between age groups ($c_{g'g} = 1$), no mixing between age groups ($c_{g'g} = \mathbf{1}\{g' = g\}$) and a model where we estimate a power transformation C^κ of the contact matrix as described in Meyer and Held (2017). See the `vignette("BNV")` in the supplementary package `HIDDA.forecasting` for how to implement these models in R.

We have fitted the models to the first four seasons (2011-W27 to 2015-W26) and evaluated the quality of forecasts in the subsequent season (2015-W27 to 2016-W26). Figure 6.7 shows the fitted mean components from the AIC-optimal model with power-adjusted contact matrix. The proportion of the incidence that can be explained by the epidemic component varies between the age groups. The endemic-epidemic decomposition actually resembles Figure 4 in Meyer and Held (2017) quite well, which is based on a different model, additionally stratified by city district. Within the epidemic part, the incorporated social contact matrix results in predominant within-group reproduction of the disease for the 0–4 and 65+ age groups, whereas the 25–44 and 45–64 year-old persons inherit a substantial proportion of cases from other age groups.

Turning to the predictive performance, Figure 6.8 shows the one-week-ahead forecasts and associated scores during the last season for the power-adjusted contact model. Average scores of the different models are compared in Table 6.2. Although the power-adjusted model gives the best AIC score on the training data, it is outperformed by the "no mixing" model on the test season, both in terms of DSS and LS. This can be explained by Table 6.3, which gives the log-likelihood contributions of the different models from a separate fit to the whole time period (training and test). We can see that the "no mixing" model has indeed a larger log-likelihood contribution in 2015/16 than all other three models. However, the power-adjusted model is best in 2011/12, 2013/14 and 2014/15 and second best in 2012/13 (after "reciprocal"). This explains why the power-adjusted model performs best in the training, but not in the test period.

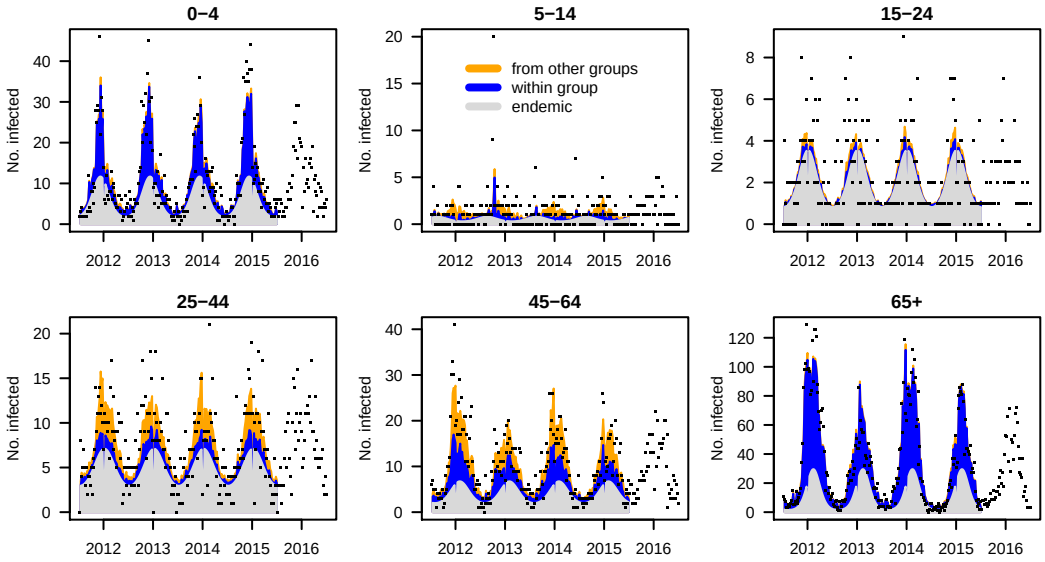


FIGURE 6.7: Estimated norovirus counts from the AIC-optimal `hhh4` model with power-adjusted contact matrix. The disease risk is additively decomposed into endemic and epidemic components. The dots show the observed counts.

TABLE 6.2: Comparison of four `hhh4` models for the age-stratified norovirus time-series from Berlin, assuming different contact matrices. The table gives the number of parameters, the AIC for the training period, and average scores of the 6×52 one-week-ahead forecasts in the last season.

	dim	AIC	DSS	LS
reciprocal	33	6051	3.031	2.399
homogeneous	33	6132	3.093	2.420
no mixing	33	6055	3.003	2.385
power-adjusted	34	6035	3.012	2.391

TABLE 6.3: Seasonal log-likelihood contributions in the norovirus models from Table 6.2, now fitted to the whole time period. The last column gives the overall AIC.

	2011/12	2012/13	2013/14	2014/15	2015/16	AIC
reciprocal	-747.53	-758.71	-763.60	-726.83	-740.38	7540
homogeneous	-762.44	-766.59	-769.68	-739.53	-745.16	7633
no mixing	-750.36	-763.29	-762.37	-721.71	-737.47	7536
power-adjusted	-744.43	-759.06	-762.01	-721.68	-738.63	7520

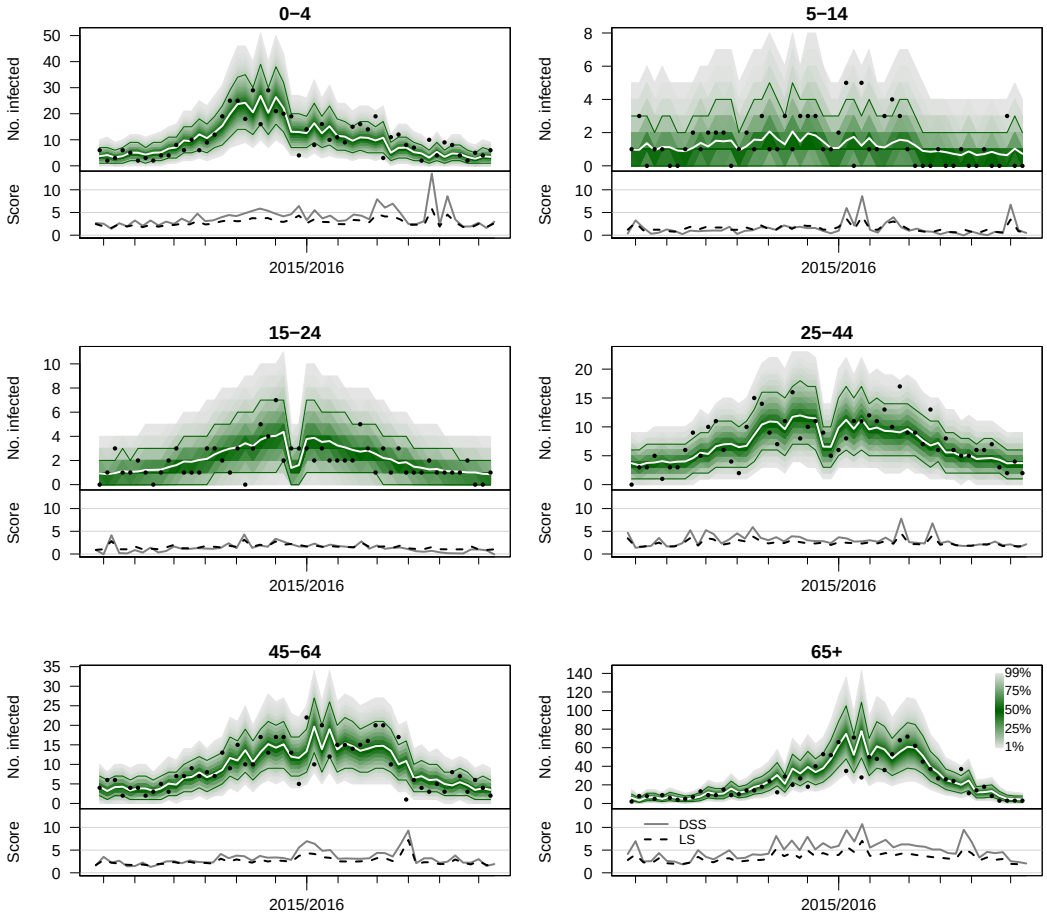


FIGURE 6.8: Age-stratified one-week-ahead forecasts of norovirus counts in Berlin based on the power-adjusted contact model. The 10% and 90% quantiles and the means are highlighted within the fan charts. The dots represent observed counts. The bottom panels show weekly scores.

Multivariate long-term predictions from the four models are evaluated in Table 6.4 for the test period. The first two columns give mDSS and logDS for multivariate predictions (of dimension 6×52) across weeks and age groups. The "no mixing" model is again best in terms of mDSS, although both the "reciprocal" and the power-adjusted model provide sharper predictions in terms of logDS. Predictions of the age-stratified total number of cases (now of dimension 6) are displayed in Figure 6.9 and the corresponding scores listed in the last two columns of Table 6.4. More uncertainty of the predictions (in age group 65+) is visible for the "no mixing" and power-adjusted models, resulting in larger values of logDS. However, the increased uncertainty seems to be benificial for the quality of the forecasts, yealding the smallest mDSS scores for these two formulations. In contrast, the "homogeneous" prediction barely covers the observed count in age group 65+, resulting in a poor mDSS score. For comparison, a naive forecaster who independently fits negative binomial distributions to the age-stratified sizes of the past four seasons would reach mDSS = 4.084 (with logDS = 3.820), which is superior to the predictions from the homogeneous contact model.

TABLE 6.4: Scaled multivariate Dawid-Sebastiani scores (mDSS) and log determinant sharpness (logDS) of long-term predictions for the last norovirus season from the four different models. Aggregated predictions refer to the final size by age group (Figure 6.9).

	Weekly		Aggregated	
	mDSS	logDS	mDSS	logDS
reciprocal	1.539	1.067	4.098	3.508
homogeneous	1.564	1.096	4.205	3.393
no mixing	1.521	1.078	4.066	3.742
power-adjusted	1.527	1.065	4.071	3.638

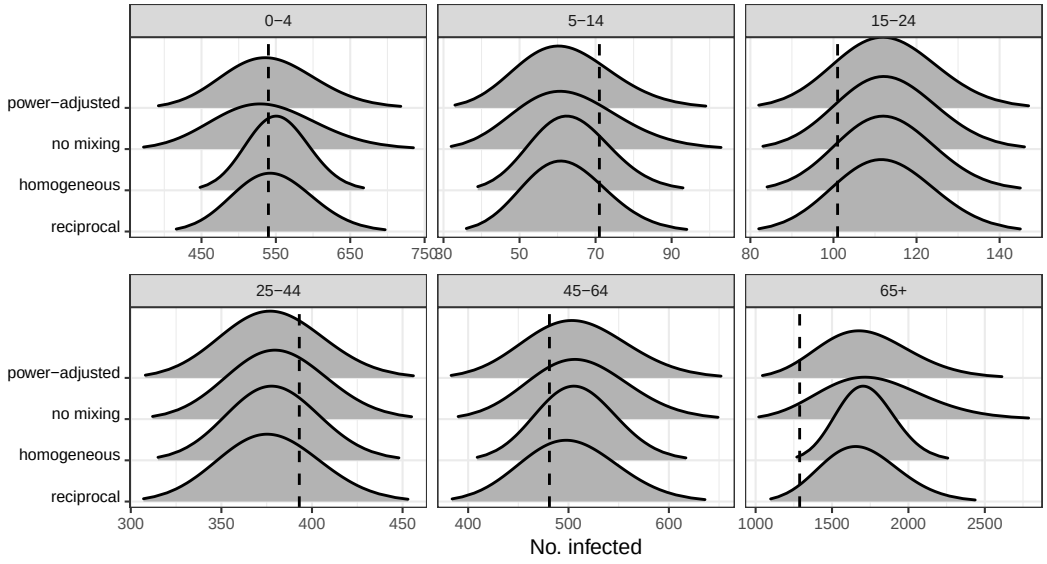


FIGURE 6.9: Forecast distributions of age-stratified sizes of the 2015/2016 norovirus epidemic from four different models. Shown are negative binomial approximations based on predictive moments, see Held et al. (2017, Appendix A) for details. The dashed vertical lines represent the observed counts.

6.5 Discussion

We have provided a non-comprehensive review of the literature on forecasting epidemics. We have focussed on statistical methods to quantify the accuracy of predictions, distinguishing between point and probabilistic forecasts. Two applications show how the different techniques can be applied to uni- and multivariate forecasts.

Inspired by similar techniques in weather forecasting and other areas of science, recent work in this area has focussed on model averaging and stacking in order to improve the predictive quality of single model forecasts (Ray and Reich, 2018). Perhaps the biggest challenge to epidemic forecasting is the incorporation of reporting delays and underreporting, as described in Chapter V.3 of this handbook. A rigorous analysis requires surveillance and internet search data archived in a way where you can actually see what was available at a given time (McIver and Brownstein, 2014). Such real-time epidemiological data should become the standard in surveillance systems, facilitating the development of novel forecasting techniques.

Acknowledgments

We are grateful to the Swiss Federal Office of Public Health (BAG) for access to the data on influenza in Switzerland. We thank Nicholas Reich for helpful comments on a previous version of this chapter.

References

Baroyan, O. V., Rvachev, L. A., Basilevsky, U. V., Ermakov, V. V., Frank, K. D., Rvachev, M. A., and Shashkov, V. A. (1971). Computer modelling of influenza epidemics for the whole country (USSR). *Advances in Applied Probability*, 3(2):224–226.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickmann, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., Velardi, P., Vespignani, A., and Finelli, L. (2016). Results from the Centers for Disease Control and Prevention's Predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16(1):357.

Biggerstaff, M., Johansson, M., Alper, D., Brooks, L. C., Chakraborty, P., Farrow, D. C., Hyun, S., Kandula, S., McGowan, C., Ramakrishnan, N., Rosenfeld, R., Shaman, J., Tibshirani, R., Tibshirani, R. J., Vespignani, A., Yang, W., Zhang, Q., and Reed, C. (2018). Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics*, 24:26–33.

Buczak, A. L., Baugher, B., Moniz, L. J., Bagley, T., Babin, S. M., and Guven, E. (2018). Ensemble method for dengue prediction. *PLOS ONE*, 13(1):e0189988.

Chretien, J. P., George, D., Shaman, J., Chitale, R. A., and McKenzie, F. E. (2014). Influenza forecasting in human populations: A scoping review. *PLOS ONE*, 9(4):e94130.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565.

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

Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An Introduction*. Cambridge University Press, Cambridge.

Del Valle, S. Y., McMahon, B. H., Asher, J., Hatchett, R., Lega, J. C., Brown, H. E., Leany, M. E., Pantazis, Y., Roberts, D. J., Moore, S., Peterson, A. T., Escobar, L. E., Qiao, H., Hengartner, N. W., and Mukundan, H. (2018). Summary results of the 2014-2015 DARPA Chikungunya challenge. *BMC Infectious Diseases*, 18(1):245.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Association*, 107(500):1410–1426.

Dunsmuir, W. and Scott, D. (2015). The `glarma` package for observation-driven time series regression of counts. *Journal of Statistical Software*, 67(1):1–36.

Farrington, C. P., Kanaan, M. N., and Gay, N. J. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2):279–295.

Funk, S., Camacho, A., Kucharski, A. J., Lowe, R., Eggo, R. M., and Edmunds, W. J. (2018). Assessing the performance of real-time epidemic forecasts: A case study of the 2013-16 Ebola epidemic. bioRxiv. `https://doi.org/10.1101/177451`.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., and Priedhorsky, R. (2014). Global disease monitoring and forecasting with Wikipedia. *PLOS Computational Biology*, 10(11):1–16.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

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

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):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(2):211–235.

Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Statistics in Medicine*, 36(22):3443–3460.

Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal*, 54(6):824–843.

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, 66(4):1295–1305.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer, Berlin.

Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., and Del Valle, S. Y. (2015). Forecasting the 2013–2014 influenza season using Wikipedia. *PLOS Computational Biology*, 11(5):1–29.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101:15124–15129.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: the `forecast` package for R. *Journal of Statistical Software*, 27(3):1–22.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Jiang, X., Wallstrom, G., Cooper, G. F., and Wagner, M. M. (2009). Bayesian prediction of an epidemic curve. *Journal of Biomedical Informatics*, 42(1):90–99.

Jordan, A., Krüger, F., and Lerch, S. (2018). Evaluating probabilistic forecasts with `scoringRules`. *Journal of Statistical Software*. In press.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton.

Lega, J. and Brown, H. E. (2016). Data-driven outbreak forecasting with a simple nonlinear growth model. *Epidemics*, 17:19–26.

Mason, S. J., Galpin, J. S., Goddard, L., Graham, N. E., and Rajartnam, B. (2007). Conditional exceedance probabilities. *Monthly Weather Review*, 135(2):363–372.

McIver, D. J. and Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLOS Computational Biology*, 10(4):e1003581.

Merkle, E. C. and Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, 10(4):292–304.

Meyer, S. (2017). ***`hhh4contacts`****: Age-structured spatio-temporal models for infectious disease counts*. R package version 0.13.0, URL `https://CRAN.R-project.org/package=hhh4contacts`.

Meyer, S. (2018). ***`HIDDA.forecasting`****: Forecasting based on surveillance data*. R package version 1.0.0, URL `https://HIDDA.github.io/forecasting`.

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

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

Moran, K. R., Fairchild, G., Generous, N., Hickmann, K., Osthus, D., Priedhorsky, R., Hyman, J., and Del Valle, S. Y. (2016). Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast. *Journal of Infectious Diseases*, 214(suppl 4):S404–S408.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):e74.

Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N., and Marathe, M. V. (2014). A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses*, 8(3):309–316.

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. J., Dredze, M., and Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Pei, S., Kandula, S., Yang, W., and Shaman, J. (2018). Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2752–2757.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ray, E. L. and Reich, N. G. (2018). Prediction of infectious disease epidemics via weighted density ensembles. *PLOS Computational Biology*, 14(2):e1005910.

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

Reich, N. G., Lauer, S. A., Sakrejda, K., Iamsirithaworn, S., Hinjoy, S., Suangtho, P., Suthachana, S., Clapham, H. E., Salje, H., Cummings, D. A. T., and Lessler, J. (2016a). Challenges in real-time prediction of infectious disease: A case study of dengue in Thailand. *PLOS Neglected Tropical Diseases*, 10(6):e0004761.

Reich, N. G., Lessler, J., Sakrejda, K., Lauer, S. A., Iamsirithaworn, S., and Cummings, D. A. T. (2016b). Case study in evaluating time series prediction models using the relative mean absolute error. *The American Statistician*, 70(3):285–292.

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., and Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLOS Computational Biology*, 11(10):e1004513.

Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334.

Shaman, J. and Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50):20425–20430.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5(5):421–433.

Steyerberg, E. (2009). *Clinical Prediction Models*. Springer, New York.

Taylor, S. and Letham, B. (2018). *prophet: Automatic forecasting procedure*. R package version 0.3.0.1, URL `https://CRAN.R-project.org/package=prophet`.

Tushar, A., Reich, N. G., Ray, E. L., and Smith, A. (2017). reichlab/flusight v2.0.0: Influenza forecasts visualizer. Zenodo. `https://doi.org/10.5281/zenodo.888171`.

Viboud, C., Boëlle, P.-Y., Carrat, F., Valleron, A.-J., and Flahault, A. (2003). Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006.

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., and Vespignani, A. (2018). The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21.

Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, 164(10):936–944.

Wei, W. and Held, L. (2014). Calibration tests for count data. *Test*, 23(4):787–805.

World Health Organization (2014). Anticipating epidemics. *Weekly Epidemiological Record*, 89(22):244.

World Health Organization, editor (2016). *Anticipating Emerging Infectious Disease Epidemics: Meeting report of WHO informal consultation*, number WHO/OHE/PED/2016.2, Geneva, Switzerland. WHO Press.

Yang, S., Santillana, M., and Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):14473–14478.