2015

Case studies in evaluating time series prediction models using the relative mean absolute error

Nicholas G Reich, *University of Massachusetts - Amherst*
Justin Lessler, *Johns Hopkins University*
Krzysztof Sakrejda, *University of Massachusetts - Amherst*
Stephen A Lauer, *University of Massachusetts - Amherst*
Sopon Iamsirithaworn, et al.

Case studies in evaluating time series prediction models using the relative mean absolute error

Nicholas G. Reich[1,*], Justin Lessler[2], Krzysztof Sakrejda[1], Stephen A Lauer[1], Sopon Iamsirithaworn[3], Derek A T Cummings[2]

[1] Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA

[2] Department of Epidemiology, Johns Hopkins University, Baltimore, MD, USA

[3] Office of Disease Prevention and Control 1, Bangkok, Thailand

[*] to whom correspondence should be addressed: nick@umass.edu

Abstract

Statistical prediction models inform decision-making processes in many real-world settings. Prior to using predictions in practice, one must rigorously test and validate candidate models to ensure that the proposed predictions have sufficient accuracy to be used in practice. In this paper, we present a framework for evaluating time series predictions that emphasizes computational simplicity and an intuitive interpretation using the relative mean absolute error metric. For a single time series, this metric enables comparisons of candidate model predictions against naïve reference models, a method that can provide useful and standardized performance benchmarks. Additionally, in applications with multiple time series, this framework facilitates comparisons of one or more models' predictive performance across different sets of data. We illustrate the use of this metric with two case studies: (1) comparing predictions of the Dow Jones Industrial Average and the NASDAQ stock indices, and (2) comparing predictions of dengue hemorrhagic fever incidence in two provinces of Thailand. These examples demonstrate the utility and interpretability of the relative mean absolute error metric in practice, and underscore the practical advantages of using relative performance metrics when evaluating predictions.

1 Introduction

Statistical prediction models play a critical role in helping people plan for the future. While the merit of evaluating predictions is widely appreciated and understood, methods implemented to evaluate predictions vary in practice.

Many statistical prediction models have a goal of predicting a single quantitative outcome, e.g. probability of 5-year cancer survival, or the number of wins of a sports team in a given season. Statistical models designed to predict the trajectory of a time series face added dimensions of complexity. For each observable unit of data (e.g. a time series observed up to a specific time), we might ask such models to predict not just one value, but a sequence of values. Additionally, if a robust and generalizable model is sought, the model must predict not just one time series effectively, but many.

These new dimensions quickly add complexity to the question of how to evaluate time series prediction models. If you are interested in evaluating predictions made at N separate time points, each at up to M time steps into the future, for L different time series, you need to make $N \cdot M \cdot L$ distinct, if correlated, predictions.

Furthermore, evaluating predictions made by time series models poses a particular set of challenges, including identifying the features of the time series you are trying to optimally predict (e.g. timing of peaks and/or troughs, cumulative totals, exact counts, etc...), accounting for correlated observations and predictions, and potentially comparing predictions from time series with different scales.

Existing research has worked to identify the pros and cons of different methods for evaluating the accuracy of time series predictions. One trend in the literature highlights advantages of using relative absolute error metrics (e.g. the relative mean absolute error, or the mean absolute scaled error) instead of squared error metrics to reduce the impact of outlying observations and to increase interpretability (Hyndman & Koehler 2006, Armstrong & Collopy 1992). In this context, several methods have been proposed to facilitate evaluation of predictions of seasonal data (see, e.g. the "naïve2" method in Makridakis & Hibon (2000)), although these methods do not appear to have been widely adopted. Additionally, the measure called "forecast skill", which relies on a relative mean

squared error calculation, has been widely used for several decades in the field of weather forecasting (Murphy 1988). Another thread of work advocates for the use of proper scoring rules for probabilistic forecasts, where the observation is evaluated against the predicted distribution (Gneiting & Raftery 2007, Czado et al. 2009, Held & Paul 2012). These methods, while having a strong theoretical foundation, are less directly comparable or interpretable and require more data (i.e. a full predictive distribution, not just a value) to be calculated.

In this paper, we present a framework for multi-step time series prediction model evaluation that emphasizes computational simplicity and an intuitive interpretation to facilitate comparisons of model performance across different time series. Specifically, we discuss the "relative mean absolute error" metric and show its utility in two distinct prediction settings: stock markets and infectious disease incidence. The relative mean absolute error (or relative MAE) is defined as the average of the absolute values of the prediction errors from one model, divided by the average of the absolute values of the prediction errors from a second model (Hyndman & Koehler 2006).

A strength of the relative MAE metric that we find particularly compelling for use in practice is how it enables standardized comparisons of candidate models with reference models. This encourages honest evaluation (i.e. a model could have very low error, but a simpler model may have similarly low error) and can help identify the strengths and weaknesses of prediction models.

Generally speaking, we conceive of reference models as being able to create reasonable (if naïve) predictions by analysts without extensive formal quantitative or statistical training. (Although we note that in practice any fitted model, simple or complex, could be used as a comparative reference.) In many fields of research or application, there may be existing standard and accepted models that would be suitable as a standard reference model.

For models predicting disease incidence (one of the examples presented in this paper), there is little standardization in the published methods used to create and evaluate predictions. In part, this reflects the wide range of scientific and planning goals in these prediction efforts, although further standardization would be valuable to the field. Appropriate reference models for predicting disease incidence could be as simple as an overall measure of central tendency (e.g. mean or median) or, for diseases that follow a seasonal pattern, an historical monthly average. For models predicting the

timing of different features of an outbreak (duration, peak, onset, etc...), reference models could be based on historical trends or trends from other nearby locations.

In Section 2, we describe a framework for facilitating comparisons of predictions for time series data. In Section 3, we present an illustrative example: predicting the daily closing values of the Dow Jones Indistrial Average and the NASDAQ stock indices. In Section 4, we present a more detailed evaluation of prediction models using a dataset with incidence of dengue fever in Thailand.

2 A generalizable metric for evaluating time series prediction models

We focus our discussion on evaluating the accuracy of time series predictions. Specifically, we are interested in summarizing a model's error for each observed value. Predicting other features of a time series may also be desirable: for example, predicting the timing of a peak, the cumulative counts, or the percentage of predictions that fall within a given percentage of the true value, to name a few. The methods defined below may be adapted for these types of metrics, although the current work focuses on implementing these methods in the context of predicting the value of as yet unobserved observations.

We consider data, $y_1, \ldots, y_T$ from a time series broken into continuous blocks labelled $k = 1, \ldots, K$. We are interested in comparing the performance of multiple models for the specific block k^*. In a prediction context, block k^* might represent data un-available at the time of fitting. Assume that we fit a suite of models to data excluding block k^* and each of these models can be used to generate predictions for any given time t. Let $\hat{\mu}^A_{t,h}$ be the (out-of-sample) predicted outcome for time t from model A, made at time $t - h$. In other words, the prediction horizon, or the number of time steps forward this prediction was made, is defined as h.

2.1 The relative mean absolute error

For a particular block of observations, the mean absolute error for model A at prediction horizon h is defined as $MAE_{A,h} = \frac{1}{|k^*|} \sum_{t \in k^*} |y_t - \hat{\mu}^A_{t,h}|$, where $|k^*|$ is the number of observations in block k^*.

Squared error metrics are commonly used in statistical model evaluation but we focus here on absolute errors as the basis for evaluating predictions, due to two distinct features of the mean absolute error metric. First, the MAE provides a very easily interpretable metric: the average error across all predictions. Squared error metrics are not as easily interpretable. Interpretability is a significant advantage when working with collaborators who are eager to understand and interpret the evaluations of the prediction models. Second, squared error metrics are sensitive to outlier prediction errors (Hyndman & Koehler 2006, Armstrong & Collopy 1992). It is known that the median minimizes expected loss when the loss function is the absolute value. This implies that the models using mean absolute error would be ranked based on the best median predictive performance, a measure that is less sensitive to outlying observations.

We define the relative mean absolute error between models A and B at horizon h as

$$relMAE_{A,B,h} = \frac{MAE_{A,h}}{MAE_{B,h}}. \tag{1}$$

This is an extension of the metric proposed by Hyndman and Koehler (2006) to account for multi-step predictions, or different prediction horizons (Hyndman & Koehler 2006). Additionally, Hyndman and Koehler specifically recommended a special form of the relative MAE, what they named the mean absolute scaled error (MASE) for use of evaluating predictions across multiple time series with different scales (Hyndman & Koehler 2006). They define the MASE as (Hyndman & Koehler 2006):

$$MASE = \frac{\sum_{t=1}^{T} |y_t - \hat{\mu}_t|/T}{\sum_{t=2}^{T} |y_t - y_{t-1}|/(T-1)}. \tag{2}$$

Heuristically, this represents the ratio of the average absolute value of the residual from the prediction model (the numerator) and the average absolute value of the residual from a naïve "reference" model. The MASE is equivalent to the relative MAE with model B taken as a simple stationary auto-regressive lag-1 (AR-1) model where the predicted value of y_t, or $\hat{\mu}_{t,h}$, is simply y_{t-1} for all values of h. However, the MASE is defined with respect to a fixed reference model which becomes meaningless when predictions are evaluated for long time horizons, particularly in periodic systems. The relative MAE avoids some of these shortcomings.

2.2 Properties of the relative MAE

The relative MAE has several desirable properties. First, the interpretation of the relative MAE for a given dataset does not depend on the scale of the data. Second, the relative MAE has an intuitive interpretation. Since the relative MAE is a ratio, a value near 1 indicates the magnitude of the two errors is approximately equal whereas a value of 2 indicates that the magnitude of prediction errors for the candidate model is twice that of prediction errors from the reference model. Third, the relative MAE is defined when there are zeroes in the data, unlike several other popular accuracy metrics (Hyndman & Koehler 2006).

Fourth, as defined above, the relative MAE easily accomodates settings with long prediction horizons, or when predictions are made many steps ahead into the future. When predictions are desired far into the future, the reference model for the original MASE will not evaluate the predictions in a meaningful way because it only considers one-step-ahead prediction errors. Therefore, the flexibility of the reference model definition in the relative MAE and of the extension to multiple prediction horizons, makes this a more useful and relevant metric in evaluating multi-step predictions of time series data. Below, in Sections 3 and 4, we systematically explore the performance of the relative MAE using two time series applications that require predictions at multiple steps forward into the future.

Finally, we also observe that the relative MAE lends itself to comparisons between any two models, not necessarily just a reference and a set of candidate prediction models. For example, a simple but important property of the relative MAE is that

$$relMAE_{A,B,h} = \frac{relMAE_{A,C,h}}{relMAE_{B,C,h}} \tag{3}$$

indicating that as long as a common reference model is used in two model comparisons (e.g. A vs. C and B vs. C), the relative MAE can be computed between the two models that were not explicitly compared (e.g. A vs. B).

2.3 Using scaled observations with relative MAE

Since data from predictions may be skewed, we discuss briefly the impact of scaling the y and $\hat{\mu}$ inputs before the calculations are made. (Note: in this section we supress the prediction horizon notation for simplicity.) The choice of scaling functions for the MAE calculation affects the interpretation of both the MAE and the relative MAE. Simply plugging in the data transformations used in model estimation (i.e. using $|\log y - \log \mu|$ in the MAE) leads to arbitrary implicit loss functions in the model evaluation. This choice should instead be made based on an implicit or explicit loss function and it leads to a generalized MAE defined as

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |f(y_t) - f(\hat{\mu}_t)| \tag{4}$$

where the function $f(\cdot)$ is some function that transforms both the predicted and observed values.

This paper so far considers all errors on the original scale of the data where f is the identity function. The loss function implied by this choice is one that emphasizes absolute differences. It implies that errors of a given magnitude are of equal importance whether the time series observations are near zero or near a large value—cost must be constant regardless of the scale.

An alternative cost function might emphasize relative error compared to a practically or scientifically meaningful reference point. In this case the cost of an error would decrease with distance from the reference point. For example a $5,000 error in the predicted value of a $10,000 stock portfolio might be a disaster whereas the same $5,000 error in a million dollar portfolio would not be meaningful.

For an implicit relative cost function with count data, we suggest calculations of the MAE on the log-scale. To do this for count data —or for other data with a meaningful baseline value— we define f in the MAE calculations as $f(x) = \log(x - b + 1)$ where b is the baseline. For count data, where $b = 0$, the modified the calculations of the mean absolute error are

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |\ln(y_t + 1) - \ln(\hat{\mu}_t + 1)| = \sum_{t=1}^{T} |\ln(\frac{y_t + 1}{\hat{\mu}_t + 1})|$$

.

With this MAE calculation, if either $\frac{y_t+1}{\hat{\mu}_t+1} = C$ or $\frac{y_t+1}{\hat{\mu}_t+1} = \frac{1}{C}$, then the mean absolute error for that observation will be the same, no matter what y_t or μ_t are. For example, this means that $(y_t, \mu_t) = (100, 110)$ has the same contribution to the MAE as $(y_t, \mu_t) = (10, 11)$ or $(y_t, \mu_t) = (11, 10)$, or any (y_t, μ_t) such that $\frac{y_t}{\mu_t} = 1.1$ or $\frac{\mu_t}{y_t} = 1.1$. Whether this is approriate depends on the application-specific implicit loss function.

3 Illustrative example: Predicting the stock market

As an example to illustrate the use of the relative MAE, we consider a series of simple models to predict the daily closing value of the Dow Jones Industrial Average (DJIA, 2014 mean/sd: 16778 / 553) and the NASDAQ Composite Index (NASDAQ, 2014 mean/sd: 4375 / 212) in 2014. For each day of trading in 2014, the model predicted the closing value for that day as well as the 29 subsequent days of trading (i.e. roughly six weeks of activity). We focus on three models that use averages of recent closing values for predictions; and we compare the performance of these models at different time horizons.

Historical data on the DJIA (from 1 January 2006 to the time of analysis) and the NASDAQ (from 05 February 1971 to the time of analysis) were obtained using the `quantmod` package for R from the Federal Reserve Economic Data system (Ryan 2014, Federal Reserve Bank of St Louis 2015*b*,*a*). This manuscript was dynamically typeset using `knitr` and R version 3.2.1 (2015-06-18) (Xie 2015, Ihaka & Gentleman 1996). The data for 2014 is shown in the left panels of Figure 1.

We defined a sequence of three simple models to predict the daily stock market closing values for both the DJIA and the NASDAQ. Each of these three models took an average of recent daily closing values as observed at day t (i.e. the 5-day average close model averages across days $t-1, t-2, ..., t-5$) and used that observed mean as the predicted closing value for day t and the subsequent 29 days. We define $\hat{\mu}_{t,h}$ as the h-step ahead predicted value of y_t, i.e. the prediction of y_t made at time $t-h$. The first model predicted for every subsequent value the last observed close; the second model used the mean closing value over the past 5 days; and, the third model used the mean closing value over the past 20 days. We refer to these models as $D1$, $D5$, and $D20$ when predicting

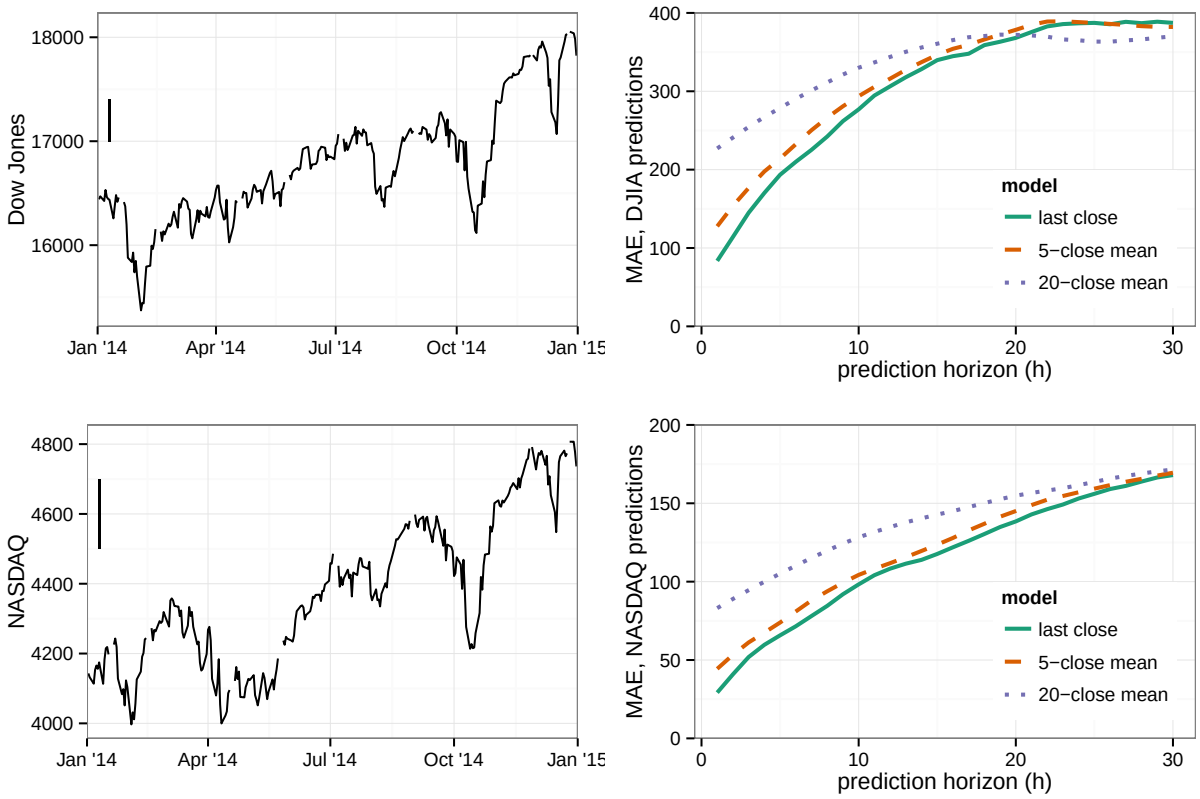


Figure 1: Sample of the two stock-market time-series, and the prediction errors at different forecast horizons. Left-hand panels: the Dow Jones Industrial Average (top) and NASDAQ (bottom) daily closing values for each day in 2014. The vertical black bar shows the height of the y-axis from the panels on the right hand side, in an attempt to illustrate the relative scale of the errors against the raw data. Right-hand panels: the mean absolute error (MAE) plotted for three models as a function of the forecast horizon, or the number of steps into the future that a prediction was made for.

the DJIA and $N1$, $N5$, and $N20$ when predicting the NASDAQ. The MAE are shown as a function of the prediction horizon, h, in the right hand panels of Figure 1.

In general, increasing the averaging window of closing values (from 1 to 5 to 20 days) reduced our ability to make accurate predictions in the short term. This pattern is seen clearly in Figure 1 and is a consequence of the high auto-correlation (over 0.5 at lags shorter than 20 days) in these time series. In other words, since there is auto-correlation in these time series, merely predicting the previous value will be in general a good prediction of the next value. However, this is less true at longer prediction horizons.

In evaluating the DJIA prediction models, we observed one step-ahead (model $D1$) mean absolute error of $MAE_{D1,1} = 83.55$. The relative MAE for the 5-day average close (model $D5$) predicting 1-day ahead was $relMAE_{D5,D1,1} = \frac{MAE_{D5,1}}{MAE_{D1,1}} = 1.53$, showing we did quite a bit worse at predicting the next day's close by using the 5 day average compared to the 1 day average (about 1.5 times the error). Looking at an even longer term average of 20 days (D20) further worsened the short term predictions. We observed $relMAE_{D20,D1,1} = 2.72$ looking 1 day into the future. This was substantially worse than model D5, calculated as $relMAE_{D20,D5,1} = 1.78$. So the 20 day average model had 78 percent more error than the 5 day average model when predicting the DJIA close 1 day into the future. Of note is that we can calculate this quantity either by directly dividing the MAE for D20 by the MAE for D5, or by dividing $\frac{MAE_{D20,1}}{MAE_{D1,1}}$ by $\frac{MAE_{D5,1}}{MAE_{D1,1}}$.

For longer term predictions, all of the models had roughly equivalent errors. The 30 step-ahead absolute error of the most recent close model (D1) was $MAE_{D1,30} = 387.29$. The relative MAE comparing the 20-day average close model (D20) and the most recent close model (D1) was $relMAE_{D20,D1,30} = 0.96$. This indicates that the absolute prediction errors for these models had approximately equal magnitudes when predicting closing values at longer horizons. Similarly, the 5-day average close model (D5) showed similar absolute prediction errors at 30-day horizons, with $relMAE_{D5,D1,30} = 0.99$.

We compared the performance of these three models in predicting the NASDAQ, and observed similar results. Model $N1$ showed the best short-term performance compared with the $N5$ and $N20$, a margin that decreased as the prediction horizon increased. Comparing the last-close model (N1) to a

five-day average close model (N5), in predicting a single day ahead we observed a relative MAE of $relMAE_{N5,N1,1} = 1.52$. And in predicting 30 days ahead with both the 5-day (N5) and 20-day (N20) average models we observed $relMAE_{N5,N1,30} = 1.01$ and $relMAE_{N20,N1,30} = 1.02$.

Despite the DJIA and NASDAQ data being separate time series during 2014, each with their own distinct scale, these relative error metrics allowed us to make direct comparisons between the model predictions on these two datasets. For example, predictions of the closing value of the DJIA from a 20-day moving average model in 2014 on average were 4% closer to the true value than a prediction using the most recently observed closing value. Similarly, predictions of the closing value of the NASDAQ from a 20-day moving average model in 2014 on average were 2% further away from the true value than a prediction using the most recently observed closing value. This demonstrated that for both of these time series, our 20-day moving average model did not add value to longer term predictions of stock market performance over a simple reference model.

4 Scientific example: Predicting infectious disease incidence

4.1 Infectious disease prediction models show disparate evaluation methods

The spread of infectious disease is a dynamic process driven by many biological and social factors. While our knowledge of these biological mechanisms (e.g. infectiousness, pathogen-on-pathogen and pathogen-environment interactions) and social/behavioral patterns (e.g. networks of social contacts and travel) has grown significantly in recent decades, predicting infectious disease patterns remains a challenging task.

We reviewed a small, non-random sample of peer-reviewed publications that focused on predicting infectious disease outbreaks. A scoping review on published research on predicting the spread of influenza showed that the types of predicted outcomes in these studies vary widely. They have included weekly or daily incidence, cumulative incidence, timing of peak incidence, timing of epidemic onset, and length of the epidemic season (Chretien et al. 2014). In one example, Shaman et al. describe a model used to prospectively predict the peak of seasonal influenza outbreaks in the U.S.

(Shaman & Karspeck 2012, Shaman et al. 2013). Retrospective evaluation showed mixed performance across a wide range of municipalities, but in many regions, predictions of the peak timing of the outbreak were accurate to within one week. As a reference comparison, they used a model that chose resampled historical peaks, and their method outperformed this reference model, with increasingly better relative performance as the flu season progressed.

In forecasting outbreaks of dengue fever, the predicted outcomes have also varied, and have included timing of and duration of outbreak and weekly incidence. In a series of papers using machine learning methods to predict outbreaks of dengue fever in the Phillipines, Buczak et al. followed a rigorous train/validate/test evaluation protocol, but used different methods for defining testing and validation sets across their two papers on this topic (Buczak et al. 2012, 2014). They did not compare their results to a reference prediction model. Hii et al. (2012) attempted to predict dengue outbreaks in Singapore. Their model predicted weekly incidence and their final model had very close correlation with the actual data. The authors made no reference model comparisons, and among many models fit to the data, featured the results from a single model that retrospectively showed good predictive performance.

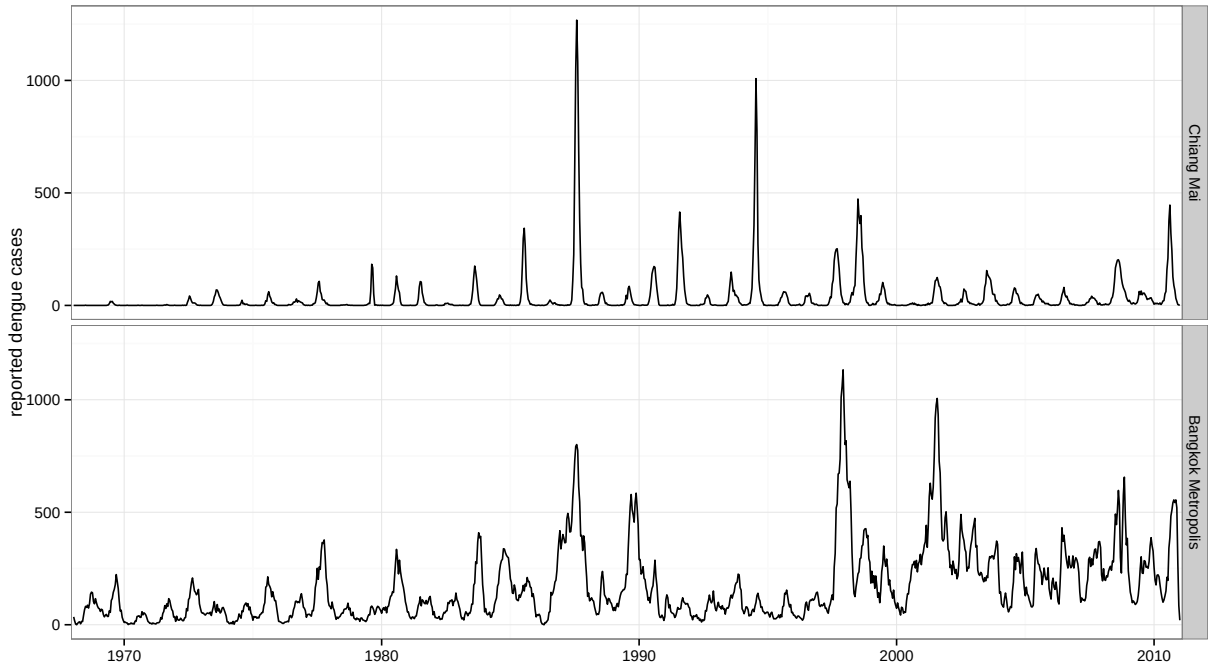
4.2 Prediction models for dengue hemorrhagic fever in Thailand

Dengue is a mosquito-borne virus that causes fever, rash, and in severe cases, internal bleeding and organ failure. Around 2.5 billion individuals on the planet live in regions where dengue is endemic (World Health Organization 2015). Dengue is carried by mosquitoes that thrive in hot and rainy weather. Therefore, in many regions where dengue circulates it exhibits a strong seasonal pattern (Campbell et al. 2013, Johansson et al. 2009).

From the Thailand Ministry of Public Health surveillance system, we obtained incidence case reports of dengue hemorrhagic fever (a severe form of dengue) for each of the 72 provinces in Thailand from January 1, 1968 to December 31, 2010. Data were aggregated and/or disaggregated (depending on the reporting scale of the raw data) to biweekly intervals. (Biweeks represent two week intervals, and are based on consistently defined 14 or 15 day intervals of time within each year.) For the purposes of this prediction exercise, we focused on making predictions for just two provinces, Bangkok and Chiang

Mai, although data from other provinces were considered as possible covariates in the prediction models (see details below). The 2010 census estimates for population in those two provinces were 8,249,117 and 1,708,564 for Bangkok and Chiang Mai, respectively (National Statistical Office of Thailand 2011). The total number of reported cases per province over the 43 years was 185,927 (Bangkok) and 35,938 (Chiang Mai). We show the complete time series in Figure 2.

Figure 2: Reported cases of dengue hemmorhagic fever for Bangkok and Chiang Mai between 1968 and 2010.



We implemented relative MAE comparisons using three different reference models and a candidate Poisson regression model for dengue hemorrhagic fever in the Thai provinces of Bangkok and Chiang Mai.

Poisson model using data from correlated provinces

We define the number of cases with onset at time t in province i as $Y_{t,i}$. Below we adopt the convention of refering to $Y_{t,i}$ as the unobserved random variable that is being modeled and $y_{t,i}$ as

observed values that may be used as covariates in the model. The model assumes that

$$Y_{t,i} \quad \sim \quad Poisson(\lambda_{t,i} \cdot [y_{t-1,i} + 1]) \tag{5}$$

where the lag-1 term $y_{t-1,i}$ is treated as an offset in this model. This formulation assumes that the model for the expected number of cases at time t can be represented by multiplying the number of cases observed at the prior time-step ($y_{t-1,i}$) by a "reproductive rate" of cases ($\lambda_{t,i}$). The explicit model below for $\lambda_{t,i}$ facilitates an intuitive interpretation: if $\lambda_{t,i} < 1$ then the number of cases is expected to decrease and if $\lambda_{t,i} > 1$ then the number of cases is expected to increase.

We explicitly modeled the expected number of cases as a generalized additive model (i.e. a generalized linear model estimated by penalized maximum likelihood) (Hastie & Tibshirani 1990)

$$\log(\lambda_{t,i} \cdot [y_{t-1,i} + 1]) \quad = \quad f_i(t) + \sum_{j \in \mathcal{C}_i} \alpha_j \log \frac{y_{t-1,j} + 1}{y_{t-2,j} + 1} + \log(y_{t-1,i} + 1) \tag{6}$$

where $f_i(t)$ is assumed to be a province-specific cyclical cubic spline and $\mathcal{C}_i$ is the set of the 3 most correlated provinces with province i (possibly including province i itself) at a one biweek lag across the entire dataset. With $\lambda_{t,i}$ as a function of the lag-1 and lag-2 terms, we have adapted the structure of an ARIMA(0,1,0) model, using a difference at 1-lag on the log scale. This captures the direction of transmission intensity (is it increasing or decreasing?) across several different locations. We note that this model is one specific parameterization of a general class of ARIMA-style models that consider different numbers of correlated provinces (i.e. not just 3) and different numbers of lag-times as predictors of the current incidence in province i at time t. The goal of this modeling exercise is to demonstrate the utility of the relative MAE metric in evaluating predictions from these and other similar time series modeling examples. A more thorough exploration of our example's model space is performed in other work (in preparation).

The autoregressive terms in the model for $\lambda_{t,i}$ approximate the reproductive rate for province j at time $t-1$, and are designed to capture the slope of recent incidence in the correlated provinces. The addition of the value 1 in the numerator and denominator ensures that the quantities are defined when zero case counts are observed. This method of adjusting for zero counts has been discussed at length,

with the interpretation of an "immigration rate" added to each observation (Zeger & Qaqish 1988).

Auto-regressive lag 1 (AR-1) model

The first reference model was a simple AR-1 model used in the definition of the mean absolute scaled error, described above. When making a h-step ahead prediction for time t using data up to time $t-h$, the predicted value was $\hat{\mu}_{t,h} = y_{t-h}$. Note that this meant that if we were generating a prediction for 13 biweeks (half a year) into the future, the predicted value was the most recently observed value. We observe that for the AR-1 reference model, the predicted value for y_t changes depending on when the prediction is made. For example, if a one-step-ahead prediction is made for time t' at time $t'-1$, the predicted value for the AR-1 reference model would be $\hat{\mu}_{t',1} = y_{t'-1}$. A two-step-ahead prediction for the same timepoint would yield a predicted value of $\hat{\mu}_{t',2} = y_{t'-2}$.

Seasonal medians model

The second reference model predicted a median seasonal value, so $\hat{\mu}_t = \text{median}(y_{\mathcal{S}_t})$. In this model, the median is calculated across all values of t that fall in the set $\mathcal{S}_t$ which contains all times in the training data with the same season as t. The seasonal reference model is time-invariant: the predictions for a particular time t' are the same no matter when the prediction occurs.

Overall median model

The third reference model predicted an overall median value, so $\hat{\mu}_t = \text{median}(y_t)$ where the median is calculated across all times in the training data. This overall median reference model is time-invariant: the predictions for time t' are the same no matter when the prediction occurs.

4.3 Model training and validation

We implemented a leave-one-year-out cross-validation procedure to create out-of-sample predictions for the mean absolute error (MAE) calculations. For all predictions, data from 1968 through 1999

was included in the training dataset. For each year from 2000 to 2009, a single full calendar year of case data was left out in turn from the training dataset and the model was fit to the remaining case data. This ensured that all predictions were made based on the same amount of training data (42 years of case data). For every biweek in the year that was left out from the training dataset, we made a set of h step-ahead predictions, where h ranged from 1 biweek to 13 biweeks (about 6 months). To construct an h step-ahead prediction for biweek t, we assumed case data up to biweek $t-h$ were complete. We then sequentially predicted case counts for the following h biweeks, up through time t. The resulting predicted case counts ($\hat{\mu}_{t,h}$) were then compared to the final, observed case count y_t in computing the MAE. Figure 4 shows the MAE calculated on the predictions from the 10 years of out-of-sample cross-validation.

These models were fit to the data using the `mgcv` package for the R statistical programming language (Wood 2011).

4.4 Results from model comparisons give insight into time series predictions

We evaluated model performance using relative MAE. Based on our previous work and our knowledge on this topic, we expected that the candidate Poisson model presented above should beat an average prediction model for the seasonal dengue incidence data. For some (if not most) provinces, we expected it to provide short- and long-term predictions that would be better than just guessing the seasonal mean. For short-term predictions, we expected the model to do better in many cases than a simple auto-regressive model. These kinds of knowledge and statements about our model informed our choices of reference models to compare with our candidate model.

Predictions of dengue fever incidence in Bangkok showed mixed results at different time scales. A sample of predictions for time points in 2005 are shown in Figure 3. This plot provides a snapshot of how each of the four models performed at one specific timepoint. Summaries of the predictions across all years are shown in Figure 4. This plot provides an overall evaluation of how the models performed across all timepoints.

The errors for predictions from the AR-1 reference model in both provinces monotonically increased

Figure 3: Example of cases and prediction errors for Bangkok for a single timepoint. This figure shows the predictions made for 1 to 13 biweeks into the future at biweek 7 in 2005. This is merely a sample of the predictions made, as predictions like these were made for each biweek in 2000 through 2009. Black bars indicate the number of dengue hemorrhagic fever cases observed when predictions were made. Grey bars indicate cases that were subsequently observed. The four lines represent predictions from the four models: seasonal medians (green circles), AR-1 (orange triangles), overall median (purple squares), and our candidate Poisson model (pink crosses).

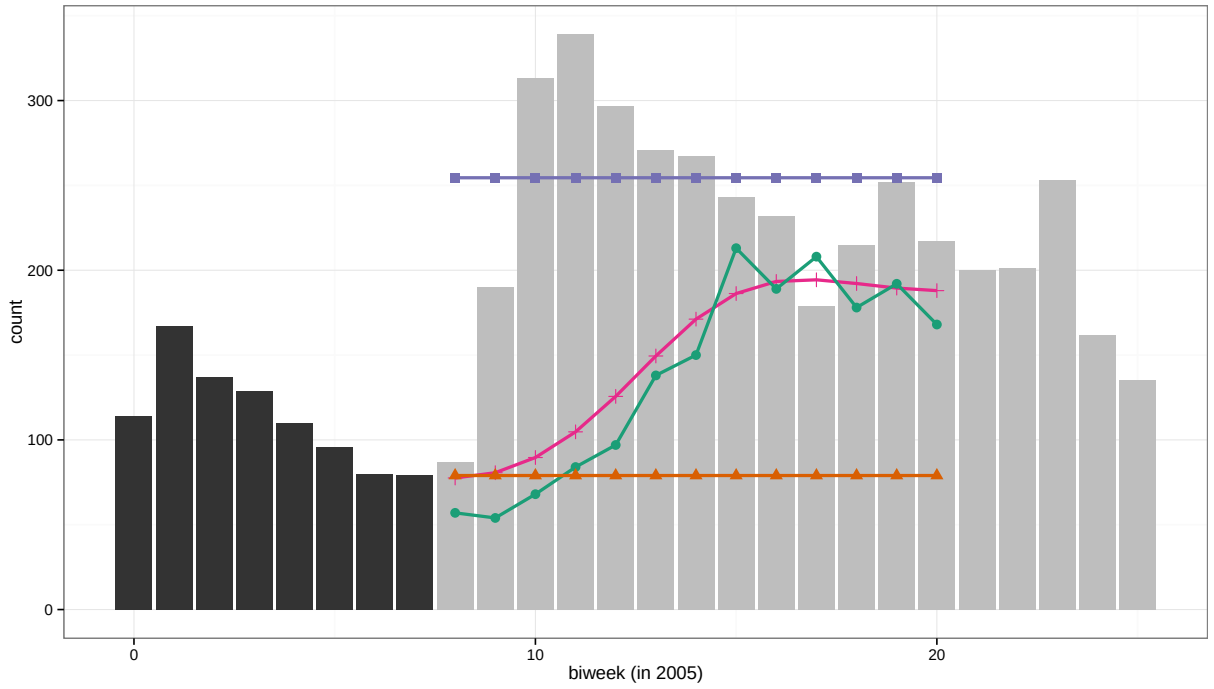
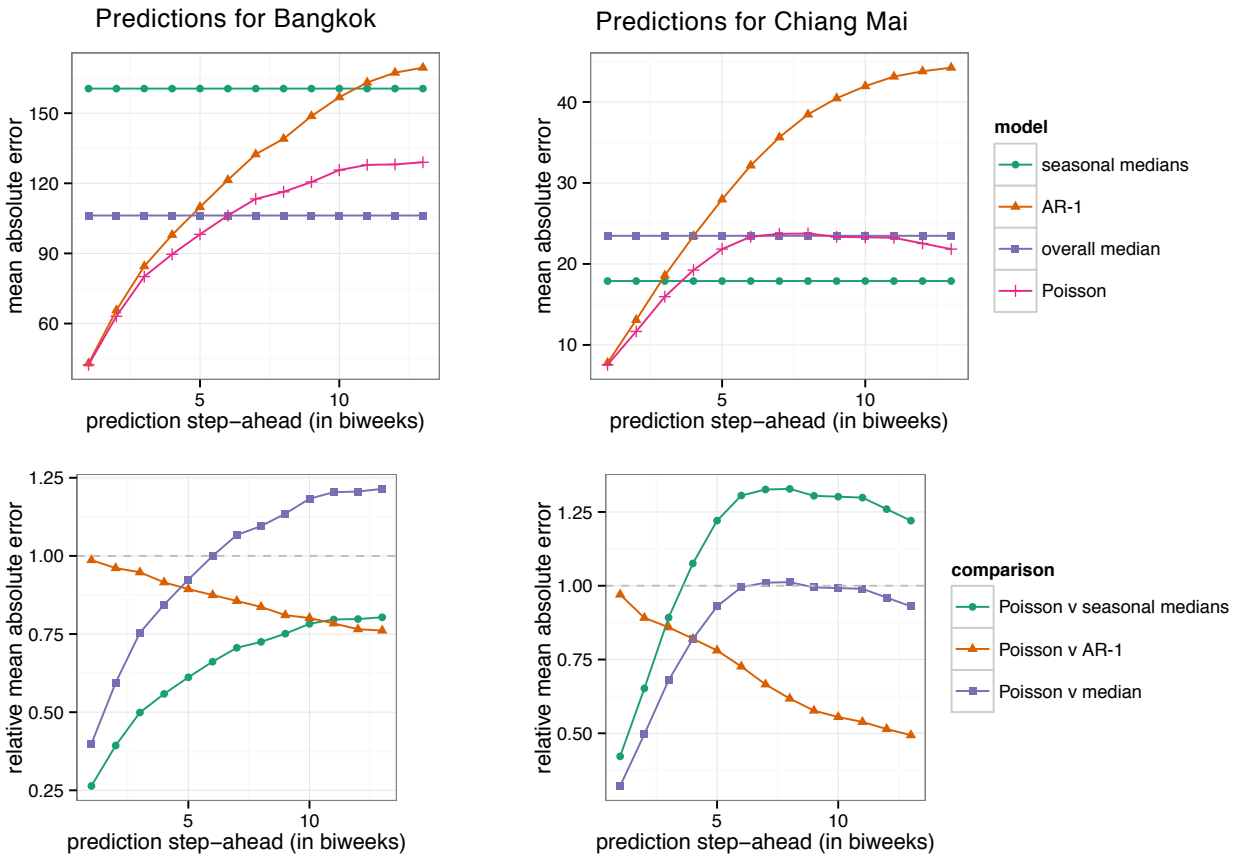


Figure 4: Mean absolute error (MAE) and relative mean absolute error (relative MAE) scores for Bangkok (left column) and Chiang Mai (right column) based on 10 years of cross-validated predictions. These metrics show the errors calculated not just for 2005 (as in Figure 3, but for all cross-validated years (2000-2009). The top two sub-figures show the mean absolute error for each model. Errors are shown for the candidate Poisson model (pink), historical medians model (green), AR-1 model (orange), and overall median model (green). The bottom two sub-figures show the relative MAE values comparing the candidate Poisson regression model to the seasonal median, overall median, and AR-1 reference models.



as the predicted time point moved further into the future. For Bangkok, the mean absolute error for the AR-1 model was 42.9 cases predicting one biweek ahead, and 169.5 cases predicting six months (13 biweeks) ahead. For Chiang Mai, the mean absolute errors for the AR-1 model were 7.8 cases and 44.2 cases for predicting two weeks and six months ahead, respectively.

Since the seasonal and overall median predicted values do not depend on recently observed values they always make the same prediction for a given timepoint. This results in the prediction errors being constant for a specific observation across prediction horizons. In Bangkok, the simple seasonal model produced predictions that were on average 51% further from the observed value than a model that predicted the median observed value for every observation ($\frac{MAE_{seasonal}}{MAE_{median}} = 1.51$). In Chiang Mai, this pattern was reversed, as the seasonal model had predictions that were on average 24% closer to the observed value ($\frac{MAE_{seasonal}}{MAE_{median}} = 0.76$). This is likely reflective of the stronger seasonal patterns of dengue in Chiang Mai, visible in Figure 2.

The Poisson model in each province had nearly equivalent performance to the AR-1 model predicting one biweek ahead ($\frac{MAE_{Poisson,1}}{MAE_{AR1,1}} = 0.99$ for Bangkok, $\frac{MAE_{Poisson,1}}{MAE_{AR1,1}} = 0.97$ for Chiang Mai). The Poisson models showed much less relative error than the AR-1 model when predicting six months ahead ($\frac{MAE_{Poisson,13}}{MAE_{AR1,13}} = 0.76$ for Bangkok, $\frac{MAE_{Poisson,13}}{MAE_{AR1,13}} = 0.49$ for Chiang Mai).

In Bangkok, across all the years studied, we observed that our Poisson model consistently outperformed both the AR-1 and seasonal median reference model, as shown by relative MAE scores below 1 for both reference model comparisons at all prediction steps ahead (see Figure 4). However, at longer prediction horizons (6 biweeks and above, for Bangkok) simply predicting an overall median provided more accuracy than any of the models considered. Predicting four biweeks out, the Poisson model achieved its best relative performance compared to all models, with predictions at least 8% closer to the truth on average than any other model. In Chiang Mai, the Poisson model made better predictions than every model only for the first three biweeks, after which the seasonal model made better predictions. This analysis shows that while in both locations our candidate Poisson models provided marginal improvement in predicting dengue in the medium-term (1-3 months), these models had equivalent or worse performance than other reference models at short and longer prediction horizons.

5 Discussion

We have shown that using the relative mean absolute error framework described above to compare prediction models can have several important advantages. First, using metrics that do not benchmark performance against a reference model (such as an auto-regressive model, or the average of recent observations) can lead to overstating the added value of predictions, even when accepted methods for evaluation are used. For example, if a candidate prediction model has very low cross-validated mean squared error that is in general a good thing. But if a simple auto-regressive model can achieve the same score, then the candidate model may not have much value. Second, comparisons against different types of reference models can help identify the strengths and weaknesses of prediction models. Third, as shown in both of our examples, using this metric also facilitates comparisons of similar modeling techniques between two different time series. This property of the relative MAE makes it particularly conducive to comparisons designed to evaluate generalizability of a given modeling approach. Finally, these comparisons can demonstrate the value of simple modeling efforts, leading to improved predictions at a lower cost. This may be especially true if the methods or data used for complex predictions are time- and/or resource-intensive.

In our infectious disease application (see Section 4), we observed similar overall patterns of errors and model comparisons for all models but the Poisson when comparing relative MAE values calculated using on the log scale and on the original data scale. We observed that calculating the MAE on the log-scale reduced the relative error of the Poisson model disproportionately among the four models considered. (Data not shown.) This may reflect the fact that the model was optimized and estimated on the log scale. This further underscores the importance of understanding the context in which the predictions are being made, and choosing a scaling method that reflects the loss functions used by decision makers.

Evaluating model predictions against different reference models can provide valuable information about model performance, especially at different time-scales. Comparing the mean absolute error for different models can tell us which model made better predictions. For example, we saw in both examples that reference models based on recent observations served as good benchmarks for other models when making short-term predictions. When making longer term predictions, models that

accounted for seasonal trends or longer-term moving averages did not always improve our models' predictive ability.

We see several concrete benefits of using the absolute error instead of the squared error (a more traditional choice in statistical error evaluation) as the basis for these comparisons. First, as others have observed, the absolute error is less prone to being influenced by several outlying points or observations. Second, we find the interpretability of the relative mean absolute error to be particularly compelling, especially in a context where the results need to be explained to a non-quantitative audience. For example, the relative MAE allows us to say that "on average, predictions from model A were p% closer to observed values than predictions from model B". No other metric that we know of provides this intuitive of an interpretation.

In the context of larger goals of developing models for infectious disease prediction in the era of "big data" (Hay et al. 2013), developing a standardized way of measuring and evaluating forecasts may play an increasingly important role. Relative metrics could serve as a cornerstone of these efforts, as they enable simple comparisons of prediction accuracy in different settings.

While we have focused on the advantages of using the relative mean absolute error as a metric to evaluate forecasts, there are some limitations and caveats that we must also present. The relative MAE focuses its model evaluation on point predictions. Other methods for prediction evaluation (such as scoring rules) allow for evaluation based on a full predictive probabilistic distribution, which take into account the uncertainty in the predictions. These methods could play an important role in distinguishing between prediction models. Additionally, understanding how model or data variability impacts the interpretation of the relative MAE would be a valuable contribution to this area of research. It would be possible to extend the relative MAE metric to include confidence intervals based on the model uncertainty.

In conclusion, we recommend the use of the relative mean absolute error metric to evaluate and compare time series predictions from both simple and complex models. This approach could assist decision makers in a wide range of settings, who need to understand and quantify the value of a multiple sets of predictions.

Funding

NGR, JL, KS, SAL, and DATC were funded by NIH NIAID grant 1R01AI102939. Additionally, NGR and DATC report funding from NIH NIAID grant R21AI115173.

References

Armstrong, J. S. & Collopy, F. (1992), 'Error measures for generalizing about forecasting methods: Empirical comparisons', *International Journal of Forecasting* **8**(1), 69–80.

Buczak, A. L., Baugher, B., Babin, S. M., Ramac-Thomas, L. C., Guven, E., Elbert, Y., Koshute, P. T., Velasco, J. M. S., Roque, Jr, V. G., Tayag, E. A., Yoon, I.-K. & Lewis, S. H. (2014), 'Prediction of High Incidence of Dengue in the Philippines', *PLOS Neglected Tropical Diseases* **8**(4), e2771.

Buczak, A. L., Koshute, P. T., Babin, S. M., Feighner, B. H. & Lewis, S. H. (2012), 'A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data', *BMC Medical Informatics and Decision Making* **12**(1), 124.

Campbell, K. M., Lin, C. D., Iamsirithaworn, S. & Scott, T. W. (2013), 'The Complex Relationship between Weather and Dengue Virus Transmission in Thailand.', *American Journal of Tropical Medicine and Hygiene* **89**(6), 1066–1080.

Chretien, J.-P., George, D., Shaman, J., Chitale, R. A. & McKenzie, F. E. (2014), 'Influenza Forecasting in Human Populations: A Scoping Review', *PloS one* **9**(4), e94130.

Czado, C., Gneiting, T. & Held, L. (2009), 'Predictive Model Assessment for Count Data', *Biometrics* **65**(4), 1254–1261.

Federal Reserve Bank of St Louis (2015*a*), 'NASDAQ OMX Group, *NASDAQ Composite Index*© [NASDAQCOM]'.
URL: *https://research.stlouisfed.org/fred2/series/NASDAQCOM*

Federal Reserve Bank of St Louis (2015*b*), 'S&P Dow Jones Indices LLC, *Dow Jones Industrial Average* [DJIA]'.
URL: *https://research.stlouisfed.org/fred2/series/DJIA/*

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, CRC Press.

Hay, S. I., George, D. B., Moyes, C. L. & Brownstein, J. S. (2013), 'Big data opportunities for global infectious disease surveillance.', *PLOS Medicine* **10**(4), e1001413.

Held, L. & Paul, M. (2012), 'Modeling seasonality in space-time infectious disease surveillance data - Held - 2012 - Biometrical Journal - Wiley Online Library', *Biometrical Journal* .

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C. & Rocklöv, J. (2012), 'PLOS Neglected Tropical Diseases: Forecast of Dengue Incidence Using Temperature and Rainfall', *PLoS neglected tropical ...* .

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

Ihaka, R. & Gentleman, R. (1996), 'R: A language for data analysis and graphics', *Journal of computational and graphical statistics* **5**(3), 299–314.

Johansson, M. A., Cummings, D. A. T. & Glass, G. E. (2009), 'Multiyear climate variability and dengue–El Niño southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: a longitudinal data analysis.', *PLOS Medicine* **6**(11), e1000168.

Makridakis, S. & Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting* .

Murphy, A. H. (1988), 'Skill Scores Based on the Mean-Square Error and Their Relationships to the Correlation-Coefficient', *Monthly Weather Review* **116**(12), 2417–2425.

National Statistical Office of Thailand (2011), Preliminary Report The 2010 Population and Housing census (Whole Kingdom), Technical report.

Ryan, J. A. (2014), 'quantmod: Quantitative Financial Modelling Framework', *R package version 04-2* .

Shaman, J. & 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.

Shaman, J., Karspeck, A., Yang, W., Tamerius, J. & Lipsitch, M. (2013), 'Real-time influenza forecasts during the 2012–2013 season', *Nature Communications* **4**.

Wood, S. N. (2011), 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society Series B-Statistical Methodology* **73**(1), 3–36.

World Health Organization (2015), 'Dengue and severe dengue'.
URL: *http://www.who.int/mediacentre/factsheets/fs117/en/*

Xie, Y. (2015), *Dynamic Documents with R and Knitr, Second Edition*, CRC Press.

Zeger, S. L. & Qaqish, B. (1988), 'Markov regression models for time series: a quasi-likelihood approach.', *Biometrics* **44**(4), 1019–1031.