Epidemics 28 (2019) 100345

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Complementing the power of deep learning with statistical model fusion: Probabilistic forecasting of influenza in Dallas County, Texas, USA

Marwah Soliman[a,1], Vyacheslav Lyubchich[b,*], Yulia R. Gel[a]

[a] Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA
[b] Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD, USA

ARTICLE INFO

Keywords:
Epidemics
Beta regression
ARIMA
LASSO
Bayesian model averaging

ABSTRACT

Influenza is one of the main causes of death, not only in the USA but worldwide. Its significant economic and public health impacts necessitate development of accurate and efficient algorithms for forecasting of any upcoming influenza outbreaks. Most currently available methods for influenza prediction are based on parametric time series and regression models that impose restrictive and often unverifiable assumptions on the data. In turn, more flexible machine learning models and, particularly, deep learning tools whose utility is proven in a wide range of disciplines, remain largely under-explored in epidemiological forecasting. We study the seasonal influenza in Dallas County by evaluating the forecasting ability of deep learning with feedforward neural networks as well as performance of more conventional statistical models, such as beta regression, autoregressive integrated moving average (ARIMA), least absolute shrinkage and selection operators (LASSO), and non-parametric multivariate adaptive regression splines (MARS) models for one week and two weeks ahead forecasting. Furthermore, we assess forecasting utility of Google search queries and meteorological data as exogenous predictors of influenza activity. Finally, we develop a probabilistic forecasting of influenza in Dallas County by fusing all the considered models using Bayesian model averaging.

1. Introduction

Influenza is one of the main causes of morbidity and mortality with up to five million cases of severe illness and 650,000 deaths worldwide each year (WHO, 2018). In the United States, the Centers for Disease Control and Prevention (CDC) report between 12,000 (during 2011–2012) and 56,000 deaths (during 2012–2013) per influenza season in 2010–2016 (Rolfes et al., 2016). Seasonal influenza significantly elevates the patient volume that stresses the health care system and contributes to emergency department (ED) crowding, which is a major cause of delays in critical treatments and increased mortality. An influenza pandemic presents a well recognized serious threat to the United States health care infrastructure (CDC, 2017).

Management actions can reduce the impact of influenza considerably. For example, the CDC reported about 5.1 million influenza like illness (ILI) cases and 3000 deaths were prevented by vaccination in the 2015–2016 influenza season (Rolfes et al., 2016). Effective management of both seasonal and pandemic influenza requires early detection of the outbreak through timely and accurate surveillance linked with a rapid response to reduce crowding (Aronis et al., 2017; Lee and Chuh, 2010; Preis and Moat, 2014; Shaman and Kohn, 2009; Spreco et al., 2017b; Spreco and Timpka, 2016).

One of the key obstacles for developing operational and reliable near-real time influenza forecasts is the lack of the most recent records on influenza activity. Indeed, even the CDC data for well monitored areas in the USA are typically two weeks behind, since time is required to confirm ILI. Two weeks is a substantial period in terms of influenza transmission. As a result, recently there has been a spark of interest in using non-traditional data from various online social media, starting with the Google Flu Trend project (discontinued in 2015; Ginsberg et al., 2009) and followed by Twitter, Wikipedia, Yahoo, and many other web search engines (Alessa and Faezipour, 2018; Allen et al., 2016; McIver and Brownstein, 2014; Polgreen et al., 2008; Santillana et al., 2015; Sharpe et al., 2016; Spreco et al., 2017a; Dugas et al., 2013). While the epidemiological data from online social media are subject to criticism due to sensitivity to fickle media interest and, hence, require calibration with offline information sources, there exists an ever growing interest in adaptive fusion of online and offline data for

* Corresponding author.
E-mail address: lyubchich@umces.edu (V. Lyubchich).

[1] This paper received the American Statistical Association (ASA) 2019 Best Student Paper Award jointly administered by three ASA Sections: the Government Statistics Section, Survey Research Methods Section, and Social Statistics Section.

https://doi.org/10.1016/j.epidem.2019.05.004
Received 6 October 2018; Received in revised form 8 March 2019; Accepted 6 May 2019
Available online 08 June 2019

enhanced near-real time epidemiological forecasting.

Furthermore, most of current epidemiological studies are based on conventional statistical models (see overviews by Biggerstaff et al., 2016, 2018; Paul and Dredze, 2017, and references therein), while utility of modern deep learning tools for infectious disease forecasting, in particular, predicting of influenza activity, is still largely under-investigated (see Liang et al., 2018; Paul and Dredze, 2017; Xu et al., 2017; Zhong and Bian, 2016).

Finally, a proper quantification of various sources of uncertainties is a core requirement for operational implementation of the developed epidemiological forecasting (see Held et al., 2017; Moran et al., 2016). Nevertheless, assessment of uncertainty in prediction of infectious disease activity is still often overlooked and systematic statistical methodology for uncertainty quantification has not yet widely penetrated epidemiological applications.

This paper aims to fill these gaps, with the multifold goal of providing an interdisciplinary insight into the predictive utility of deep learning, the role of online and offline data fusion, and statistical multi-model combination as uncertainty quantification tool, in application to forecasting influenza in Dallas County. The key contributions of the paper are:

- We investigate utility of deep learning with feedforward neural networks (DL with FNN) for ILI prediction, in application to forecasting influenza in Dallas County.
- We develop fully probabilistic forecasting of ILI, using Bayesian model averaging (BMA). To our knowledge, we are the first to use continuous ranked probability score (CRPS) as weights for BMA in epidemiological applications.
- We assess the proposed statistical forecasting methodology and a fusion of online and offline epidemiological data sources, in application to influenza prediction in Dallas County, and to our knowledge, this study is the first to assess influenza activity in Dallas County.

The remainder of the paper is organized in four major sections: data description, statistical methodology, case study, and discussion. In Section 2, we provide information on the collected influenzal, online social media, and atmospheric data. We introduce our statistical and machine learning forecasting methodology in Section 3. Section 4 is devoted to validation of the proposed modeling approaches to prediction of seasonal influenza in Dallas County, TX, USA. Finally, the paper is concluded with discussion and future work in Section 5.

2. Data description

2.1. Influenza data

The ILI data are provided by the Dallas County Public Health Informatics and represent the ILI percentage among visits of the Dallas County hospitals. ILI cases are defined according to the Texas Department of State and Health Services by fever (⩾100 °F or 37.8 °C, oral or equivalent), cough and/or sore throat, and without a known cause other than influenza (Texas DSHS, 2017). The data are weekly seasonal, from the first week of October to the last week of April, for seasons of 2011–2012 up to 2017–2018 (Fig. 1).

The data quality was controlled by the Department of Public Health of Dallas County as follows:

- *Validation/continuity* check: every 24 hours data must be received from each hospital (to assure connections are live and operational).
- *Staging* checks (each message checked as received): valid patient visit and patient ID, patient category (emergency room or inpatient), valid chief complaint, other critical HL7 standards for proper binning visit record; invalid messages filtered. (Here HL7 refers to the set of international standards for transfer of clinical and administrative data between software applications used by various health care providers.)
- *Hospital data quality*: statistical analysis performed monthly with minimum requirements.
- *Timeliness*: 90% of records received within 24 h.
- *Completeness*: 90% of records contain all required data elements.
- *Validity*: 80% of data elements received must contain valid data (numerical, coded, free-text, etc.).

With exception of 13 hospitals out of 110, all data were received in real time.

2.2. Google search data

We collected weekly data on normalized search frequency in Dallas and Fort Worth area using seven terms related to influenza: *cold*, $G_1(t)$; *cough*, $G_2(t)$; *fever*, $G_3(t)$; *flu*, $G_4(t)$; *h3n2*, $G_5(t)$; *influenza*, $G_6(t)$, and *sore throat*, $G_7(t)$. The data come from the Google trends website for the period from October 2011 to January 2018 (Fig. 2).

Each individual ith time series $G_i(t)$ ($i = 1, \ldots, 7$) may be very noisy, but combined these time series can potentially reflect the true influenza activity. As Figs. 1 and 2 suggest, generally, the dynamics of the Google trend search terms and of the ILI are well-aligned.

2.3. Atmospheric data

Meteorological data have been used frequently for modeling influenza activity (Basile et al., 2018; Lowen and Steel, 2014; Wu et al., 2016). We used Weather Underground (2017) to access daily data on absolute maximum temperature, $W_1(t)$; mean temperature, $W_2(t)$; absolute minimum temperature, $W_3(t)$; mean dew point, $W_4(t)$; mean relative humidity, $W_5(t)$, and daily high pressure, $W_6(t)$, during October 2011–January 2018. The data were aggregated weekly and aligned with the ILI data. To account for weather shocks, we calculated average temperature difference, $W_7(t) = W_2(t) - W_2(t-1)$, and average humidity difference, $W_8(t) = W_5(t) - W_5(t-1)$ (Fig. 3).

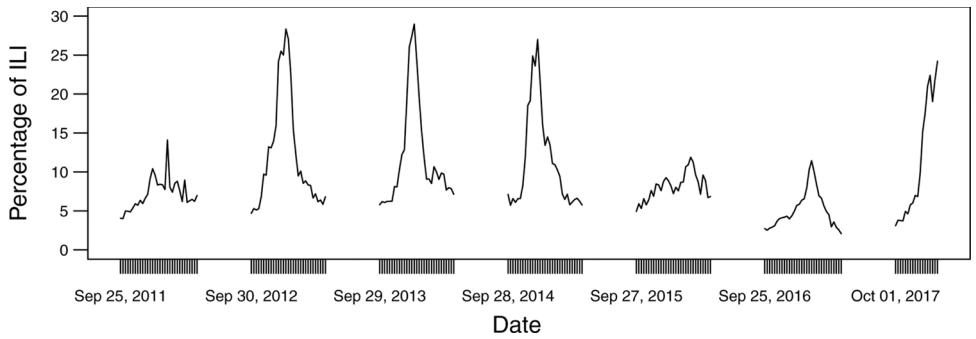

Fig. 1. Emergency department influenza like illness (ILI), October 2011–January 2018.

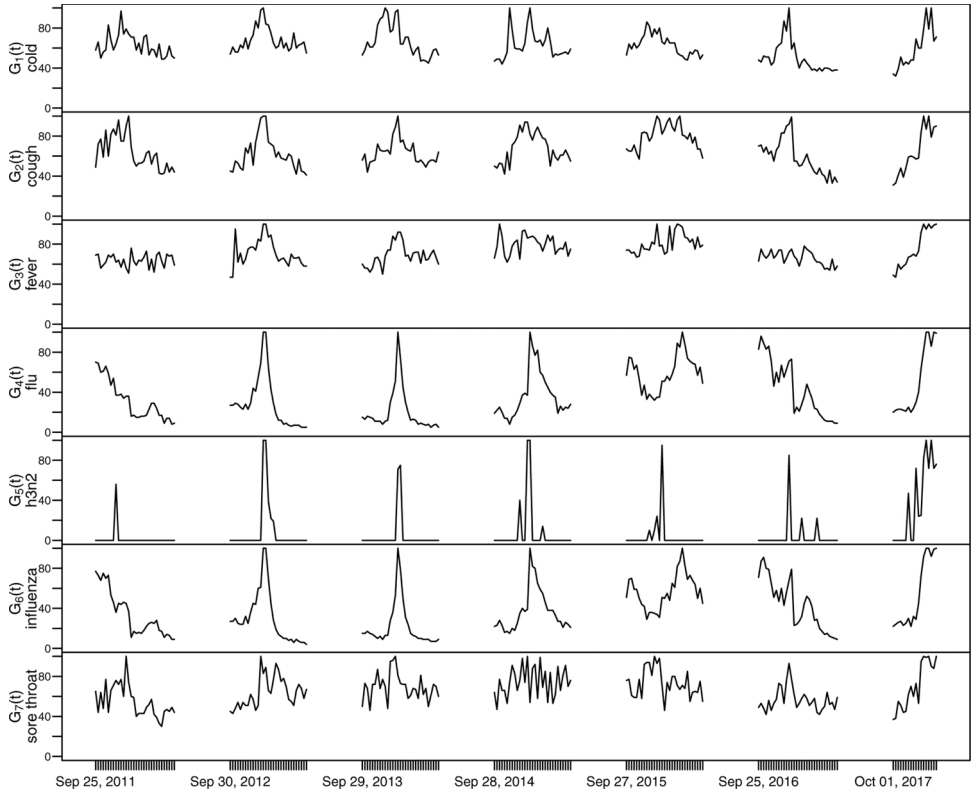

Fig. 2. Normalized frequency of Google search queries related to influenza, October 2011–January 2018.

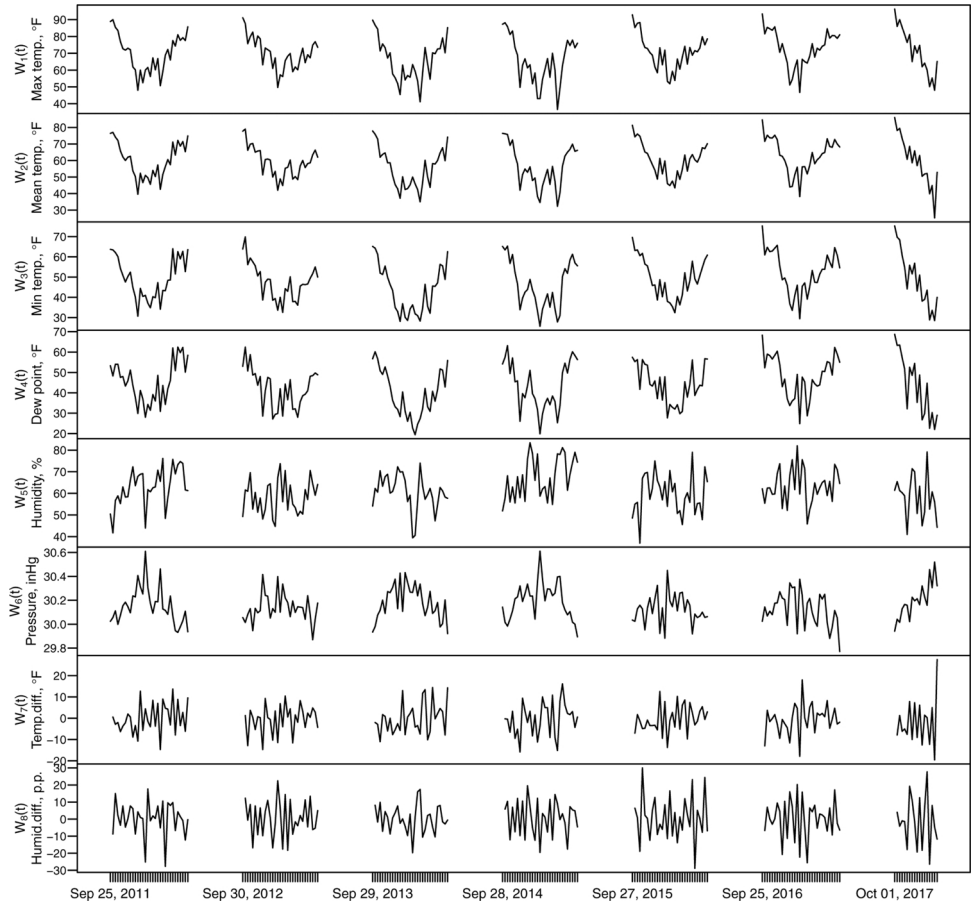

Fig. 3. Time series plots of the weather variables, October 2011–January 2018.

To account for possible delayed effects, we also incorporate lagged atmospheric and Google trend variables. Lagged variables are selected based on cross-correlation with the ILI data and further analysis of the predictive utility. All the resulting variables are summarized in Table 1.

3. Methodology for epidemiological forecasting

3.1. Beta regression

Since the ILI data are presented as fractional data in the interval (0, 1), we can employ the method of beta regression, similar to Guolo and Varin (2014). The beta regression model is based on a re-parameterization of the beta density in terms of the mean and precision parameters (Cribari-Neto and Zeileis, 2009; Ferrari and Cribari-Neto, 2004). The beta density is usually expressed as

$$f(y;\, p,\, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1,$$

where $p > 0$, $q > 0$, and $\Gamma(\cdot)$ is the gamma function. By setting $\mu = p/(p+q)$ and $\phi = p + q$, Ferrari and Cribari-Neto (2004) proposed a different parameterization:

$$f(y;\, \mu,\, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$, $\phi > 0$, $y \sim Beta(\mu,\, \phi)$. Hence, mean and variance are given by $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \mu(1-\mu)/(1+\phi)$, respectively. The parameter ϕ is known as the precision parameter since, for fixed μ, the larger ϕ the smaller the variance of y (ϕ^{-1} is a dispersion parameter).

Let $y_1, \ldots, y_n$ be a random sample of weekly ILI data such that $y_i \sim Beta(\mu_i,\, \phi)$, $i = 1, \ldots, n$. The beta regression model is defined as $g(\mu_i) = x_i^\top \beta$, where $\beta = (\beta_0, \ldots, \beta_k)^\top$ is a $k \times 1$ vector of regression coefficients ($k < n$), $x_i = (1, \ldots, x_{ik})^\top$ is the vector of k regressors. Here $g(\cdot)$ is the link function that maps (0, 1) into $\mathbb{R}$. The beta regression model was fit using R package `betareg` (Zeileis et al., 2018):

$$g(y_{t+h}) = \alpha_0 + \sum_{i=1}^{8}\sum_{l=0}^{1} \alpha_{i,l} W_i(t-l) + \sum_{i=1}^{7}\sum_{l=0}^{2} \beta_{i,l} G_i(t-l) + \sum_{l=2}^{4} \gamma_l y_{t-l} + \epsilon_{t+h},$$

where $g(\cdot)$ is a logit link function; y_t is the ILI activity; $h = 1, 2$ are forecasting horizons; α_0 is the intercept term; $\alpha_{i,l}$ and $\beta_{i,l}$ are regression coefficients for the weather variables $W_i(t)$ and Google trend data $G_i(t)$, respectively; γ_l are coefficients for the autoregressive relationships between current and past ILI activity, and ϵ_t is the error term.

3.2. Box–Jenkins model

Here we consider a standard benchmark autoregressive integrated moving average (ARIMA(p, 0, q)) model with exogenous regressors (Box and Jenkins, 1968):

$$y_{t+h} = \alpha_0 + \underbrace{\sum_{i=1}^{p} \gamma_i y_{t-i}}_{\text{AR part}} + \underbrace{\sum_{j=1}^{q} \theta_j \epsilon_{t+h-j}}_{\text{MA part}} + \underbrace{\epsilon_{t+h}}_{\text{error}} + \underbrace{\sum_{i=1}^{H}\sum_{l=0}^{R} \alpha_{i,l} W_i(t-l) + \sum_{k=1}^{M}\sum_{l=0}^{L} \beta_{k,l} G_k(t-l)}_{\text{exogenous regressors}},$$

where γ_i, $i = 1, \ldots, p$ and θ_j, $j = 1, \ldots, q$ are autoregressive and moving average coefficients, respectively; ϵ_t is white noise (i.e., $\epsilon_t \sim WN(0, \sigma^2)$). The model orders p, q, H, M, R and L are selected based on the Akaike information criterion (AIC; Bozdogan, 1987), using R package `forecast` (Hyndman et al., 2018). In our case we find that $p = 4$, $d = 0$, $q = 1$, $H = 8$, $M = 7$, $R = 1$, $L = 2$ is a suitable choice for the model order. Furthermore, due to unavailability of the ILI data for the most recent 2 weeks, we set $\gamma_1 = 0$.

Here p is a number of autoregressive (AR) terms, q is a number of moving average (MA) terms, and d is a number of non-seasonal differences; $H = 8$, $M = 7$ are the numbers of atmospheric and Google trend variables, respectively; $R = 1$, $L = 2$ represent the number of lags.

3.3. Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator (LASSO) is a penalized linear regression by Tibshirani (1996), in which the L_1 penalty is used for both fitting and penalization of the coefficients. The objective of LASSO is to prevent the model from overfitting due to either collinearity of the regressors or high dimensionality. The coefficients of a LASSO model were estimated using R package `penalized` Goeman et al. (2018) by minimizing:

$$\sum_t \left[\ln(y_{t+h}) - \alpha_0 - \sum_{p=1}^{27} a_p X_{p,t}\right]^2 + \lambda \sum_{p=1}^{27} |a_p|,$$

where λ is the penalty parameter and $X_{p,t}$ represents all regressors, $X_{p,t} = \{G_1(t), G_1(t-1), \ldots, G_7(t), W_1(t), W_1(t-1), \ldots, W_8(t), y_{t-2}, y_{t-3}, y_{t-4}\}$, see the complete list in Table 1. LASSO has been widely used for modeling influenza dynamics (see, for instance, Guo et al., 2017; Hayate et al., 2016; Lu et al., 2018; Volkova et al., 2017; Xu et al., 2017).

3.4. Deep learning – feedforward neural network

Deep learning (DL) is a machine learning tool that is based on learning data representation and originally rooted in the methodology of neural networks (Goodfellow et al., 2016; LeCun et al., 2015; Schmidhuber, 2015). DL has proven to exhibit a high utility in a broad range of applications, from computer vision and natural language processing to analysis of complex network, bioinformatics, and finance (Goodfellow et al., 2016; Gupta and Raza, 2018; Nielsen, 2015; Skansi, 2018; Zhao et al., 2017). Nevertheless, applications of DL for epidemiological forecasting and, in particular, influenza studies are still limited (see Xu et al., 2017, and reference therein).

Table 1
Summary of predictors for the ILI activity $y(t)$.

Google variables	Atmospheric variables	ILI activity
$G_1(t)$ and $G_1(t-1)$ for *cold*	$W_1(t)$ and $W_1(t-1)$ is maximum temperature	$y(t-2)$ is ILI at time $t-2$
$G_2(t)$ and $G_2(t-1)$ for *cough*	$W_2(t)$ is mean temperature	$y(t-3)$ is ILI at time $t-3$
$G_3(t)$ and $G_3(t-1)$ for *fever*	$W_3(t)$ is minimum temperature	$y(t-4)$ is ILI at time $t-4$
$G_4(t)$ and $G_4(t-1)$ for *flu*	$W_4(t)$ and $W_4(t-1)$ is mean dew point	
$G_5(t)$ and $G_5(t-2)$ for *h3n2*	$W_5(t)$ and $W_5(t-1)$ is mean relative humidity	
$G_6(t)$ and $G_6(t-1)$ for *influenza*	$W_6(t)$ is daily high pressure	
$G_7(t)$ for *sore throat*	$W_7(t)$ is average temperature difference	
	$W_8(t)$ is average humidity difference	
Total 13 variables	Total 11 variables	Total 3 variables

In this project, we employ a feedforward neural network (FNN) architecture for DL, where information moves from input nodes to hidden layers to output nodes; there are no cycles or loops in the network (Schmidhuber, 2015; Zhang and Zong, 2015). The input variables are the Google trend data ($G_i(t-l)$; $i = 1, \ldots, 7$; $l = 0, a, 2$), the atmospheric data ($W_i(t-l)$; $i = 1, \ldots, 8$; $l = 0, 1$), and the offline records of ILI in the previous weeks y_{t-2}, y_{t-3}, and y_{t-4}. We trained FNNs using R package `h2o` (LeDell et al., 2018) and selected the optimal FNN structure using cross-validation. For our data, the selected number of hidden layers was 2, with 75 hidden nodes each, and the learning rate which yielded the best performance in terms of mean squared error was 0.005.

3.5. Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) is a nonparametric regression model (Friedman, 1991) that has no assumptions about the form of relationship between the response and predictors. The model constructs this relationship from a set of coefficients and basis functions that are entirely determined from the data. MARS performs a forward/backward stepwise approach to determine the knot points in a data set, which defines each basis function using generalized cross-validation (GCV) criterion. MARS model can be written as follows:

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x),$$

where $\hat{f}(x)$ is the estimated function, $B_m(x)$ is the basis function, a_m is the coefficient, M is number of the basis functions, and x are the regressors (i.e., the Google trend, weather variables, and the past ILI activity).

The basis function is defined as:

$$B_m^{(q)}(x) = \prod_{k=1}^{K_m} [s_{\mathrm{km}}(x_{v(k,m)} - t_{\mathrm{km}})]_+^q,$$

where $s_{km} = \pm 1$, K_m is the number of factors (interaction order) in the mth basis function, $v(k, m)$ label the predictor variables and t_{km} represent values on the corresponding variables, q is the order of the spline, and "+" means positive part.

The truncated power spline of first order is

$$(x-t)_+ = \begin{cases} x-t, & \text{if } x > t, \\ 0, & \text{otherwise.} \end{cases}$$

To our knowledge, MARS has not been used for influenza forecasting yet. We used R package `mda` (Hastie et al., 2017).

3.6. Variable selection

We have applied all the variables (Table 1) to all the models above. In beta regression and ARIMA, we used statistically significant variables only, since LASSO, MARS and deep learning do the variable selection or flexibly readjust coefficients automatically. In beta regression, the significant variables are ILI data with lags 2 and 4 ($y(t-2)$ and $y(t-4)$), average temperature, temperature difference ($W_2(t)$, $W_7(t)$), and Google time series for cold, fever, flu, influenza, and lagged by one week time series for cough and flu ($G_1(t)$, $G_3(t)$, $G_4(t)$, $G_6(t)$, $G_2(t-1)$, and $G_4(t-1)$). In ARIMA, the significant variables are ILI data with lags 2, 3 and 4 ($y(t-2)$, $y(t-3)$ and $y(t-4)$), average temperature, temperature difference, and lagged maximum temperature ($W_2(t)$, $W_7(t)$, and $W_1(t-1)$), and Google series for cold, fever, flu, influenza, and lagged series for cold ($G_1(t)$, $G_3(t)$, $G_4(t)$, $G_6(t)$, and $G_1(t-1)$).

3.7. Bayesian model averaging

Efficient operational implementation of health risk mitigation strategies requires proper assessment of various uncertainty sources associated with the delivered epidemiological forecasts. Nevertheless, a critical step of uncertainty quantification is still often overlooked in many health surveillance and forecasting studies. We propose to evaluate forecasting uncertainty by developing a weighted multi-model ensemble of future epidemiological states. That is, we apply the Bayesian model averaging (BMA), which allows us to combine multiple epidemiological models with weights corresponding to their most recent prediction performance (Hoeting et al., 1999; Raftery et al., 2005; Raftery and Painter, 2005). To better quantify prediction uncertainty, we propose two probabilistic forecasts based on the BMA methodology and two types of model weights.

Our first BMA approach is to define model weights via root mean square error (RMSE) of fitted $\tilde{y}_t$ by each model on the training set of data y_t ($t = 1, \ldots, n$), i.e.,

$$\mathrm{RMSE}(k) = \sqrt{\frac{\sum_{i=1}^{n}(y_t - \tilde{y}_t)^2}{n}},$$

where n is the size of training data set; RMSE(k) corresponds to the kth model, namely, for beta regression $k = 1$; for ARIMA $k = 2$; for LASSO $k = 3$; for MARS $k = 4$, and for deep learning $k = 5$. The resulting RMSE(k) leads to the following corresponding weight of the kth model in BMA:

$$\frac{1/\mathrm{RMSE}(k)}{\sum_{i=1}^{5} 1/\mathrm{RMSE}(i)}.$$

Let the predicted values from each model be $Pred(k)$ for $k = 1, \ldots, 5$. Then the BMA forecast with RMSE as weights is given by

$$\mathrm{Pred}_{\mathrm{RMSE}} = \sum_{k=1}^{5} \frac{1/\mathrm{RMSE}(k)}{\sum_{i=1}^{5} 1/\mathrm{RMSE}(i)} \mathrm{Pred}(k).$$

We follow a similar algorithm for BMA with continuous ranked probability score (CRPS) weights. That is, we start from calculating CRPS for k-th model ($k = 1, \ldots, 5$) based on truncated normal distribution. Then the resulting BMA forecast with CRPS as weights is given by

$$\mathrm{Pred}_{\mathrm{CRPS}} = \sum_{k=1}^{5} \frac{1/\mathrm{CRPS}(k)}{\sum_{i=1}^{5} 1/\mathrm{CRPS}(i)} \mathrm{Pred}(k).$$

3.8. Evaluation metrics

We use three standard statistical measures of accuracy: RSME, mean absolute percentage error (MAPE), and mean absolute error (MAE) (Liang et al., 2018; Lu et al., 2018; Yang et al., 2017). Let $\hat{y}_i$ be the forecast value and y_i be the corresponding real value, then

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \\ \mathrm{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \\ \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \end{aligned}$$

where n is the number of out-of-sample forecasts produced by a model.

In addition, we measured the correlation between the observed and forecasted values using Pearson correlation coefficient

$$r(y, \hat{y}) = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Lower RMSE, MAPE, MAE and higher Pearson correlation coefficient imply better forecasting performance.

Finally, we employ the CRPS to assess performance of probabilistic (or weighted multi-model ensemble) forecast delivered by the BMA

Table 2
Performance summary for the five models for prediction ILI one week ahead.

Method	RMSE	MAE	MAPE	$r(y, \hat{y})$
Beta regression	8.12	4.15	0.38	0.74
ARIMA	5.40	3.57	0.40	0.74
Deep learning	4.14	2.95	0.32	0.83
LASSO	5.38	3.54	0.38	0.79
MARS	5.14	3.42	0.37	0.75

Table 3
Performance summary for the five models for prediction ILI two weeks ahead.

Method	RMSE	MAE	MAPE	$r(y, \hat{y})$
Beta regression	8.03	4.48	0.44	0.69
ARIMA	5.59	3.79	0.43	0.74
Deep learning	4.89	3.60	0.38	0.76
LASSO	6.17	4.29	0.48	0.73
MARS	6.05	4.03	0.44	0.72

approach (Gneiting et al., 2005). CRPS can be viewed as a generalization of the MAE, which assesses performance of a point forecast, to measuring the accuracy of a probabilistic forecast. CRPS is widely used in atmospheric studies, social sciences, and finance (Alfieri et al., 2014; Chu and Coimbra, 2017; Fan et al., 2016; Gneiting and Raftery, 2007; Scheuerer and Hamill, 2015; Thorey et al., 2017). CRPS is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} [F(y) - H(y - x)]^2 \mathrm{d}y,$$

where $H(y - x)$ denotes the Heaviside step function and is 0, when $x > y$, and 1 otherwise, and F is the cumulative distribution function associated with an empirical probabilistic forecast. In finite samples, CRPS takes the form (Gneiting and Raftery, 2007):

$$\mathrm{CRPS}(F, x) = \frac{1}{2} E_F |X - X'| - E_F |X - x|,$$

where X and X' are independent copies of an integrable random variable following the distribution F. The lower the CRPS, the better the accuracy.

4. Case study of forecasting flu activity in Dallas County, TX

We use an adaptive form of out-of-sample forecast where the model is trained throughout the analysis, by using a 20-week sliding window for all considered models (that is, we use only 20 most recent weeks to train the models). The window length of 20 weeks is found to deliver the most competitive RMSE. The out-of-sample forecast validation started from September 2012 to January 2018.

In our analysis, we consider the following models: beta regression, ARIMA, LASSO, DL with FNN, and MARS. In the Box–Jenkins approach, ARIMA(4,0,1) is selected as the optimal model using **auto.arima** function in R (Hyndman et al., 2018). In LASSO, the tuning parameter λ is selected using cross-validation, that is, we choose λ that delivers a minimum mean cross-validated error ($\lambda = 0.0620$ for 1-week, and $\lambda = 0.0823$ for 2-week ahead predictions). In FNN, the number of hidden layers are 2 with 75 hidden nodes, and the optimal DL structure is selected using cross-validation.

4.1. Individual point forecasts

Tables 2 and 3 show that point forecasts by the deep learning model (FNN) outperform forecasts delivered by the other models, based on the relatively low errors (RMSE, MAE, and MAPE) and relatively high

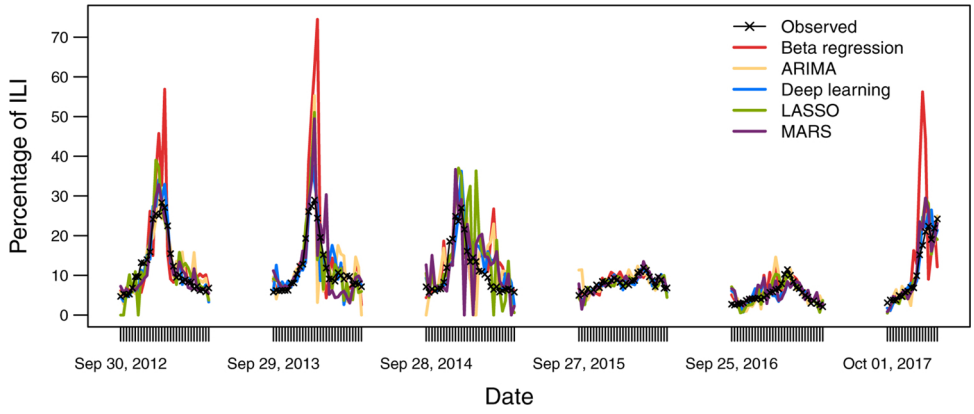

Fig. 4. One week ahead forecasts.

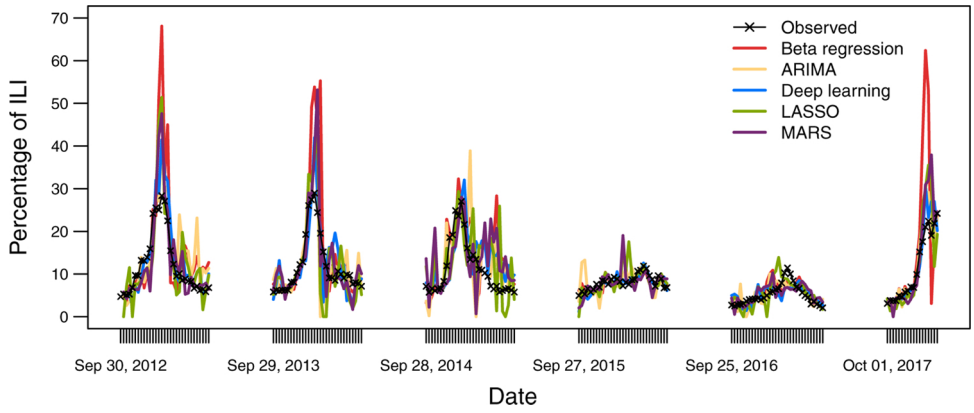

Fig. 5. Two weeks ahead forecasts.

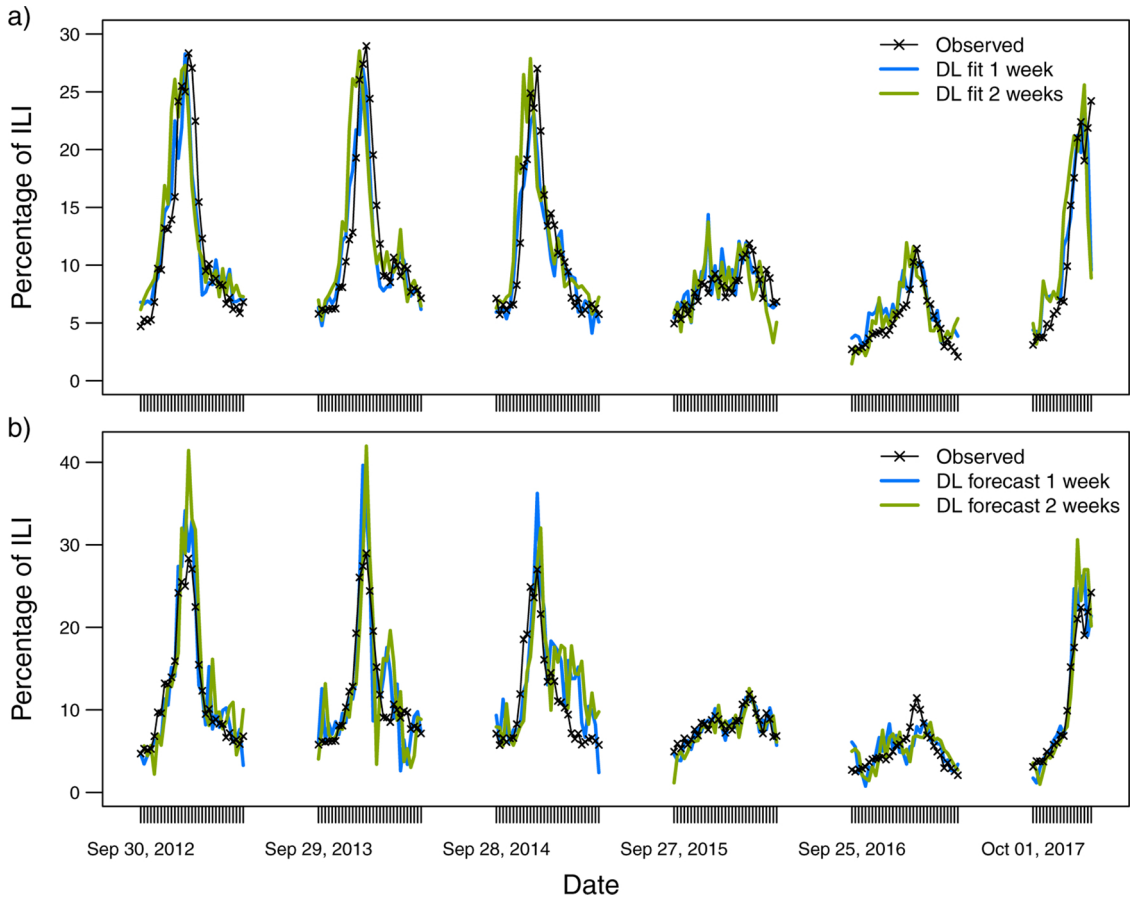

Fig. 6. Performance of the FNN deep learning procedure based on (a) fitted ILI percentages, and (b) 1 and 2 weeks ahead forecasts of the ILI percentages.

correlation of deep learning forecasts with the actual data. The next two best methods are MARS for 1-week ahead forecasts (Table 2) and ARIMA for 2-week ahead forecasts (Table 3), however, the differences between these two methods are small. Figs. 4 and 5 show point forecasts 1 and 2 weeks ahead by all individual methods. We observe that deep learning appears to deliver the most competitive forecasting performance than other approaches in terms of best fitting of the ILI curve.

The curve fits delivered by the deep learning procedure are shown in Fig. 6a, while the forecasts of 1 and 2 weeks ahead are in Fig. 6b. As expected, we observe smaller errors for the in-sample fit (Fig. 6a) than for out-of-sample forecasts (Fig. 6b).

4.2. BMA point forecasts

We consider two types of BMA-based forecasts, that is, the weighted multi-model ensembles, delivered by BMA using RMSE or CRPS as weights. To calculate RMSE, MAE, MAPE and correlation coefficient of the BMA-based probabilistic forecasts, we use the mean of the corresponding weighted multi-model ensemble. Notice that BMA weights are adaptively recalculated at each time point t. Hence, weights and the corresponding model ranking scores vary over time. Nevertheless, we find that on average DL tends to deliver one of the highest performance scores that are consistent over BMA with RMSE and BMA with CRPS and forecasting horizons. The next performance ranking scores are followed by LASSO and MARS which weights, however, vary substantially over type of BMA and forecasting horizons.

Table 4 shows that the BMA with CRPS as weights is better in modeling the ILI curve than BMA with RMSE as weights, through comparing the RMSE, MAE, and MAPE for both one week and two weeks forecast. By comparing Table 4 with Tables 2 and 3, we observe that BMA with CRPS weights outperforms deep learning with FNN and all other individual methods in 1-week ahead forecasts, and both BMA methods outperform the individual methods in forecasting ILI two weeks ahead.

Table 4
Predictive performance of BMA using RMSE and CRPS as weights.

Forecast horizon	Weight	RMSE	MAE	MAPE	$r(y, \hat{y})$
One week	RMSE	4.15	2.78	0.29	0.85
One week	CRPS	3.73	2.59	0.27	0.84
Two weeks	RMSE	4.39	3.17	0.34	0.79
Two weeks	CRPS	4.33	3.11	0.32	0.78

4.3. Interval forecasts

Since the response variable, ILI percentage, is bounded between 0 and 100, we construct the probabilistic forecast of ILI by embedding the point forecasts (delivered by beta regression, ARIMA, LASSO, DL with FNN, and MARS) with a truncated normal distribution, using BMA. The probability density function of truncated normal distribution is defined as (Forbes et al., 2011)

$$f(y; \mu, \sigma, a_1, a_2) = \frac{\phi(\frac{y-\mu}{\sigma})}{\sigma\left(\Phi(\frac{a_2-\mu}{\sigma}) - \Phi(\frac{a_1-\mu}{\sigma})\right)},$$

where μ and σ are mean and standard deviation, respectively; a_1 and a_2 are the lower and upper limits which y can attain (i.e., $y \in [a_1, a_2]$); $\phi(\xi) = (2\pi)^{-1/2}\exp(-\xi^2/2)$ is the probability density function of the standard normal distribution, and $\Phi(y) = (1 + \mathrm{erf}(y/\sqrt{2}))/2$ is its cumulative distribution function. To construct prediction intervals of the ILI percentage, we used $a_1 = 0$ and $a_2 = 100$.

Table 5
CRPS score for all models for one week and two weeks prediction.

Method	CRPS score one week	CRPS score two weeks
Beta regression	3.61	4.15
ARIMA	3.02	3.29
Deep learning	2.59	3.25
LASSO	3.25	3.99
MARS	2.79	3.36
BMA-RMSE	2.41	2.73
BMA-CRPS	1.89	2.23

We evaluate the quality of interval forecasts using CRPS, which indicates that for both 1- and 2-week ahead prediction horizons, BMA-based weighted multi-model ensembles with CRPS as weights delivers the most competitive performance, i.e., the lowest CRPS for 1-week forecasts (Table 5). The next best is BMA with RMSE as weights, followed by DL with FNN, while for 2-week we can see that BMA with CRPS as weights is still the best followed by BMA with RMSE as weights then DL with FNN. Figs. 7 and 8 show the BMA-CRPS based predictions for different confidence levels, overlaid with the observed ILI percentages.

5. Discussion

In this study, we evaluated the forecasting utility of deep learning for predicting ILI in Dallas County. In addition to historical ILI records, we have considered Google search queries and meteorological data as exogenous predictors of influenza activity. The DL approach has been compared to more conventional statistical methods such as beta regression, autoregressive integrated moving average (ARIMA), least absolute shrinkage and selection operators (LASSO), and a non-parametric multivariate adaptive regression splines (MARS) models. Furthermore, we have developed the probabilistic forecast of ILI activity, namely, the weighted multi-model ensembles of ILI, based on Bayesian model averaging (BMA) with RMSE and CRPS as weights.

Our results indicate that in terms of evaluation metrics for point forecasts (i.e., RMSE, MAE, and MAPE), DL and the BMA-based multi-model ensemble of ILI forecasts yield a similar competitive performance, outperforming all other considered models. In terms of assessment of probabilistic forecasts, the BMA-based multi-model ensemble with CRPS as weights provides the highest predictive accuracy, delivering CRPS of about 20% lower than the next best approach BMA-RMSE, and 29% lower than DL.

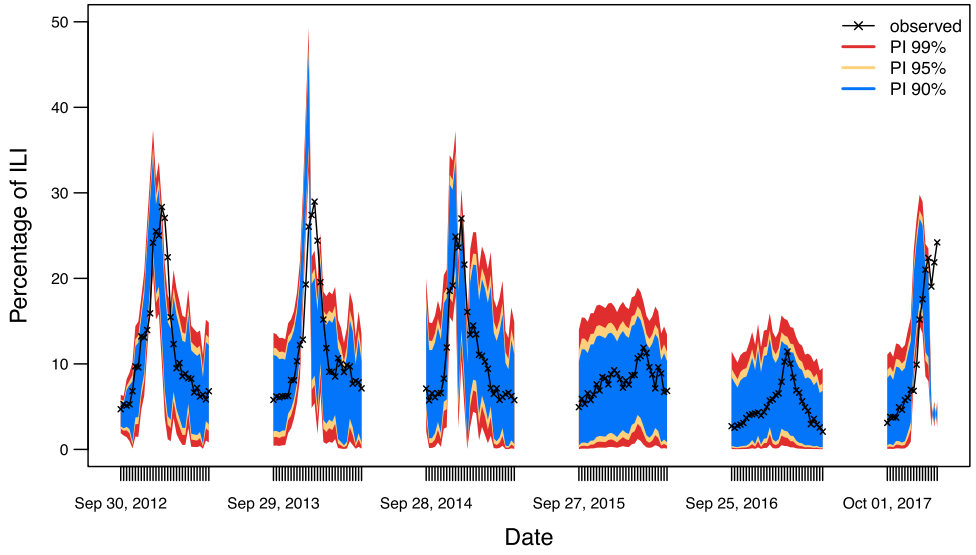

Fig. 7. One week ahead prediction intervals (PI) of BMA with CRPS weights, and observed ILI.

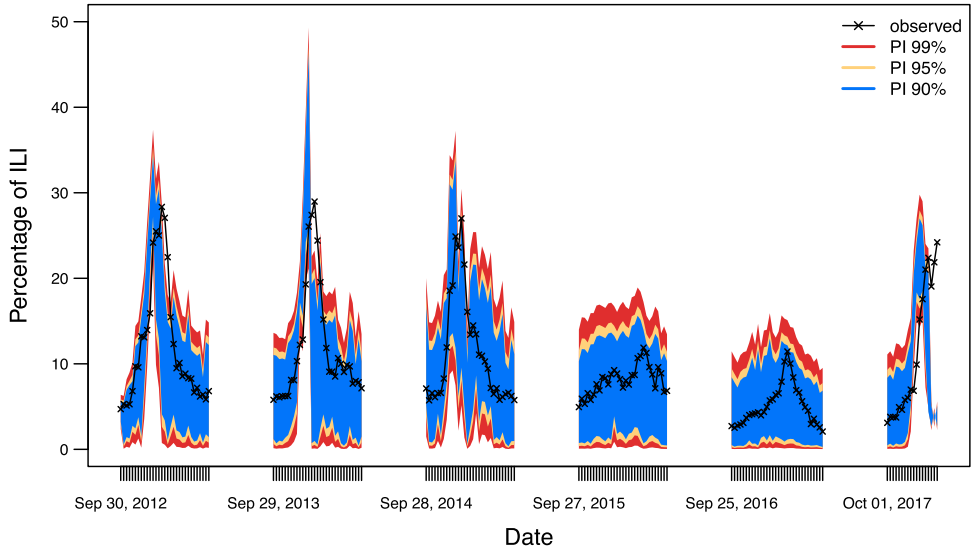

Fig. 8. Two weeks ahead prediction intervals (PI) of BMA with CRPS weights, and observed ILI.

These findings suggest that harnessing multi-model strengths via statistical model fusion appears a promising direction for enhancing current epidemiological prediction. In the future, we plan to complement the proposed methodology by integrating spatial data relevant to ILI activity and by applying tools of geometric deep learning for space-time epidemiological forecasting, with a particular focus on non-seasonal influenza and climate sensitive infectious diseases.

References

Alessa, A., Faezipour, M., 2018. A review of influenza detection and prediction through social networking sites. Theor. Biol. Med. Model. 15 (1), 2.

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., Salamon, P., 2014. Evaluation of ensemble streamflow predictions in Europe. J. Hydrol. 517, 913–922.

Allen, C., Tsou, M.H., Aslam, A., Nagel, A., Gawron, J.M., 2016. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. PLOS ONE 11 (7), e0157734.

Aronis, J.M., Millett, N.E., Wagner, M.M., Tsui, F., Ye, Y., Ferraro, J., Haug, P.J., Cooper, G.F., 2017. Detecting overlapping outbreaks of influenza. Online J. Public Health Inform. 9 (1), e14.

Basile, L., de la Fuente, M.O., Torner, N., Martínez, A., Jané, M., 2018. Real-time predictive seasonal influenza model in Catalonia, Spain. PLOS ONE 13 (3), e0193651.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I.C.H., Hickmann, K.S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.H., et al., 2016. Results from the centers for disease control and prevention's predict the 2013-2014 Influenza Season Challenge. BMC Infect. Dis. 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., et al., 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics 24, 26–33.

Box, G.E.P., Jenkins, G.M., 1968. Some recent advances in forecasting and control. J. R. Stat. Soc. Ser. C (Appl. Stat.) 17 (2), 91–109.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52 (3), 345–370.

CDC, 2017. National Pandemic Strategy. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/pandemic-resources/national-strategy/index.html.

Chu, Y., Coimbra, C.F., 2017. Short-term probabilistic forecasts for direct normal irradiance. Renew. Energy 101, 526–536.

Cribari-Neto, F., Zeileis, A., 2009. Beta Regression in R. Department of Statistics and Mathematics, WU Vienna University of Economics and Business.

Dugas, A.F., Jalalpour, M., Gel, Y.R., Levin, S., Torcaso, F., Igusa, T., Rothman, R.E., 2013. Influenza forecasting with Google flu trends. PLOS ONE 8 (2), e56176.

Fan, F., Bell, K., Infield, D., 2016. Probabilistic weather forecasting for dynamic line rating studies. In: Power Systems Computation Conference (PSCC). 2016, IEEE. pp. 1–7.

Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. J. Appl. Stat. 31 (7), 799–815.

Forbes, C., Evans, M., Hastings, N., Peacock, B., 2011. Statistical Distributions, 4th ed. John Wiley & Sons, Hoboken, NJ.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.

Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133 (5), 1098–1118.

Goeman, J., Meijer, R., Chaturvedi, N., Lueder, M., 2018. penalized: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model. R package version 0.9-51. https://CRAN.R-project.org/package=penalized.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning, vol. 1 MIT Press, Cambridge.

Guo, P., Zhang, J., Wang, L., Yang, S., Luo, G., Deng, C., Wen, Y., Zhang, Q., 2017. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. Sci. Rep. 7, 46469.

Guolo, A., Varin, C., 2014. Beta regression for time series analysis of bounded data, with application to Canada Google® flu trends. Ann. Appl. Stat. 8 (1), 74–88.

Gupta, T.K., Raza, K., 2018. Optimizing Deep Neural Network Architecture: A Tabu Search Based Approach. arXiv preprint arXiv:180805979.

Hastie, T., Tibshirani, R., Leisch, F., Hornik, K., Ripley, B.D., 2017. mda: Mixture and Flexible Discriminant Analysis. R package version 0.4-10. https://CRAN.R-project.org/package=mda.

Hayate, I., Wakamiya, S., Aramaki, E., 2016. Forecasting word model: Twitter-based influenza surveillance and prediction. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers 76–86.

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

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–417.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2018. forecast: Forecasting Functions for Time Series and Linear Models. R package version 8.4. https://CRAN.R-project.org/package=forecast.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava, M., 2018. h2o: R Interface for 'H2O'. R package version 3.20.0.8. https://CRAN.R-project.org/package=h2o.

Lee, A., Chuh, A.A.T., 2010. Facing the threat of influenza pandemic-roles of and implications to general practitioners. BMC Public Health 10 (1), 661.

Liang, F., Guan, P., Wu, W., Huang, D., 2018. Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015. PeerJ 6, e5134.

Lowen, A.C., Steel, J., 2014. Roles of humidity and temperature in shaping influenza seasonality. J. Virol. 88 (14), 7692–7695.

Lu, F.S., Hou, S., Baltrusaitis, K., Shah, M., Leskovec, J., Sosic, R., Hawkins, J., Brownstein, J., Conidi, G., Gunn, J., 2018. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the Boston Metropolis. JMIR Public Health Surveill. 4 (1), e4.

McIver, D.J., Brownstein, J.S., 2014. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput. Biol. 10 (4), e1003581.

Moran, K.R., Fairchild, G., Generous, N., Hickmann, K., Osthus, D., Priedhorsky, R., Hyman, J., Del Valle, S.Y., 2016. Epidemic forecasting is messier than weather forecasting: the role of human behavior and internet data streams in epidemic forecast. J. Infect. Dis. 214 (suppl_4), S404–S408.

Nielsen, M., 2015. Neural Networks and Deep Learning. Determination Press, USA.

Paul, M.J., Dredze, M., 2017. Social monitoring for public health. Synth. Lect. Inf. Concepts Retr. Serv. 9 (5), 1–183.

Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D., Weinstein, R.A., 2008. Using internet searches for influenza surveillance. Clin. Infect. Dis. 47 (11), 1443–1448.

Preis, T., Moat, H.S., 2014. Adaptive nowcasting of influenza outbreaks using Google searches. R. Soc. Open Sci. 1 (2), 140095.

Raftery, A.E., Painter, I.S., 2005. BMA: an R package for Bayesian model averaging. R News 5 (2), 2–8.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174.

Rolfes, M.A., Foppa, I.M., Garg, S., Flannery, B., Brammer, L., Singleton, J.A., Burns, E., Jernigan, D., Reed, C., Olsen, S.J., Bresee, J., 2016. Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States. https://www.cdc.gov/flu/about/disease/2015-16.htm.

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

Scheuerer, M., Hamill, T.M., 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Mon. Weather Rev. 143 (4), 1321–1334.

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117.

Shaman, J., Kohn, M., 2009. Absolute humidity modulates influenza survival, transmission, and seasonality. Proc. Natl. Acad. Sci. USA 106 (9), 3243–3248.

Sharpe, J.D., Hopkins, R.S., Cook, R.L., Striley, C.W., 2016. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. JMIR Public Health Surveill. 2 (2), e161.

Skansi, S., 2018. Introduction to Deep Learning. Springer International Publishing, New York.

Spreco, A., Timpka, T., 2016. Algorithms for detecting and predicting influenza outbreaks: metanarrative review of prospective evaluations. BMJ Open 6 (5), e010683.

Spreco, A., Eriksson, O., Dahlström Ö, Cowling, B.J., Timpka, T., 2017a. Integrated detection and prediction of influenza activity for real-time surveillance: algorithm design. J. Med. Internet Res. 19 (6), e211.

Spreco, A., Eriksson, O., Dahlström Ö, Timpka, T., 2017b. Influenza detection and prediction algorithms: comparative accuracy trial in Östergötland county, Sweden, 2008-2012. Epidemiol. Infect. 145 (10), 2166–2175.

Texas DSHS, 2017. Texas Department of State Health Services. http://www.dshs.texas.gov/idcu/disease/influenza/.

Thorey, J., Mallet, V., Baudin, P., 2017. Online learning with the continuous ranked probability score for ensemble forecasting. Q. J. R. Meteorol. Soc. 143 (702), 521–529.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B (Methodol.) 58 (1), 267–288.

Volkova, S., Ayton, E., Porterfield, K., Corley, C.D., 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PLOS ONE 12 (12), e0188941.

Weather Underground. https://www.wunderground.com.

WHO, 2018. Influenza (Seasonal). World Health Organization (accessed 05.03.19). https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).

Wu, X., Lu, Y., Zhou, S., Chen, L., Xu, B., 2016. Impact of climate change on human infectious diseases: empirical evidence and human adaptation. Environ. Int. 86, 14–23.

Xu, Q., Gel, Y.R., Ramirez Ramirez, L.L., Nezafati, K., Zhang, Q., Tsui, K.L., 2017. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. PLOS ONE 12 (5), e0176690.

Yang, S., Santillana, M., Brownstein, J.S., Gray, J., Richardson, S., Kou, S., 2017. Using electronic health records and internet search information for accurate influenza forecasting. BMC Infect. Dis. 17 (1), 332.

Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., 2018. betareg: Beta Regression. R package version 3.1-1. https://CRAN.R-project.org/package=betareg.

Zhang, J., Zong, C., 2015. Deep neural networks in machine translation: an overview. IEEE Intell. Syst. 30 (5), 16–25.

Zhao, Y., Li, J., Yu, L., 2017. A deep learning ensemble approach for crude oil price forecasting. Energy Econ. 66, 9–16.

Zhong, S., Bian, L., 2016. Predicting influenza dynamics using a deep learning approach. International Conference on GIScience Short Paper Proceedings, vol. 1.