The Report committee for Anurekha Ramakrishnan
Certifies that this is the approved version of the following report:

Predicting influenza hospitalizations

APPROVED BY

SUPERVISING COMMITTEE:

Lauren Ancel Meyers, Supervisor

Paul Damien, Co-Supervisor

Predicting influenza hospitalizations

by

Anurekha Ramakrishnan, B.E.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2012

Dedicated to Appa, Amma, Chittu, and my Husband.

Acknowledgments

I would first like to thank my supervisor Prof. Lauren Ancel Meyers. This report would have not been possible without her constant source of guidance and support throughout the course of my research. I am grateful and count myself lucky to have got the opportunity to work with her as a research assistant.

I thank Prof. Paul Damien for his advice and supervision.

My sincere thanks to Prof. Joydeep Ghosh for providing me with the wonderful opportunity of working with him and his students as a research assistant for more than a year and a half. I learned many cutting edge technologies and algorithms during this period.

I am grateful to Prof. Daniel Powers for providing me an opportunity to pursue Masters in Statistics at UT. These two years have been a mighty learning experience and no words can express how thankful I am.

I thank each and every faculty and staff at the Division of Statistics and Scientific Computation who have supported me to make my two years of graduate study a memorable experience. Special thanks to Dr. Mary Parker for her remarkable teaching which stirred my interest in the Theory of Statistics, Dr. Matt Hersh and Vicki Keller.

Finally, I thank my parents, my brothers and my husband. I am grateful

for the unconditional love and support that they have always provided. My husband, Anand Ramalingam has been my pillar of strength throughout the course of my graduate studies. I would not have been able to complete my graduate studies without his constant encouragement and emotional support during trying times.

I am grateful to many people for their kindness and I am sorry if I have missed acknowledging them here.

Predicting influenza hospitalizations

Anurekha Ramakrishnan, M.S.Stat.
The University of Texas at Austin, 2012

Supervisor: Lauren Ancel Meyers
Co-Supervisor: Paul Damien

Seasonal influenza epidemics are a major public health concern, causing three to five million cases of severe illness and about 250,000 to 500,000 deaths worldwide. Given the unpredictability of these epidemics, hospitals and health authorities are often left unprepared to handle the sudden surge in demand. Hence early detection of disease activity is fundamental to reduce the burden on the healthcare system, to provide the most effective care for infected patients and to optimize the timing of control efforts. Early detection requires reliable forecasting methods that make efficient use of surveillance data. We developed a dynamic Bayesian estimator to predict weekly hospitalizations due to influenza related illnesses in the state of Texas. The prediction of *peak* hospitalizations using our model is accurate both in terms of number of hospitalizations and the time at which the peak occurs. For 1-to 8 week predictions, the predicted number of hospitalizations was within 8% of actual value and the predicted time of occurence was within a week of actual peak.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Seasonal Influenza causes severe illness and life threatening complications, especially in the high-risk group, which include children younger than two years, adults 65 years or older, and people with comorbidities. Influenza epidemics occur yearly during autumn and winter resulting in three to five million cases of severe illness and about 250,000 to 500,000 deaths worldwide [12].

These epidemics can be unpredictable and severe and hence early detection of disease activity is important in reducing the impact. Early detection requires regular surveillance and most importantly reliable forecasting methods that make efficient use of the surveillance data. This helps hospitals and other health care services anticipate and be well prepared.

In the literature, techniques from statistics, machine learning and natural language processing have been adapted to develop real-time tracking of influenza like illness. A linear statistical model using log-odds to analyze large numbers of Google search queries [6], optimal statistical interpolation [9], classical statistical methods such as regression [2], Bayesian models [17], particle filtering [11], and state space tracking approach based on particle learning [4] have been applied to real time learning and surveillance of infectious diseases.

Techniques from machine learning and natural language processing have been applied to mine and analyze data from online social networks such as `facebook` and micro-blogging site `twitter` to do real-time surveillance of epidemics [1, 3].

In this paper, we use a Kalman filter [7, 8] for predicting weekly hospitalizations due to influenza related illnesses in the state of Texas. The advantage of Kalman filtering, a dynamic Bayesian estimator, is that it not only provides real-time prediction but it also has the ability to correct itself from past mistakes leading to improved performance as the size of the training data (hospitalization data) increases. But hospitalization data is not available every week; there is a lag of six months to one year in the data availability. Hence, we forecast weekly hospitalizations using correlated time series data, which are available within a two-week lag.

The model, developed in this paper, predicts weekly hospitalizations many weeks into the future. We present forecast results for both timing and magnitude of the *peak* hospitalizations. To illustrate the power of the proposed method, we note that it has an error less than 8% in terms of peak magnitude and is within one week of peak occurrence when we forecast hospitalizations eight weeks from the actual peak occurence. In addition, the model output can be used for decision making in public health service delivery such as the usage of limited resources like beds, ventilators, medicines in a hospital.

Chapter 2

Hospitalizations predictors

We consider four time series that could be related to influenza related hospitalizations and thus predict hospitalizations: Google flu trends, Influenza like illnesses network (ILINet) surveillance data, humidity reports, and school calendars. These data sources have been shown to be useful and efficient in forecasting flu activities and hospitalizations in literature. Google flu trends, a website that tracks the frequency of `google` search queries related to ILI, has been shown to strongly correlate with ILINet data from the national level down to the city level [6]. The transmission of influenza virus and seasonality of influenza outbreaks are highly correlated with absolute humidity [19]. In addition, the transmission of influenza virus is influenced by the school calendars because of the high proportion of contacts in the school environment.

2.1 Google Flu Trends Data

The relative frequency of certain web search queries is highly correlated with ILINet data in the US [6]. Therefore we use Google Flu Trends as a predictor in our algorithm to predict the influenza hospitalization for the state. Data is downloaded from the Google flu trends site. Since these data are

available either for major cities or states, we use the Google Flu trends data for the state of Texas. This data is available within a one week lag, hence it comes in very handy during real-time forecasts.

2.2 ILI Data

CDC reports and tracks several types of metrics related to flu activity in the US, including hospitalizations, mortality and outpatient visits due to influenza-like-illnesses (ILI) on a weekly basis. CDC guidelines defines ILI as fever of 100 degrees F (or higher) and a cough and/or sore throat in the absence of a known cause other than influenza. A network of 2,400 sites (health departments, laboratories, vital statistics offices, health care providers, and emergency departments) in over 122 cities and 50 states, in charge of sending reports on flu activity to CDC is called the US outpatient Influenza-Like-Illness Surveillance network (ILINet). In this study we use the ILINet weekly reports for the state of Texas. This data is downloaded from the DSHS ILI surveillance page website real-time.

2.3 School Data

Because the transmission of the influenza virus is highly correlated with the school calendars, one of the predictor variables that we used in our forecasting model is the school calendar (i.e. the weekly number of days schools are scheduled to be open). We used school calendars of the largest school districts located in each health service regions (HSR) that are available online.

For each of the HSRs we formed the time series for the school calendar predictor, with the school calendars of aforementioned school districts that include the years from 2004 to 2010. We calculated the percent of days in a week that schools are open to convert our school data into weekly aggregated time series.

2.4 Humidity Data

Epidemiological studies show that the transmission of influenza virus and seasonality of influenza outbreaks are strongly correlated with absolute humidity [19]. Therefore we use the weekly humidity reports for related regions as one of the factors that can be defined as a predictor of influenza in the Kalman filter. The data that is used is obtained from the National Climatic Data Center and was converted into weekly averages by taking the averages of the reporting stations located in each health service regions. Then, relative humidity is converted into absolute humidity (Kg Water / Kg Dry Air) using known relationships between temperature and the total mass of water held in the air at 100% saturation.

Chapter 3

Mathematical Model

In this chapter we present the Kalman filter by motivating it as a recursive (also known as online) version of the linear regression model [20]. The linear regression model is given by:

$$\begin{aligned} y &= \theta_0 + \theta_1 f_1 + \ldots + \theta_n f_n + v \\ y &= F\Theta + v \end{aligned} \tag{3.1}$$

Here the unknown but *fixed* variables are $\theta_0, \ldots, \theta_n$ and the input variables[1] are $f_1, \ldots, f_n$. Assume we have m such observations, $y_1, \ldots, y_m$.

The goal in linear regression is to estimate the variables $(\theta_0, \ldots, \theta_n)$. The most commonly used estimator for inference is the ordinary least squares. The least squares method minimizes the sum of squared residuals leading to an analytical expression for the estimated value of the unknown variable Θ:

$$\Theta = (F^T F)^{-1} F^T y \tag{3.2}$$

[1] Also called predictors in Machine Learning literature and referred to as covariates or independent variables in Statistics literature. Conventionally, these are denoted as $x_1, x_2, \cdots, x_n$. But we prefer our notation since the collection of these covariates is denoted as F.

In terms of computation, the least squares method can be thought of as a batch algorithm[2]. This is not very efficient when we get data in real time which motivates the need for a recursive (online) algorithm.

3.1 State Space Model

To develop an online version of the linear regression model we need to understand a State Space model, which we motivate as follows. Suppose we receive the $m+1$ sample $y_{(m+1)}$ and the corresponding $(m+1)$ input denoted as $F_{(m+1),.}$, The linear regression model in Eq. (3.1) must take into account the $m+1$ sample. The $m+1$ sample will cause changes to the values of the unknown variables, $\Theta = (\theta_0, \ldots, \theta_n)$. Now we make a key and useful assumption, namely the vector Θ of unknowns is *random*, not *fixed*. This subtle distinction puts us immediately into a Bayesian framework. Assume then that Θ evolves linearly and is also corrupted by a Gaussian noise. Denote this model for random Θ by:

$$\Theta_t = G\Theta_{t-1} + w_t \tag{3.3}$$

where the index t refers to time[3] and the noise in the system is given by $w_t \sim \mathcal{N}(0, W_t)$. The above equation is called the *system* equation.

The impact of a new estimate for Θ at time $t+1$ then leads to a change

[2]Batch algorithm needs all the data at once.
[3]In our discussion, $t = m + 1$.

in the linear regression equation in Eq. (3.1). The updated equation is:

$$y_t = F\Theta_t + v_t \tag{3.4}$$

The noise in the observation is given by $v_t \sim \mathcal{N}(0, V_t)$. The above equation is called the *observation* equation. It is assumed that the system and observation noises are independent. Eq. (3.3) and Eq. (3.4) together constitute the State Space model.

3.2 Kalman Filter

Denoting $\mathbb{P}\{.\}$ to mean a probability model, the goal of Kalman Filtering is to find $\mathbb{P}\{\Theta_t|y_1, \ldots, y_t\}$. This is obtained using a Bayesian update [10]:

$$\mathbb{P}\{\Theta_t|y_1, \ldots, y_t\} \propto \mathbb{P}\{y_t|\Theta_t, y_1, \ldots, y_{t-1}\} \times \mathbb{P}\{\Theta_t|y_1, \ldots, y_{t-1}\} \tag{3.5}$$

Let us discuss what happens at time t. At $t-1$, the knowledge about Θ_{t-1} is given by:

$$(\Theta_{t-1}|y_1, \ldots, y_{t-1}) \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1}) \tag{3.6}$$

Given we have complete knowledge of the system at $t-1$, we look forward to t in two steps:

1. prior to observing y_t;

2. after observing y_t.

3.2.1 Prior to observing y_t

Before observing y_t, our best choice for Θ_t is given by the system equation Eq. (3.3), $\Theta_t = G\Theta_{t-1} + w_t$. Since Θ_{t-1} is Gaussian and characterized by Eq. (3.6), using the fact that a linear transformation of a Gaussian is Gaussian, we get:

$$(\Theta_t | y_1, \ldots, y_{t-1}) \sim \mathcal{N}(G\mu_{t-1}, R_t) \tag{3.7}$$

where $R_t = G\Sigma_{t-1}G^T + W_t$.

3.2.2 After observing y_t

After observing y_t, one gets the posterior distribution of Θ_t, which is also Gaussian and is given by:

$$(\Theta_t | y_1, \ldots, y_t) \sim \mathcal{N}(\mu_t, \Sigma_t) \tag{3.8}$$

where

$$\mu_t = G\mu_{t-1} + R_t F^T (V + FR_t F^T)^{-1} e_t \tag{3.9}$$

$$\Sigma_t = R_t - R_t F^T (V + FR_t F^T)^{-1} FR_t \tag{3.10}$$

and $e_t = y_t - FG\Theta_{t-1}$

Interpreting μ_t: If the variance of the error $(V + FR_t F^T)$ is large, then use the system's prediction; else incorporate the error which is weighted by the covariance of the error and the prediction given by $R_t F^T$.

Interpreting Σ_t: the variance decreases when we see a sample. This agrees with intuition: as we take more samples, the variance in our estimation should go down.

Note that once Θ_t is obtained, we can calculate y_t using Eq. (3.4). For example, to make a one step ahead prediction of y_t combine Eq. (3.7) with Eq. (3.4) to obtain the following *predictive* probability distribution for y_t:

$$\begin{aligned}(\Theta_t|y_1,\ldots,y_{t-1}) &\sim \mathcal{N}(G\mu_{t-1}, R_t)\\ y_t &\sim \mathcal{N}(FG\mu_{t-1}, FR_tF^T + V)\end{aligned} \tag{3.11}$$

Chapter 4

Implementation Details

In this chapter we provide the details of implementing the above model to predict flu related hospitalizations. In our case, the target (y_t) in Eq. (3.4) is the number of hospitalizations. The input variables or predictors ($F = [f_1, \ldots f_n]$) are given by:

- Google flu trends
- ILINet (Influenza Like Illness Network data)
- Humidity
- School calendar

From the previous chapter, the aims of the analysis is to learn about the coefficients (Θ) of the predictors, and to minimize the error in the predicted number of hospitalizations at each time step.

We implemented the Kalman filter in `R` [16] using the `dlm` package [13, 14]. Some of the text preprocessing was done using `python` [15]. The details of the `R` implementation is given next. The Kalman filter has the following components:

- *Predictors* The predictors discussed above (Google flu trends, ILINet, Humidity and School calender) are implemented using `dlmModReg(predictor)`.

- *Seasonality* The data are collected every week. We need predictors to capture the seasonal component of the weekly nature of the data. This is implemented using the function `dlmModSeas(52)`.

- *Trend* Since there is no overall trend in growth we model this using a random walk plus noise component. This is implemented using `dlmModPoly(order = 1)`.

To facilitate prediction, we lag the predictors by k weeks while we train the filter. To illustrate this consider $k = 4$ weeks. This means, while training the filter, we predict hospitalizations in a given week using the predictors $k = 4$ weeks back. More concretely, in the training phase the hospitalizations in week 11 is predicted using Google flu trends values in week 7. This means given this week's Google flu trends data we can predict hospitalizations $k = 4$ weeks from now. We note that the prediction for k weeks ahead is done using Eq. (3.11), which is run k times. Since we are predicting, we do not have access to the actual hospitalization data and thus cannot do the correction step of Eq. (3.8).

4.1 Feature Selection

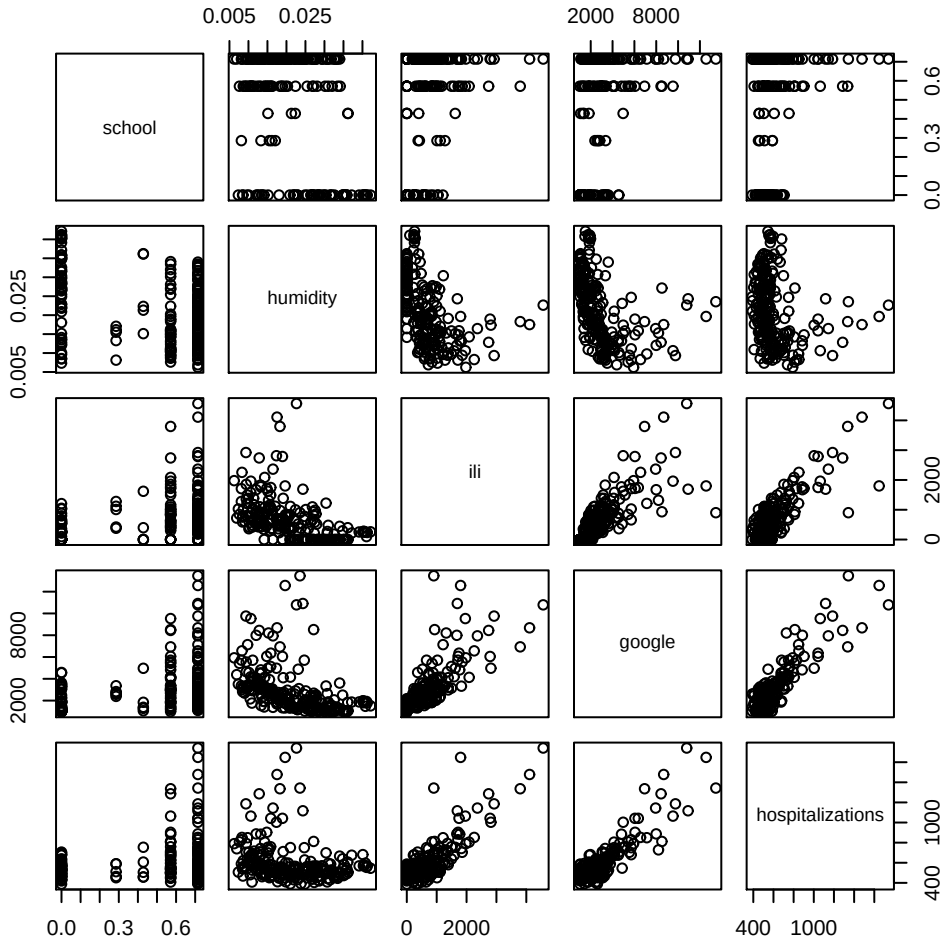
We have four predictors for our model namely: `school`, `humidity`, `ili`, `google`. The correlation of various predictors with `hospitalizations` is shown in Table 4.1.

Table 4.1: Correlation matrix of various predictors with `hospitalizations`. Note that of all predictors, `google` has the highest correlation 0.88 with `hospitalizations`.

	school	humidity	ili	google	hospitalizations
school	1.00	-0.35	0.35	0.32	0.21
humidity	-0.35	1.00	-0.50	-0.44	-0.24
ili	0.35	-0.50	1.00	0.77	0.81
google	0.32	-0.44	0.77	1.00	0.88
hospitalizations	0.21	-0.24	0.81	0.88	1.00

The correlation between various predictors and `hospitalizations` is shown graphically in Fig. 4.1. We can observe that only `ili` and `google` correlate well with `hospitalizations`. Hence we can simplify by choosing these two predictors without losing any accuracy.

Figure 4.1: Scatterplot matrix of all the predictors and hospitalizations. We illustrate reading this matrix through an example. The $(5, 4)$ scatterplot has `google` in the x-axis and `hospitalizations` in y-axis. Thus to read the x-axis look vertically and look at the label in the diagonal. Similarly to lookup y-axis scan horizontally. It is clear from the matrices that `google` and `ili` correlate well with `hospitalizations`. One can observe that the predictors `humidity` and `school` do not correlate well with `hospitalizations`.



Chapter 5

Results

In this chapter, we discuss the performance of our model and compare its performance to two other models namely: Method of Analogues and Serfling. Method of Analogues is a non-parametric method forecasting method developed to predict weather, later adapted for epidemiological projections [21]. It was used to predict flu seasons across France over 10 years. Their results were much more accurate than classical statistical methods such as autoregressive methods. Serfling regression is a parametric regression technique developed to model excess influenza mortality [18] and is widely used in the literature as the baseline model [22].

We present multiple steps ahead forecasts of weekly hospitalizations in addition to the timing of the peak and magnitude of the peak measures.

5.1 Model Performance

The model was fit to historical data from October 2005 to October 2007 and tested for its ability to forecast the state of Texas influenza-related hospitalizations during the 2008-2009 influenza season. We made a series of predictions starting at different time points during the epidemic (between 8

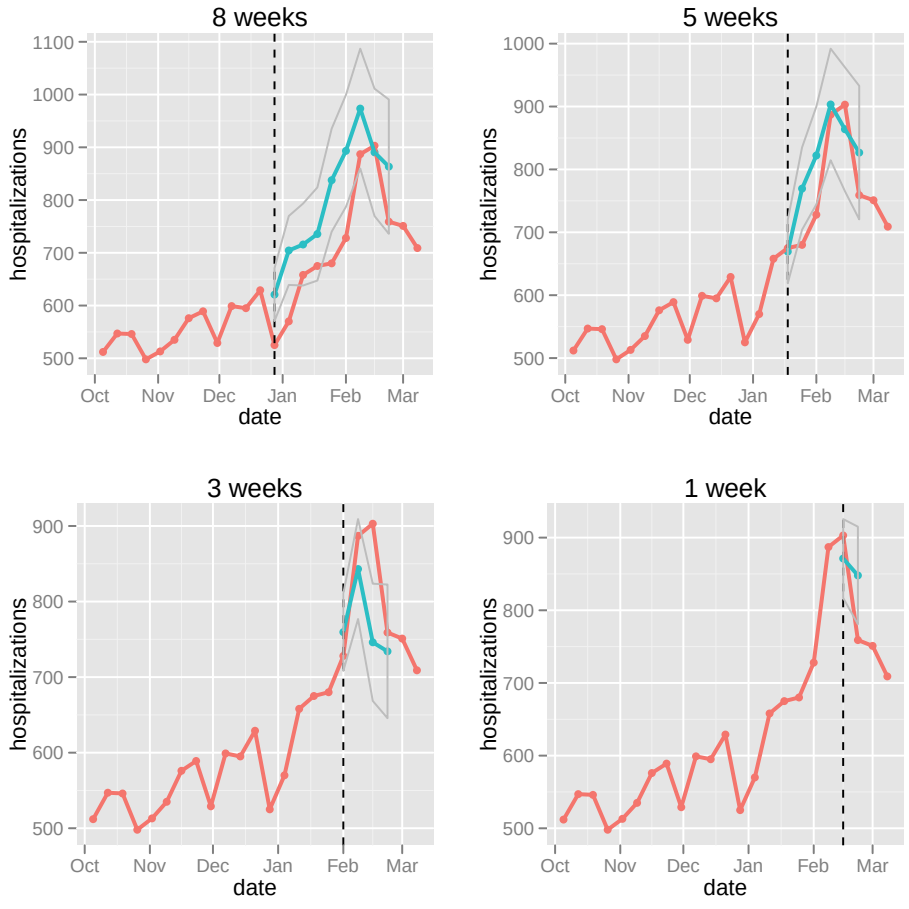
and 1 week prior to peak hospitalizations). The estimate of the peak is offset by one week for predictions earlier than one-week. This is due to the fact that in the past (i.e. the training data) the peak week is usually the 6th week. Since this is a dynamic model it also has the ability to correct itself from past mistakes, through the system equation, leading to improved performance as the size of the training data increases.

Fig. 5.1 shows the predictive power of the model as we approach the peak in 4 time steps. The top left figure shows the 8-week ahead prediction, where we see that the prediction of the peak is offset by a week and the peak value is a little overshot. The next figure shows the 5-week ahead prediction of the peak, where again the peak is offset by a week, but the predicted peak value is very close to the actual value. The next horizon is 3 weeks ahead; here again, the peak is offset by a week and the peak value has decreased from the value in the earlier horizon. The last is the one week ahead prediction where the model captures the peak correctly and the peak value is also close to the actual value.

5.2 Comparison

We compare our results with the results obtained using Method of Analogues and Serfling regression. We provide a brief overview of these two methods before presenting the comparison.

Figure 5.1: Prediction using Kalman Filter The actual hospitalizations values are in red while the predictions using Kalman filter is in blue. The 95% prediction interval for the Kalman filter prediction is shown by a grey bounding box.



5.2.1 Method of Analogues

Method of Analogues works by comparing historical data to current incidence and projects an average value for the n nearest neighbors. Briefly, influenza activity in week t is defined as a vector of influenza incidence from week $t-l$ through week t. We then compare all historical vectors of length $l+1$ and identify the n nearest neighbors using euclidean distance. A k week ahead forecast is accomplished by averaging the kth week ahead from the selected nearest neighbors.

5.2.2 Serfling Regression

Serfling regression fits a local regression term and a seasonal forcing term through a wave function:

$$I_t = \theta_0 + \theta_1 t + \sum \alpha_i \cos\theta + \sum \beta_i \sin\theta \tag{5.1}$$

where I_t is influenza incidence at time t and θ is a linear function of t.

Both the methods were implemented using the time series package in `R`. We can see that the MOA model does reasonably well, however when we compare it in terms of peak week and peak value estimation, the Kalman filter model does better.

Table 5.1 gives the peak error due to Kalman filter, Method of Analogues and Serfling Regression. The results are provided for the year 2008-2009.

Fig 5.2 shows the predictive power of the model graphically. The top left plot in that figure shows 8-week ahead prediction, we see that magnitude

Table 5.1: Comparison of the Kalman filter with Method of Analogues and Serfling Regression. The results are provided for the year 2008-2009. The column "Start of Forecast" has the number of weeks the method forecasts. For example, consider the first row in the Kalman filter results: 8 denotes we have started forecasting 8 weeks before the peak. Since the peak occurred at week 7 of 2009, this means we started forecasting at week 52 of 2008. Note that Kalman filter is the best among the all methods both in terms of peak value predicted as well as time at which the peak is going to occur. Note that the error for the predicted peak value is $\leq 8\%$; the week at which peak occurs is off by at most 1 week in case of Kalman filter. Note that in case of Methods of Analogues, in case of 8 week forecast even though the peak value is very nearly correct the time at which the peak is predicted is off by 2 weeks.

	Start of Forecast	Peak Value	Peak Week	Peak Week Error	% Peak Error
	Observed	903	7	-	-
Kalman filter	8	970.73	6	1	7.50
	5	901.26	6	1	-0.19
	3	863.25	6	1	-4.40
	1	879.34	7	0	-2.62
Method of Analogues	8	905.17	5	2	0.24
	5	735.04	4	3	-18.60
	3	780.65	6	1	-13.55
	1	924.36	8	-1	2.36
Serfling Regression	8	663.99	5	2	-26.47
	5	660.53	5	2	-26.85
	3	660.80	5	2	-26.82
	1	662.16	5	2	-26.67

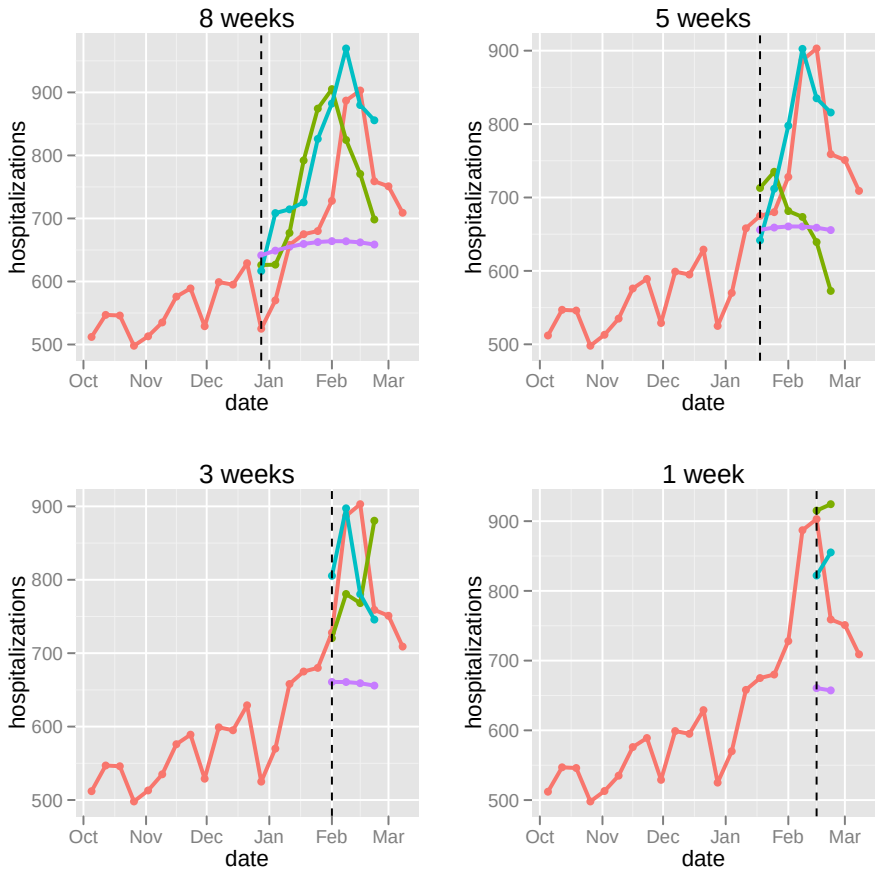
of the peak is a little overshot while the time at which peak occurs is off by a week. The top right panel shows the 5-week ahead prediction of the peak, where the predicted peak value is very close to the actual value but the peak is offset by a week. Bottom left and bottom right panels show the 3-week ahead and 1-week ahead prediction of the peak. The estimate of the peak is offset by one week for predictions earlier than one-week. This is due to the fact that in the past (i.e. the training data) the peak week is usually the 6th week.

Serfling's regression does not capture the dynamics of the system thus resulting in large errors in predicting magnitude of the peak values. While the Method of Analogues (MOA) does capture the dynamics it suffers from large errors in predicting the time at which the peak occurs. For example, consider the case of the 8-week forecast using the Method of Analogues. The peak predicted value is 905 which is very close to the actual value of 903 but the predicted peak is during week 5 which is 2 weeks off from the actual peak which is week 7. Also MOA is not consistent in tracking the hospitalizations. In the 8-week ahead prediction, MOA does track the hospitalizations but in the case of 5-week ahead prediction (top right panel, green color)

Thus Kalman filter is the best among the methods both in terms of predicting the magnitude as well as the time at which peak value is going to occur. Note that the error for the predicted peak value is $\leq 8\%$; the week at which the peak occurs is off by at most one week in case of the Kalman filter.

Next we evaluate the performance of our model for the 8-week ahead forecast using mean absolute percentage error (MAPE) and root mean square

Figure 5.2: Prediction using method of analogues (MOA) plotted in green. and prediction using Serfling regression is plotted in purple. The actual hospitalizations values are in red while the predictions using Kalman filter is in blue. Note that three weeks before the peak hospitalization prediction (lower left pane) MOA predicts the peak a week after it has occurred while Kalman filter predicts a week before it occurs. Also the prediction using Serfling's method does not capture the dynamics of the system.



error (RMSE). The results are tabulated in Table 5.2.

Table 5.2: Kalman filter error metrics. The hospitalization data is from year 2005, week 40 to end of year 2009. The results corresponding to 2007–2008 row are calculated as follows: The training data is from year 2005, week 40 till year 2007, week 51. The testing data for which the error metrics: RMSE, MAPE are reported runs from year 2007, week 52 till year 2008, week 7. The forecasts for 2007-2008 are not as good as those for 2008-2009 season, this is simply due to the size of the training data. LooM denotes "leave one-out method" and uses the entire training data from 2005, week 40 till end of 2009 but leaving an year (say 2006) out for training. This is repeated for years 2007, 2008, 2009 as training data and the results are averaged.

	RMSE	MAPE	Peak Week Error
2007-2008	317.26	39.46	0
2008–2009	90.18	7.50	-1
LooM	155.50	15.65	0.25

We have hospitalization data starting from year 2005, week 40 to the end of year 2009. The results corresponding to 2007–2008 row are calculated as follows: The training data is from year 2005, week 40 till year 2007, week 51. The testing data for which the error metrics RMSE and MAPE are reported runs from year 2007, week 52 till year 2008, week 7. Overall the performance is very encouraging. However, the forecasts for 2007-2008 are not as good as those for the 2008-2009 season; this is simply due to the size of the training data.

5.3 Discussion

For strategic planning and stock piling in the health care industry, we require reasonably accurate forecasts of hospitalizations which are ideally one-month ahead. But the actual hospitalization data are not available every week and it is usually available every six months or in some cases once every year. Thus models such as Method of Analogues and Serfling Regression which depend on past hospitalization data cannot provide accurate forecasts in real time due to the lack of availablility of the recent hospitalization data. Our model overcomes the lack of recent hospitalization data by using time series such as Google Flu Trends, ILINet as predictors of hospitalizations. These predictors correlate well with hospitalization data and are updated in real time on the internet.

Chapter 6

Conclusion

In conclusion, the proposed Kalman filter is the best among the methods compared in this work both in terms of peak hospitalization value predicted as well as the time at which the peak hospitalization is going to occur. The error for the predicted peak value is less than 8% for predictions from 1 to 8 weeks; the week at which peak occurs is off by at most 1 week. In comparison, Method of Analogues suffers upto 19% peak error and can be off by upto 3 weeks; Serfling is even worse in terms of peak error by underestimating it by 26% and the timing can be off by 2 weeks.

Bibliography

[1] Lingji Chen, Harshavardhan Achrekar, Benyuan Liu, and Ross Lazarus. Vision: Towards Real Time Epidemic Vigilance through Online Social Networks. In *First ACM Workshop on Mobile Cloud Computing and Services: Social Networks and Beyond (MCS)*, June 2010.

[2] Benjamin J Cowling, Irene O L Wong, Lai-Ming Ho, Steven Riley, and Gabriel M Leung. Methods for monitoring influenza surveillance data. *International Journal of Epidemiology*, 35(5):1314–1321, October 2006.

[3] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *KDD Workshop on Social Media Analytics*, 2010.

[4] Vanja Dukić, Hedibert F. Lopes, and Nicholas G. Polson. Tracking Flu Epidemics Using Google Flu Trends and Particle Learning. *Working Paper, University of Chicago*, 2011.

[5] Rodica Gilca, Gaston De Serres, Danuta Skowronski, Guy Boivin, and David L. Buckeridge. The Need for Validation of Statistical Methods for Estimating Respiratory Virus Attributable Hospitalization. *American Journal of Epidemiology*, 170(7):925–936, 2009.

[6] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epi-

demics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.

[7] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of ASME: Journal Of Basic Engineering*, 82:35–45, 1960.

[8] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Transactions of ASME: Journal Of Basic Engineering*, 83:95–108, 1961.

[9] A. Krishnamurthy, L. Cobb, J. Mandel, and J. Beezley. Bayesian Tracking of Emerging Epidemics Using Ensemble Optimal Statistical Interpolation (EnOSI). In *Proceedings of the Joint Statistical Meetings (JSM)*, pages 3471–3485, August 2010.

[10] Richard J. Meinhold and Nozer D. Singpurwalla. Understanding the Kalman Filter. *American Statistician*, 37(2):123–127, 1983.

[11] Jimmy Boon Som Ong, Mark I-Cheng Chen, Alex R. Cook, Huey Chyi Lee, Vernon J. Lee, Raymond Tzer Pin Lin, Paul Ananth Tambyah, and Lee Gan Goh. Real-Time Epidemic Monitoring and Forecasting of H1N1-2009 Using Influenza-Like Illness from General Practice and Family Doctor Clinics in Singapore. *PLoS ONE*, 5(4):e10036, 04 2010.

[12] World Health Organization. Influenza (Seasonal) Factsheet, 2009.

[13] Giovanni Petris. An R Package for Dynamic Linear Models. *Journal of Statistical Software*, 36(12):1–16, 2010.

[14] Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli. *Dynamic Linear Models with R.* useR! Springer-Verlag, New York, 2009.

[15] Python Development Core Team. *Python.* Python Software Foundation.

[16] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

[17] Paola Sebastiani, Kenneth D. Mandl, Peter Szolovits, Isaac S. Kohane, and Marco F. Ramoni. A Bayesian dynamic model for influenza surveillance. *Statistics in Medicine*, 25(11):1803–1816, 2006.

[18] Robert E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494–506, June 1963.

[19] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248, March 2009.

[20] H. W. Sorenson. Least Squares Estimation: Gauss to Kalman. *IEEE Spectrum*, pages 63–68, July 1970.

[21] Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the Spread of Influenza Epidemics by

the Method of Analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.

[22] Wenger, Julia B. AND Naumova, Elena N. Seasonal Synchronization of Influenza in the United States Older Adult Population. *PLoS ONE*, 5(4):e10187, 04 2010.

[23] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1997.

Vita

Anurekha Ramakrishnan was born in Madras (now Chennai), Tamil Nadu, India. She received the Bachelor of Engineering degree in Electrical and Instrumentation Engineering from the University of Madras and then worked in Cognizant Technology Solutions as a software engineer. In August 2010, she joined the graduate program in the Department of Statistics and Scientific Computation at the University of Texas at Austin. Currently she is working under the supervision of Prof. Lauren Ancel Meyers in the area of mathematical epidemiology.

Permanent address: anurekha.ramakrishnan@utexas.edu

This report was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.