Influenza-Like Illness(ILI) Forecasting Based on Historical Flu Rate and Corresponding Web Search Query Information

Master's Thesis by

Yiyang Su

MSC WEB SCIENCE AND BIG DATA ANALYTICS

Department of Computer Science

Univerity College London

yiyang.su.17@ucl.ac.uk

Supervisors:

Dr. Lampos Vasileios .

Univerity College London

v.lampos@ucl.ac.uk

Prof. Cox, Ingemar.

Univerity College London

i.cox@ucl.ac.uk

September 2018

This report is submitted as part requirement for the MSc in Web Science and Big Data Analytics at University College London. It is substantially the result of my own work except where explicitly indicated in the text.

Abstract

Influenza is one of the most common infectious diseases we will meet. By the data from World Health Organization, up to 650,000 people died in the world every year because of influenza related diseases. There is an interest in forecasting the prevalence of influenza one or more weeks in advance. Early detection and rapid response can help control the spread of influenza. In addition because of the huge difference between the peak of a flu season and the low of a flu season, a good prediction of the flu season will help hospitals and doctors get prepared for the potential patients. In this thesis, I trained linear models (elastic net regression, ridge regression and least square regression) and obtain the flu rate estimations which are used in autoregression models as input later. I examined various methods for time series forecasting, specifically autoregressive models (seasonal and non seasonal autoregressive integrated moving average models), and mainly focusing on ARIMAX (autoregressive integrated moving average with exogenous variable as input) models.

Acknowledgements

First of all, I do appreciate Dr. Lampos Vasileios's encouragement and guidance helping me starting this project, and his help throughout the whole project.

Secondly, I would like to thank Prof. Cox, Ingemar for offering me the precious opportunity with this interesting project together with his valuable suggestions about this work.

Finally, thanks Google for giving an authorization of using the search query frequency data from Google Health Trend API, the Royal College of General Practitioners and Public Health England for giving the flu rate data (influenza-like illness ILI rate).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter a general background introduction will be contained. Then a brief overview of this report will be followed.

1.1 Background

In Cambridge English dictionary, the word ”forecast” means *”a statement of what is likely to happen in the future”*, and the word phase ”time series” means ” *a list of numbers relating to a particular activity, which is recorded at regular periods of time and then studied. Time series are typically used to study, for example, sales, orders, income, etc.*” Time series forecasting is to predict what is likely happen in the future based on the study of the time series. Technically forecasting is different from estimating. Estimating is usually used to deal with the present situations under specific assumptions, while forecasting attempts to predict what will happen in the future. Individuals can make estimation based on their own judgment, for example salesman can estimate their sales for current week without sufficient historical data and information. A company can forecast their total sales of next 5 years based on historical data, sales trend, sales promotion events and many other information all together.

Flu is highly infectious and dangerous. The spreading of influenza is fast in the peak of a flu season. Huge amount of people can be infected in a short time, for example, from the news released by Department of Health and Mental Hygiene (New York City), 11683 influenza infection was reported in the week that ended on January 28th, 2018. Also influenza is dangerous for vulnerable groups if not treated carefully. From the information [1] provided by World Health Organization, there are up to 650,000 annually death due to influenza like illness. Learning when the flu season would be starting and when will the peak season arriving is quite useful in real world, because early detection of the influenza and quickly appropriate treatment may reduce influenza [2].

The basic way to do a time series forecast is to use only the historical value of this time series to make prediction. For well distributed (data generated by or follows a certain distribution well) and random noise data, the basic way is good enough to give good prediction results. However, the fact is that influenza rate is not perfected fitted to any known distribution. An alternative way is to supplement the historical value of the time series with additional information related to the time series. In my experiments, both the above two ways will be compared.

Influenza-like illness (ILI) is a medical diagnosis of possible influenza or other illness causing a set of common symptoms. Traditional ILI estimates are made by Public Health England (PHE). They are based on data from the Royal College of General Practitioners (RCGP), which provides data weekly on the number of patients visiting doctors with ILI. This thesis examines methods to forecast the RCGP and PHE ILI rates on each day for a continuous period. In this thesis, all ”flu rates” are referring to ILI rates.

The social media post, search engine logs and other on-line user activities are called the on-line user-generated content (UGC). It is useful to use UGC because it contains information not captured by traditional channels [3]. UGC such as related search query frequency has been found beneficial when predicting the flu rate [4]. In my experiments, the Google search query frequency will be used (details in Section 3.1 Data Sets). Google search query frequency is the frequency of queries that users search every day on flu related information via Google Search Engine by Google Health Trends API.

In this thesis, ILI rate forecasting based on daily historical ILI rates and corresponding Google search query frequencies via several methods are compared. The methods are: Auto Regression Integrated Moving Average (ARIMA) and Seasonal Auto Regression Integrated Moving Average (SARIMA) which makes prediction only on historical ILI, ARIMAX with exogenous variables and SARIMAX with exogenous variables which make predictions based on historical ILI rates and selected Google search query frequencies or estimation result from linear models (elastic net regression, ridge regression and least square regression).

1.2 Structure of this Report

The first chapter is a general introduction of the background. Chapter 2 will gives formulations and background knowledge involved with the formulas. Chapter 3 will present the approach implemented and experiments settings. chapter 4 will give the results of each approach implemented, analyze and compare the results of different approached. Chapter 5 will give the conclusion and discuss more on this topic.

Chapter 2

Formulation and Definition

In this chapter the detail formulation of each model and the brief explanation will be presented.

2.1 Linear Models

Predicting ILI rates is a regression problem. For a regression problem, a function f maps a feature space X_{mn} to a target variable y_n where m is the number of candidate search queries and n is the number of days in the sample. As described in Section 3.1 Dataset, our input space X_{mn} is the frequency of m search queries during n days and y_n is the vector representing daily flu rate of n days.

Here is a simple example of linear estimation (use only search query to estimate ILL rate): If we have first 7 day's ILI rate and first 8 day's 3 search query frequencies to estimate the ILI rate of the 8th day. We fit a function f maps X_{3x7} to y_7, then with the parameters of f and the 8th day's search query frequency X_{3x1}, we can estimate the 8th day's ILI rate $\widehat{y}$.

A time series is a sequence of measurement of same variable over time. The observed flu rate y_n is a time series of n days daily flu rate. And our input space X_{mn} each X_{in} where $i \in m$ is a time series of daily search query frequency for search query i on day n.

2.1.1 Elastic Net Model

The reason we select elastic net regression is because we can not handle too many search queries in the whole experiments, especially for autoregressive models. The elastic net regression helps us perform a natural query selection during the regression. Variables in our input space X_{mn} with different search query frequency may correlated with each other. If two search query frequencies are highly correlated with each other, then the elastic net regression may pick only one of them with non zero weight.

Elastic Net [5] is widely used in many research areas. It is a combination of L1-normalization (LASSO [6]) and L2-normalization (Ridge[7]). An elastic net regression can be represented by:

$$\underset{\widehat{w}}{\operatorname{argmin}}(\|y - X\widehat{w}\|_2^2 + 0.5 * (1-\alpha)\lambda \ \|w\|_2^2 + \alpha\lambda\|w\|_1) \tag{2.1}$$

where λ and α control the level of regularization, α is the L1 learning rate and λ is the trade off between L1 norm and L2 norm. λ and α is tuned by selecting minimum root mean square error (RMSE) with a good Pearson Correlation (performance metrics is discussed in section 3.2). The weight matrix w is the weight matrix for each features of input X and y is the real ILI rate. Our target is to find the optimal parameter $\widehat{w}$ that minimizes the equation (2.1).

Since a good pair of λ and α in Elastic Net regression need a lot of time to tune, the common used way is to fix one and tune only one parameter to find out a good pair of λ and α. Lasso regression, Ridge regression and Least Square regression are the simplified form of Elastic net regression. Lasso regression only consider the L1 norm regularization, ridge regression only deal with the L2 norm regularization and Least Square does not include any regularization. These three regression models are easy to find the optimal parameter as in equation (2.2) (2.3) only λ need to be tuned. The following is the objective function of Lasso regression, ridge regression and least square regression.

Lasso regression is to optimize the following objective function:

$$\underset{\widehat{w}}{\operatorname{argmin}}(\|y - X\widehat{w}\|_2^2 + \lambda \ \|w\|_1) \tag{2.2}$$

Ridge regression is to optimize the following objective equation:

$$\underset{\widehat{w}}{\operatorname{argmin}}(\|y - X\widehat{w}\|_2^2 + \lambda \ \|w\|_2^2) \tag{2.3}$$

Least Square is to optimizing the following objective function:

$$\underset{\widehat{w}}{\operatorname{argmin}}(\|y - X\widehat{w}\|_2^2) \tag{2.4}$$

2.2 Autoregressive Models

Autoregressive models use the value of previous time series to predict the future time series.

In AR (p) model, the present value is the sum of p weighted past values.

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + u_t \tag{2.5}$$

where δ is a constant, u_t is the white noise and X_t is the time series data.

2.2.1 ARIMA (Auto-regressive integrated moving average)

ARIMA has been broadly used for forecasting time series for a long time. It is a non-seasonal model proposed by Box and Jenkins (1970) [8, 9].

ARIMA is quite powerful for time series with clear trend and random noise. It is evaluated based on auto-regressive moving average (ARMA) model. ARMA model is a combination of Auto regression(AR) and Moving average(MA) models.

MA(q) model assumes that the present value is the moving average from q steps before.

$$X_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + ... + \theta_q u_{t-q} \tag{2.6}$$

A back shift operator B is a useful notation when working with time series lags. $X_t B = X_{t-1}$. (”lag” in time series means the time delay, for example, lag=1 means all information you can access is 1 day before the objective day.)

Applying back shift operate for equation (2.6) we can have equation (2.7).

$$X_t - \mu = u_t(1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q) = \Theta(B) u_t \tag{2.7}$$

where μ is the average, u_t is the white noise and X_t is the time series data.

ARMA(p,q) is the simple combination of AR(p) and MA(q):

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + ... + \theta_q u_{t-q} \tag{2.8}$$

Applying back shift notation on equation (2.8) we can have the following equations.

$$\Phi(B)X_t = (1-\theta_1 B+\theta_2 B^2+...+\theta_q B^p)X_t \tag{2.9}$$

$$\Phi(B)X_t = \Theta(B)u_t+\theta \tag{2.10}$$

ARIMAX(p,d,q) is an extension of ARMA(p,q)
it takes d times differential of ARMA(p,q):

$$\Phi(B)\Delta^d X_t = \Theta(B)u_t\Delta^d+\theta \tag{2.11}$$

$\Delta X_t = X_t - X_{t-1} = X_t - BX_t = (1-B)X_t$
$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1} = (1-B)X_t - (1-B)X_{t-1} = (1-B)^2 X_t$
$\Delta^d X_t = (1-B)^d X_t$
Thus ARIMA(p,d,q) can be represented by:

$$(1-\sum_{i=1}^{p}\phi_i B^i)(1-B)^d X_t = (1+\sum_{j=1}^{q}\theta_j B^j)\varepsilon_j$$

where ε_i is the noise. p is the number of autoregressive terms, d is the number of non-seasonal difference, q is the number of moving average term and B is the back shift operator.

2.2.2 SARIMA(Seasonal Autoregressive integrated moving average)

A SARIMA model can be presented as :
$ARIMA(p,d,q)(P,D,Q,s)$ where the lower case (p,d,q) represents the non-seasonal part of the model, and the (P,D,Q,s) represents the seasonal part of the model.
P is the number of seasonal AR terms, D is the number of times of seasonal differencing, Q is the number of seasonal MA terms, s is the time span of repeating seasonal pattern.
SARIMA is the ARIMA including seasonal terms. Seasonal term (P,D,Q,s), (P,D,Q) is similar to non-seasonal part of ARIMA(p,d,q) and simply multiply to original

equation and undertake back-shifts of seasonal period s.

SARIMA(p,d,q)(P,D,Q,s) can be written as:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)\varepsilon_t \tag{2.12}$$

ε_t is the noise term and B is the back shift operator,

$\phi(B) = 1 - \phi_1(B) - \phi_2(B^2) - ... - \phi_p(B^p)$

$\Phi(B) = 1 - \Phi_1(B) - \Phi_2(B^2) - ... - \Phi_P(B^P)$

$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$

$\Theta(B) = 1 + \Theta_1 B + \Theta_2 B^2 + ... + \Theta_Q B^Q$

2.2.3 ARIMAX

ARIMAX is ARIMA with X as exogenous variables [10]. The exogenous variables include information apart from the original time series. ARIMAX models absorbing information from exogenous variables may be better than ARIMA model with complementary information containing in the exogenous variables.

For time series data y_t and M exogenous data X_t, $ARIMAX(p,d,q)$ can be represented as :

$$\Delta^d y_t = \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{m=1}^{M} \beta_m X_{m,t} + \varepsilon_t \tag{2.13}$$

$\varepsilon_t \sim N(0,\delta^2)$, p is the number of autoregressive term, d is the degree of differencing, q is the number of moving average term.

The expanded expression[11] of ARIMAX is:

$$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_M X_{M,t} + \frac{1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p} \tag{2.14}$$

With the expanded expression (2.14) we can understand why adding more exogenous variables may cause worse forecasting, since each exogenous variable we add may cause the weight β_i of most efficient variables decrease.

2.2.4 SARIMAX

SARIMAX is the SARIMA model including exogenous variables or ARIMAX model including seasonal term [11].

For time series y_t and M exogenous variable X_t SARIMAX can be represented by:

$$y_t = \beta_0 + \beta_1 X_{1,t} + \ldots + \beta_M X_{M,t} + (\frac{1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p})(\frac{1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}}{1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}}) \tag{2.15}$$

In equation (2.15): p is the number of non-seasonal AR terms, d is the number of times of non-seasonal differencing, q is the number of non-seasonal MA terms, P is the number of seasonal AR terms, D is the number of times of seasonal differencing, Q is the number of seasonal MA terms, s is time span of repeating seasonal pattern. β is the weight of exogenous variables.

2.3 Literature Review

In [12], the author comparing the predictive power among ARIMA and Random Forest models on predicting outbreaks of avian influenza H5N1 in Egypt. Random forest is a machine learning algorithm that can be used for time series forecasting. They obtained the Avian influenza outbreak data from the on-line EMPRES-i Global Animal Disease Information System and daily temperature and relative humidity data from the Weather Underground website from 2005-12-08 to the 2012-10-28. They put the historical H5N1 outbreak data and the weather information together as the input, then compared the performance of ARIMA and random forest models and concluded that random forest model is better for prediction H5N1 outbreak in Egypt.

In [4], the author explains how to detect influenza using Google search engine query data. They found that the possibility a doctor visits a patient with ILI symptoms is highly correlated with relative frequency of certain search queries one day before. With this clues they can use the Google search query information to detect the spread of influenza in areas with a large population of web search users.

Chapter 3

Methods and Experiments Setup

In this chapter the details of setup of each method and experiment will be presented.

3.1 Datasets

The original search query frequency data is from Google Health Trends API, The flu rate data (influenza-like illness ILI) is from the Royal College of General Practitioners (RCGP) and Public Health England (PHE).

A search query frequency is the probability of a daily total searches of a query for a specific region and temporal resolution normalized by the sum of daily searches of all candidate queries at the same day. In this work search frequency data is the selected 1000 most related queries (which are most correlated to flu activities) from 35572 search queries [13], daily from August 24, 2005 to August 23, 2017.

RCGP and PHE provides daily ILI rate from doctor consultation reporting per 100,000 people in England from August 24, 2005 to August 23, 2017.

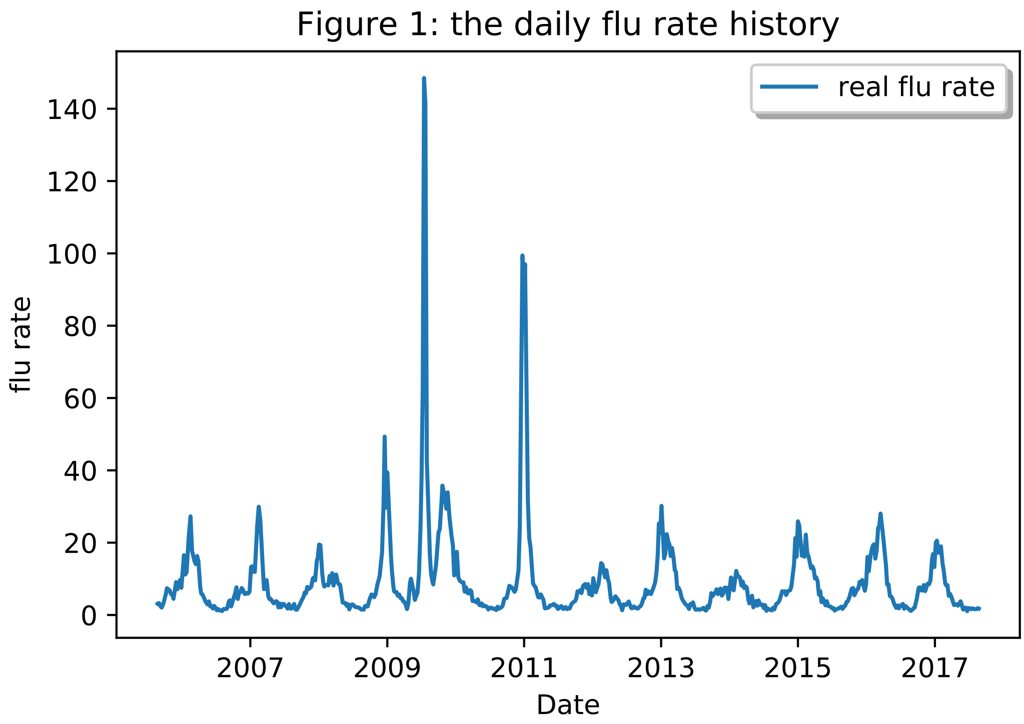

Figure 3.1: The daily ILI rate of England from August 24, 2005 to August 23, 2017

In oder to obtain a more general view of how each model performs. We form two data sets such that we can see the model's general performance on the average result of these two data sets.

In the experiments following two data sets are formed:

Dataset1 using the first 3653 days data as training and validation set (August 24, 2005 to August 23, 2015), and the next 365 days as testing set (August 24, 2015 to August 23, 2016.)

Dataset2 using the first 4018 days data as training and validation set (August 24, 2005 to August 24, 2016) and the next 365 days as testing set (August 24, 2016 to August 23, 2017)

3.2 The Performance metrics

In order to compare the performance of different models, we use the following four metrics: mean absolute error (MAE) in section 3.1.1, root mean squared error (RMSE) in section 3.1.2, Pearson Correlation in section 3.1.3 and Akaike information criterion (AIC) in section 3.1.4

3.2.1 Mean Absolute Error(MAE)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y}|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{3.1}$$

where n is the total number of samples, y_i is the observed value, $\widehat{y}$ is the predicted value and e_i is the error term. [14]

MAE gives a direct view of how precise our prediction is.

3.2.2 Root Mean Squared Error(RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y})^2}{i}} \tag{3.2}$$

where n is the total number of samples, y_i is the observed value and $\widehat{y}$ is the predicted value. [15]

RMSE may be affected by extreme values since we square the error before root the sum of them. RMSE is an adjust way from MAE to present a more general error. In my experiments RMSE is the major reference of how accurate a prediction is.

3.2.3 Pearson Correlation(CORR)

$$r_{X,Y} = \frac{cov(X,Y)}{\delta_X \delta_Y} \tag{3.3}$$

where $cov(X,Y)$ is the covariance of X and Y, δ_X is the standard deviation of X and δ_y is the standard deviation of y. [16]

In this work, it is useful to know the trend of flu season. Pearson Correlation measures the linear relationship between our prediction and real ILI rate, it states how the movement of our prediction and real ILI rates are associated. More fluctuate predictions may give better MAE and RMSE, but it is not what we want here. Thus even very low MAE and RMSE but poor correlation is not a good prediction. The prediction of a model have very high correlation with the real flu rate indicates that this model predicts the flu season trend very well.

3.2.4 Akaike Information Criterion(AIC)

The AIC formula for regression models:

$$AIC = 2k + nlog(\frac{RSS}{n}) \tag{3.4}$$

where k is the number of estimated parameters in the model, n is the number of observations and RSS is the residual sums-of-squares.

AIC is widely used in model selection as an estimator to measure the relative quality of statistical models [17, 18]. In my experiments, AIC is the major consideration of how well an autoregressive model is.

3.3 Linear Models

3.3.1 Settings

3.3.1.1 Train-validation Splitting

As mentioned in section 3.1 Datasets: The original data set forms two data sets: Dataset1 and Dataset2.

For linear models, divide approximately 10% of each training set as validation set. Since we find the flu rate only change very little from day to day, we choose two continuous periods of days as validation set. If we choose validation days randomly, the error will be very low since we know the ILI rate on the day before and the day after and the change from day to day is very small.

Dataset1: A set of 3653 days (approximately 10 years) as training and validation set, thus choose two 180 continuous periods from $\frac{1}{3}$ of the set (day 1218 to day 1398) and $\frac{2}{3}$ of the set (day 2436 to day 2616) as the validation set and the use the rest of this set as training set. The test set for dataset1 is the 11th year from August 24, 2015 to August 23, 2016.

Data set2: A set of 4018 days (approximately 11 years) as training and validation set, thus choose two 200 continuous periods from $\frac{1}{3}$ of the set (day 1340 to day 1540) and $\frac{2}{3}$ of the set (day 2679 to day 2879) as validation set and the remaining of the set as training set. The test set for dataset2 is the 12th year from August 24, 2016 to August 23, 2017.

3.3.1.2 Candidate Query Selection

(a) From the Google health trend API as described in Section 3.1 Datasets, we obtained 1000 search query frequencies. These 1000 search query frequencies are in a descending order with the correlation to ILI activity.
(b) We need to choose a subset of n queries as 1000 is too large. For each model we use the first 500 queries frequencies, the first 250 query frequencies and the first 100 queries frequencies in progress. These 500, 250, 100 query frequencies are the candidate query frequencies used as my original input.
(c) In my experiments of linear models, I will apply a Pearson Correlation Filter to reserve only the part of candidate query frequencies that are correlated with flu rate(y) than a fixed lower bound.
For example, from the 1000 search query frequencies, 500 candidate search query frequencies are chosen as input, after the Pearson Correlation filter setting the correlation lower bound equals 0.2 (which means only the candidate queries which their correlation with flu rate (y) is greater than or equal to 0.2 can be reserved), 222 search query frequencies are left, then use these 222 search query frequencies as the input of the model.

3.3.1.3 Standardization

Standardization is to put different variables on the same scale. Standardization will be useful when dealing with high order terms and interaction terms of polynomials. For models where the hyper-parameters are well tuned the difference of error between standardizing or not is tiny. In our case, we do not have high order to interaction terms in our search query frequencies data and experiments results showing that standardization does not help improve our prediction error. For convenience, in all my experiments, I will not standardize the data.

3.3.2 Elastic Net Models

As the formula in Section 2.1.1 Elastic Net Model, α and λ are the hyper-parameters of the elastic net model. Since there are infinite combinations of α and λ and our purpose is to find a generally good set of elastic net parameters which can provide a good estimation. In order to obtain a good estimation result from elastic net regression model quickly, I fixed L1-rate α equals 0.3 and tuning for best λ (this will give a local minimum error) .

3.3.3 Ridge and Least Square Models

In my experiments with elastic net regression not fixing L1 ratio α, many results indicate α should equal 0, which is ridge ($\alpha = 0$) regression. Conduct Ridge and Least Square regression under the same settings as elastic net models and analyze the performance of them.

3.4 Autoregressive Models

3.4.1 Settings

For this part, there is no need to draw validation sets from the training set since we choose model based on the Akaike Information Criterion (AIC) [17].

3.4.2 ARIMA and SARIMA

For these two models, first tuning (p,d,q) for ARIMA, (p,d,q)(P,D,Q,s) for SARIMA, select best model based on AIC.

For the SARIMA model, analyzing autocorrelation graph and the decomposition of flu rate as trend, seasonal and residual term may help find appropriate periods (s term).

Selecting the model with lowest AIC and make prediction on different lags. lags equals n meaning that using the data n days before (fixing training size and shift the time window) to predict the current day's flu rate. For example, for Dataset1, lags=10 means using first 3644 days to predict the flu rate of the 3653th day, then the next 3644 days (second day to 3645th day) to predict the flu rate of the 3654th day and so on.

In my experiments lags=1,5,10,14,28 will be compared.

3.4.3 ARIMAX and SARIMAX

First use the autocorrelation plot to find out reasonable periods.

Then assume flu rate can be decomposed into trend, seasonal and residual term with the above reasonable periods, plot the decomposition to see if the period works well.

I first try to use first 1, 3 ,10 search queries (which 1, 3, 10 most correlated queries to flu activity) frequencies as the base result for comparison, then use historical average flu rate as exogenous variable, finally use the estimation result from elastic net models as exogenous variable.

For the part use the prediction from elastic net models as exogenous variable, use the first 4 years apply elastic net model to predict the 5th year's flu rate, save the prediction, then use the 4 years data (second year to 5th year) to predict the 6th year, save the 6th year's prediction and so on until the end of training set. Putting all the estimation results from elastic net model together (5th year, 6th year,...) as the exogenous variable of ARIMAX model, then tuning the new ARIMAX model and make predictions.

The same as ARIMA and SARIMA, in my experiments I will compare the results of different lags(1,5,10,14,28).

Chapter 4

Results and Analysis

In this chapter, the result of each experiments discussed in previous section will be presented and followed by an analysis of the model based on the results.

4.1 Results of Linear Models

Note:

Test set1 is the test set in dataset1 from August 24, 2015 to August 23, 2016 and the test set2 is the test set in dataset2 from August 24, 2016 to August 23, 2017.

MAE means mean absolute error, RMSE means root mean square error, r means correlation.

We can not determine which model is the best by only the result from Dataset1 or Dataset2 because we change the training length and testing period together. But by comparing the average error of the two datasets we can obtain the general performance of the model.

		Test Set 1				Test Set 2				Average of 2 sets		
	Candidate Queries	number of picked queries	MAE	RMSE	r	number of picked queries	MAE	RMSE	r	MAE	RMSE	r
Elastic Net	500	42	2.698	3.388	0.912	77	2.703	3.293	0.914	2.701	3.340	0.913
	250	39	2.739	3.382	0.907	6	3.499	4.019	0.826	3.119	3.700	0.867
	100	26	2.905	3.562	0.890	5	3.491	4.007	0.827	3.198	3.784	0.858
Ridge	500	222	2.751	3.332	0.915	212	2.818	3.402	0.920	2.785	3.367	0.918
	250	163	2.643	3.245	0.914	149	3.295	3.718	0.902	2.969	3.482	0.908
	100	65	3.260	3.894	0.871	56	3.433	3.972	0.891	3.347	3.933	0.881
Least Square	500	222	2.515	3.270	0.900	212	2.691	3.380	0.911	2.603	3.325	0.906
	250	163	2.438	3.162	0.905	149	2.862	3.492	0.913	2.650	3.327	0.909
	100	65	2.665	3.388	0.894	56	3.115	3.728	0.910	2.890	3.558	0.902

Table 4.1: Results of Linear Regression Models

Table 4.1 shows the result of Linear regression models (the Elastic net regression, the ridge regression and the least square regression models). The results with blue is the lowest average error or highest average correlation of the two datasets among the three models. Candidate queries are the original input size of the queries and the number of picked queries is the number of non zero weights of candidate queries after the natural feature selection performed by elastic net regression.
The least square regression with 500 candidate queries has the lowest average MAE= 2.603 and lowest average RMSE= 3.325 among the three models. This result shows that the regularization term is not that important for linear estimation. Further more, the results from Table 4.1 may because least square regression is the easiest one to tune and the elastic net models are not optimized. The elastic net models are generally good and theoretically best. As described in Section 3.3.2, the L1-ratio α is fixed at 0.3 and the Pearson correlation filter is set a lower boundary of 0.2. Only local best parameter of the model is found. The local best parameters are not guaranteed to be the absolute best parameter. If change the L1-ratio, we can get better results by tuning (e.g. For 500 candidate, Pearson correlation filter $r >= 0.1$, L1-ratio α = 0.2, λ = 9.9, gives result MAE= 2.481, RMSE = 3.133, correlation r = 0.916 which is better than the results we fixing $r >= 0.2$, L1-ratio = 0.3. However, most of my experiments shows that the L1 ratio λ of elastic net should be close to zero which indicates least square regression. The estimation from elastic net regression should be at least same as the estimation by least square regression because least square is only one possible variation of elastic net). We can not promise any parameter pairs of elastic net is absolute best because too many possible parameters combinations need to tune. Therefore, fixing L1-ratio and Pearson correlation filter boundary can help find a locally best λ such that we can obtain a good estimation by the model quickly.

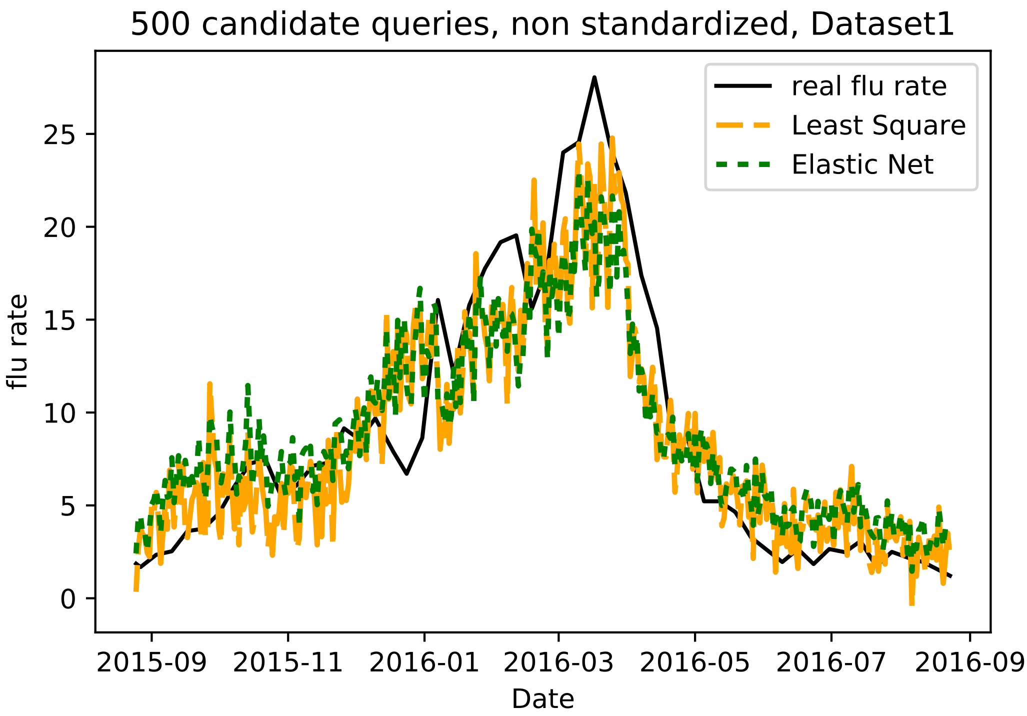

Figure 4.1: Daily estimation of flu rate based on Dataset1

Figure 4.1 shows the estimation results of least square regression and elastic net regression on dataset1 with 500 candidate queries. The black line is the real flu rate, the dashed orange line is the estimation made by least square regression and the dashed green line is the estimation made by elastic net regression.

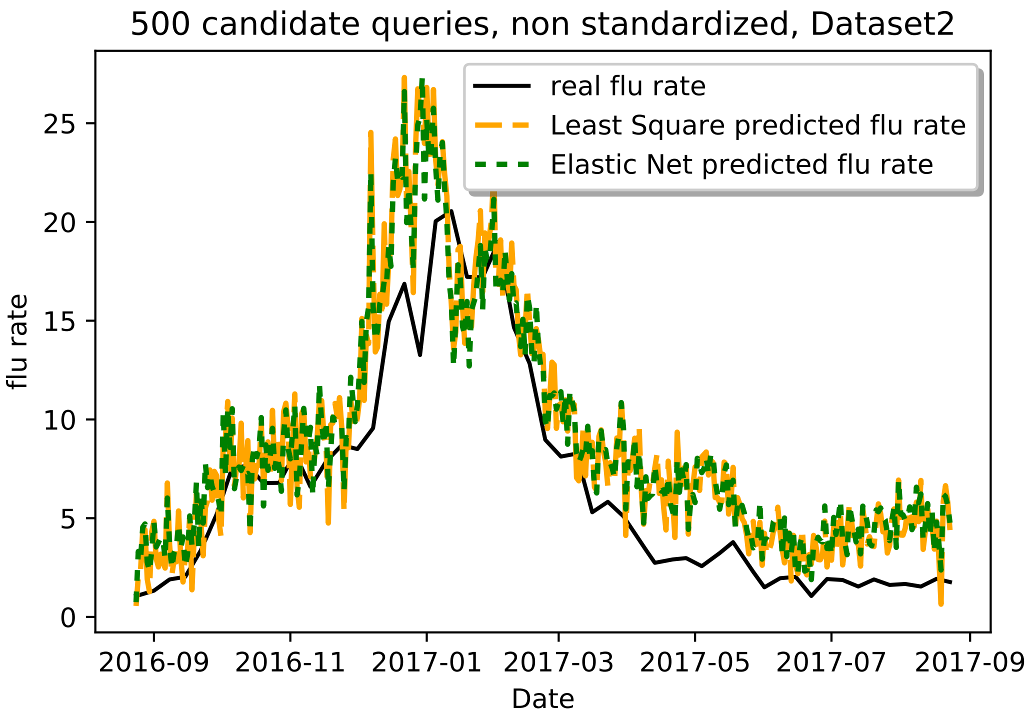

Figure 4.2: Daily estimation of flu rate based on Dataset2

Figure 4.2 shows the estimation results of least square regression and elastic net regression on dataset2 with 500 candidate queries. The black line is the real flu rate, the dashed orange line is the estimation made by least square regression and the dashed green line is the estimation made by elastic net regression.

From Table 4.1, the least square model with 500 candidate queries (real input is 222 queries after Pearson correlation filter) has the lowest average MAE= 2.603 and lowest average RMSE= 3.325 among the three models. From Figure 4.1 and Figure 4.2 we can observe that the elastic net and the least square estimations are similar in most times. The least square estimation (The orange line) fluctuates more than elastic net estimations. As we known the fact is that the daily change of ILI rate is very small, the elastic net estimation may be more useful as an input to auto-regression model.

4.2 Results of Autoregressive Models

4.2.1 Analysis of Autocorrelation

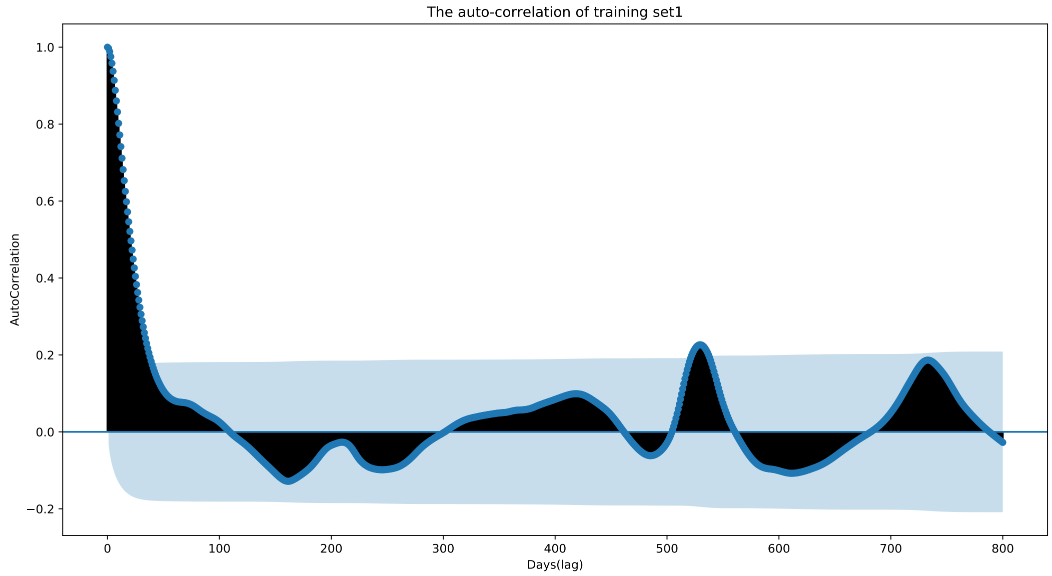

Figure 4.3: The autocorrelation plot of first 800 lags

The Figure 4.3 is part of the autocorrelation plot of the training set in dataset1 (there is no big difference between the autocorrelation plot of the training set of dataset1 and dataset2). The horizontal value is the number of lags (days) and the vertical value is the correlation measurements. The shadow area is the 95 % confidence interval. The lags with autocorrelation outside the shadow area (confidence interval) may indicate possible periodicity choices.

Autocorrelation measures the similarity of one time series against a delay of this time series which has a time lag n and give out statistics to determine whether they are correlated or not. The analysis of autocorrelation is an effective way to find the patterns in the time series. [19]

From Figure 4.3, we notice that lags from 1 to 42, lags from 526 to 534 and lags from 724 to 744 have autocorrelation values outside the cone shadow area. lag=1 has the highest autocorrelation= 0.97036 against the present time series, this is because for the flu rate data, the change from day to next day is quite small. lag= 530 has the local maximum Autocorrelation= 0.2056, and lag= 733 has the local maximum Autocorrelation= 0.1912. lag= 733 which is nearly 2 years may be a good choice of period. The first Autocorrelation= 0 happens at lag between 108 and 109, but 2*108 is still negative, it does not indicate half the period. The first positive peak is near 420, but 820 is negative, so 420 is not a good choice of period. 1 year is positive and 2 year is nearly a peak which indicates 1 year may be a choice of period. From the result of Autocorrelation, the flu rate data does not show strong periodicity characteristic, but 1 year and 2 year may be useful period choices.

4.2.2 Analysis of Decomposition

A trend T exists if there is a long term increasing or decreasing direction on the dataset. A seasonal S exists when a typical pattern occurs several times in a time series over fixed period. A noise term is the residual after trend and seasonal terms been removed.

Assume the flu rate y can be decomposed into trend T, Seasonal term S and the noise term ε.

$$y_i = T_i + S_i + \varepsilon_i$$

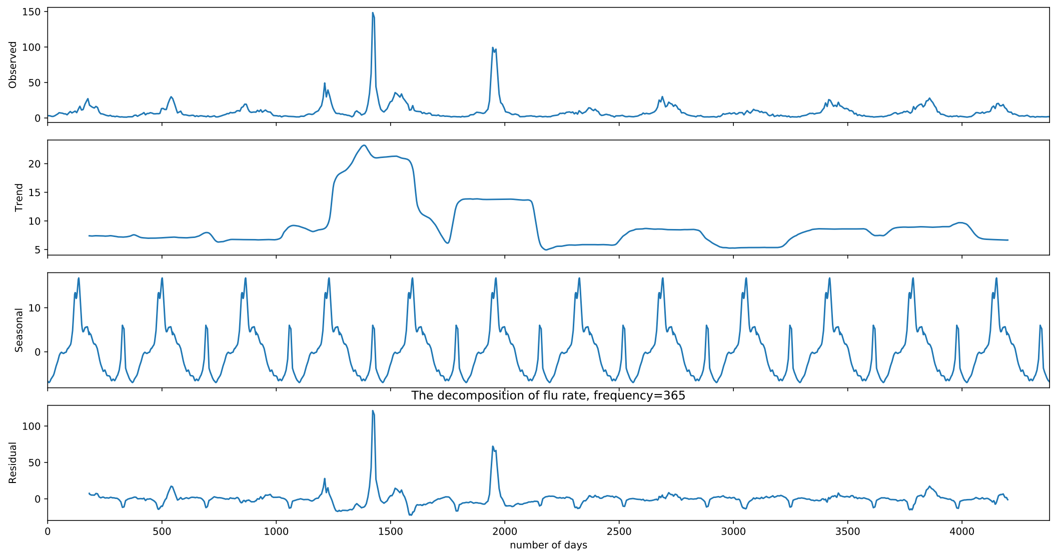

Figure 4.4: The decomposition of flu rate, period=365

Figure 4.4 shows the decomposition of flu rate assume the period is 365 days. The top sub-figure shows the original data before decomposition. Below the top sub-figure is the decomposition of trend. The third sub-figure from top shows the decomposition of seasonality. The bottom sub-figure shows the residual plots which is the original data minus trend and seasonality.

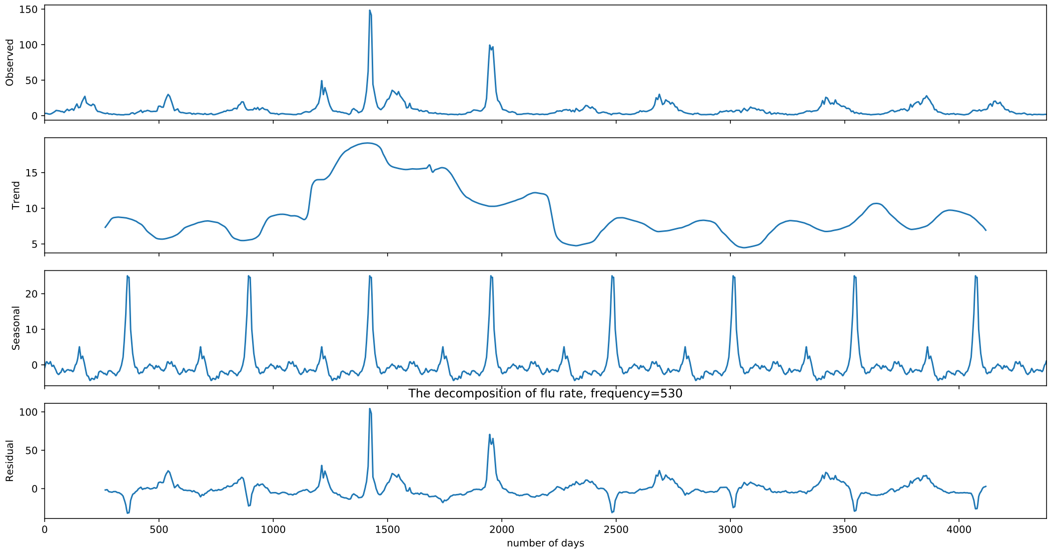

Figure 4.5: The decomposition of flu rate, period=530

Figure 4.5 shows the decomposition of flu rate assume the period is 530 days. The top sub-figure shows the original data before decomposition. Below the top sub-figure is the decomposition of trend. The third sub-figure from top shows the decomposition of seasonality. The bottom sub-figure shows the residual plots which is the original data minus trend and seasonality.

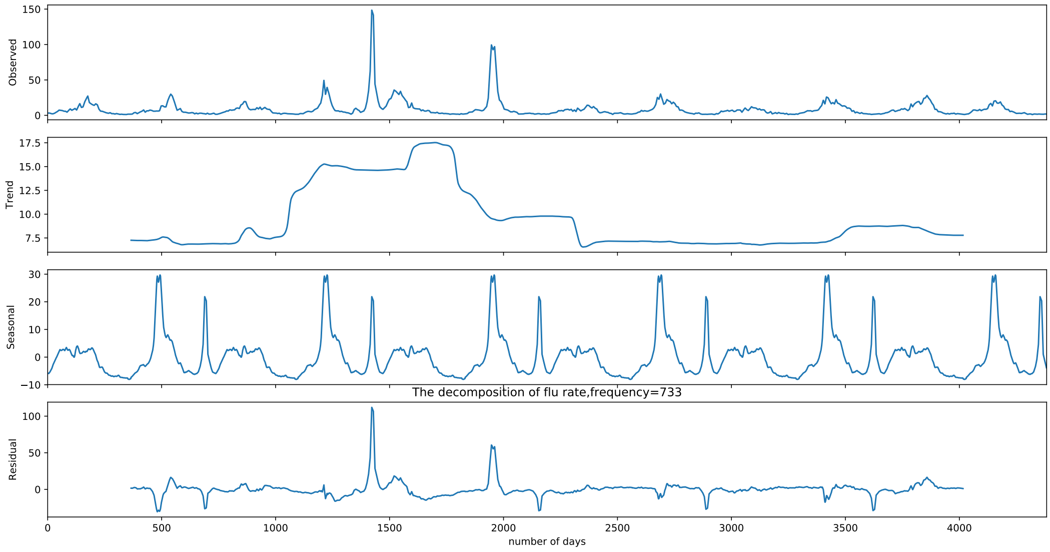

Figure 4.6: The decomposition of flu rate, period=733

Figure 4.6 shows the decomposition of flu rate assume the period is 733 days. The top sub-figure shows the original data before decomposition. Below the top sub-figure is the decomposition of trend. The third sub-figure from top shows the decomposition of seasonality. The bottom sub-figure shows the residual plots which is the original data minus trend and seasonality.

Figure 4.5, 4.6 and 4.7 shows the decomposition of flu rate assuming the period is 365, 530 and 733 days. Neither of them indicates a possible clear trend function. They all present seasonal pattern over fixed period.

The results of decomposition indicates that there is no clear trend function for the flu rate, but include a seasonal term may help improve the models.

4.2.3 Analysis on Periodically Average Flu Rate

A recent paper shows that the previous year's weekly average flu rate is useful when predicting weekly flu rate. [20]. This part obtained the estimation based on the historical average flu rate of 365 days period, 530 period and 733 days period. This 3 periods are the possible good choice of period concluded in Section 4.2.1.

		Test Set1			Test Set2		
	Period	MAE	RMSE	r	MAE	RMSE	r
Average Model	365	5.332	7.451	0.370	3.071	4.026	0.848
	530	5.682	7.183	0.186	6.791	8.833	0.052
	733	3.735	5.596	0.683	5.017	8.350	0.709

Table 4.2: Result Table of Using only Periodical Average

Table 4.2 shows the estimation error of using the historical average flu rate of period 365 days, 530 days and 733 days directly as the estimation. The values with blue is the lowest MAE and RMSE or highest Correlation.

From Table 4.2, the estimation result is not as good as the estimation made linear models (500 candidate query, elastic net regression has MAE= 2.698, RMSE= 3.388, r= 0.912 for dataset1 and MAE= 2.703, RMSE= 3.293, r= 0.914 for dataset2) which indicates query information is useful. It may be beneficial to use 365 days and 733 days historical average flu rate as input to auto-regression models.

4.2.4 Results of ARIMA models

	Test Set1			Test Set2		
lag	MAE	RMSE	CORR	MAE	RMSE	r
1	0.058	0.143	0.9998	0.041	0.108	0.9998
5	0.679	1.066	0.9894	0.481	0.798	0.9899
10	1.796	2.461	0.9448	1.300	1.847	0.9465
14	2.339	3.175	0.9113	1.895	2.486	0.9074
28	4.396	5.574	0.7106	3.521	4.081	0.7856

Table 4.3: Result Table of ARIMA(1,1,14)

Table 4.3 shows the prediction error of ARIMA(1,1,14) on dataset1 and dataset2. lag means how many days ahead our prediction is made, for example, lag= 5 means we use the information till now to forecast the flu rate 5 days later.

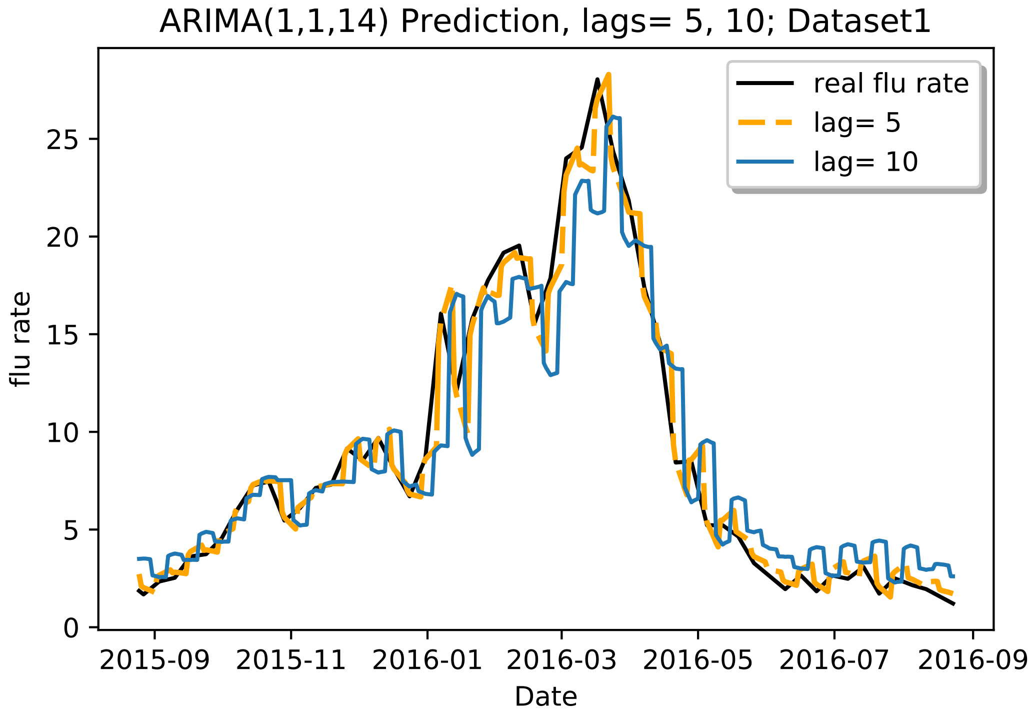

Figure 4.7: ARIMA(1,1,14) forecasting based on Dataset1

Figure 4.7 shows the flu rate forecasting result of 5 days and 10 days future by ARIMA (1,1,14) model on Dataset1. As in Figure 4.7, the black line is the real flu rate, the orange dashed line is the prediction made 5 days ahead (lag= 5), and the the solid blue line shows the forecasting of 10 days ahead (lag= 10).

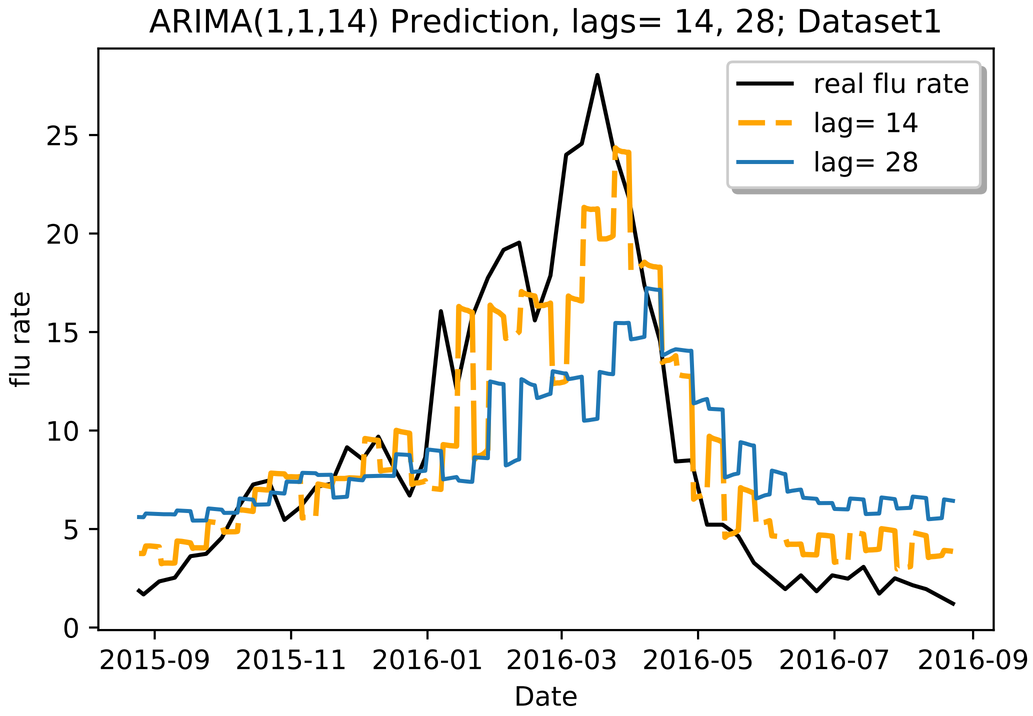

Figure 4.8: ARIMA(1,1,14) forecasting based on Dataset1

Figure 4.8 shows the flu rate forecasting result of 14 days and 28 days future by ARIMA (1,1,14) model on Dataset1. As in Figure 4.8, the black line is the real flu rate, the orange dashed line is the prediction made 5 days ahead (lag= 14), and the the solid blue line shows the forecasting of 10 days ahead (lag= 28).

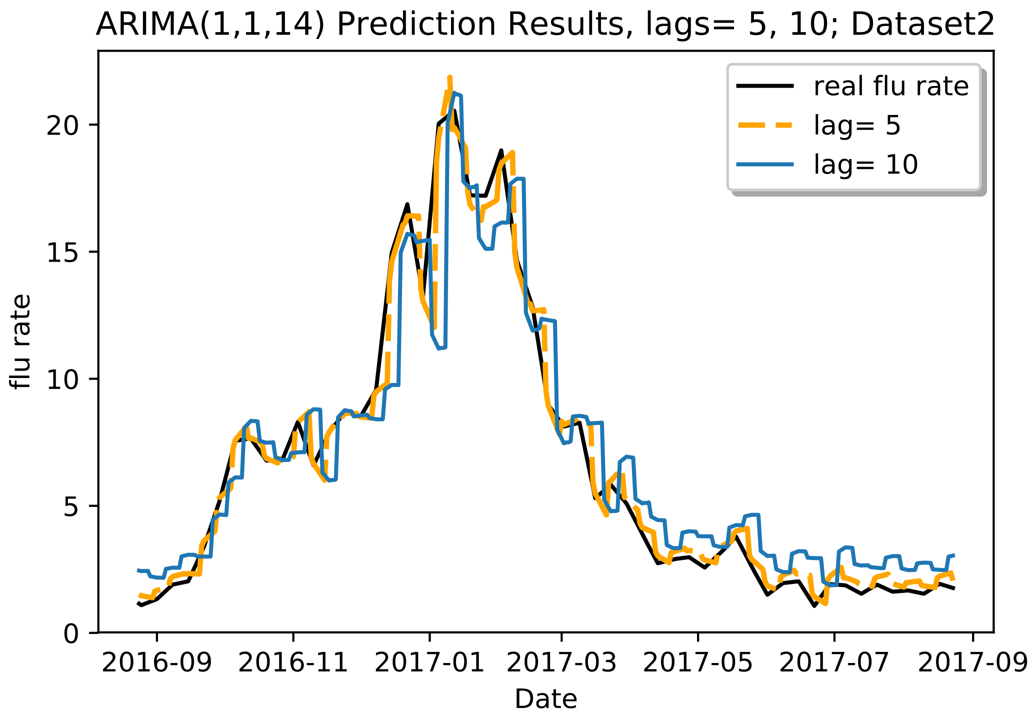

Figure 4.9: ARIMA(1,1,14) forecasting based on Dataset2

Figure 4.9 shows the flu rate forecasting result of 5 days and 10 days future by ARIMA (1,1,14) model on Dataset2. As in Figure 4.9, the black line is the real flu rate, the orange dashed line is the prediction made 5 days ahead (lag= 5), and the the solid blue line shows the forecasting of 10 days ahead (lag= 10).

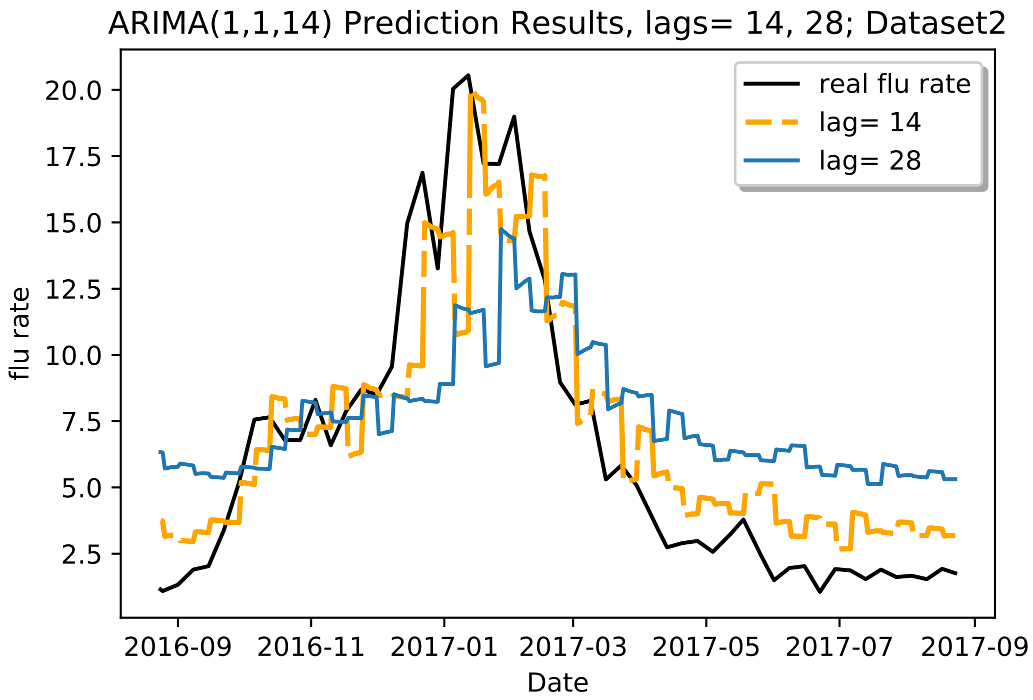

Figure 4.10: ARIMA(1,1,14) forecasting based on Dataset2

Figure 4.10 shows the flu rate forecasting result of 14 days and 28 days future by ARIMA (1,1,14) model on Dataset2. As in Figure 4.10, the black line is the real flu rate, the orange dashed line is the prediction made 5 days ahead (lag= 14), and the the solid blue line shows the forecasting of 10 days ahead (lag= 28).

From Figure 4.7, 4.8, 4.9, 4.10 we can see that the closer future we forecast the more accurate our prediction is. Since lag equals one gives forecasting result nearly identical to real flu rate (Table 4.3 presents very small error for lag= 1), the plot for lag equals one is not presented.

4.2.5 ARIMAX models

4.2.5.1 Using the first query (the most correlated search query with flu activity) as exogenous variables

	Test Set1			Test Set 2			
ARIMAX (1,2,15)	MAE	RMSE	CORR	ARIMAX (2,1,15)	MAE	RMSE	r
lag=1	0.054	0.155	0.9998	lag=1	0.062	0.133	0.9997
lag=5	0.652	1.114	0.9883	lag=5	0.500	0.865	0.9878
lag=10	1.856	2.589	0.9438	lag=10	1.174	1.854	0.9432
lag=14	2.341	3.273	0.9155	lag=14	1.609	2.432	0.9005
lag=28	3.968	5.600	0.7864	lag=28	2.851	3.786	0.7470

Table 4.4: Result Table of ARIMAX with the first query as exogenous variable

Table 4.4 shows the forecasting error of ARIMAX(1,2,15) on dataset1 and ARIMAX(2,1,15) on dataset2 with the query which is most correlated to flu activity as input exogenous variable. lag means how many days ahead our prediction is made, for example, lag= 5 means we use the information till now to forecast the flu rate 5 days later.

From Table 4.4, the ARIMAX model with first query as exogenous variable gets a generally better result for both datasets than ARIMA models, for example, dataset2, lag= 28, ARIMA(1,1,14) has MAE= 3.512, RMSE= 4.081 and here ARIMAX (1,2,15) has MAE= 2.851 and RMSE= 3.786 for lag= 28 dataset2 (see Table 4.3).

4.2.5.2 Using first 3 queries (the 3 most correlated search queries with flu activity) as exogenous variables

	Test Set1			Test Set 2			
ARIMAX (1,2,15)	MAE	RMSE	CORR	ARIMAX (2,1,14)	MAE	RMSE	r
lag=1	0.059	0.154	0.9997	lag=1	0.061	0.131	0.9997
lag=5	0.652	1.126	0.9883	lag=5	0.488	0.859	0.9880
lag=10	1.830	2.540	0.9424	lag=10	1.158	1.844	0.9449
lag=14	2.266	3.181	0.9116	lag=14	1.582	2.401	0.9055
lag=28	3.612	5.318	0.7640	lag=28	2.661	3.649	0.7744

Table 4.5: Result Table of ARIMAX with first 3 queries as exogenous variable

Table 4.5 shows the forecasting error of ARIMAX (1,2,15) on dataset1 and ARIMAX (2,1,14) on dataset2 with the 3 queries which are most correlated with flu activity as input exogenous variable.

From Table 4.5, ARIMAX model with first 3 search queries (the 3 queries that are most correlated to flu activity) as exogenous variable, the model gets better result for both datasets than ARIMAX model with the first query which is most correlated to flu activity as exogenous variable (see Table 4.4).

4.2.5.3 Using first 10 queries (the 10 most correlated search queries with flu activity) as exogenous variables

	Test Set1			Test Set 2			
ARIMAX (1,2,14)	MAE	RMSE	CORR	ARIMAX (2,1,14)	MAE	RMSE	r
lag=1	0.056	0.157	0.9998	lag=1	0.067	0.133	0.9997
lag=5	0.658	1.140	0.9883	lag=5	0.503	0.862	0.9878
lag=10	1.859	2.590	0.9440	lag=10	1.182	1.858	0.9426
lag=14	2.351	3.279	0.9158	lag=14	1.619	2.436	0.8993
lag=28	3.987	5.618	0.7882	lag=28	2.894	3.785	0.7417

Table 4.6: Result Table of ARIMAX with first 10 queries as exogenous variable

Table 4.6 shows the forecasting error of ARIMAX (1,2,14) on dataset1 and ARIMAX (2,1,14) on dataset2 with the 10 queries which are most correlated to flu activity as exogenous variables.

From Table 4.6, ARIMAX model with first 10 search queries (the 3 queries that are most correlated to flu activity) as exogenous variable, the model gets worse result for both datasets than ARIMAX model with the first 3 queries as exogenous variable (see Table 4.5). As more input variables adding to the ARIMAX model, the weight for each variable gets smaller and smaller. Therefore we need to select useful informations as input.

4.2.5.4 Using Historical Average Flu Rate as Exogenous Variable in ARIMAX

The historical average is the estimation results obtained in Section 4.2.3. For each history period of 365 days and 733 days, calculate the average flu rate of each day in the period and then use the result as input exogenous variable in ARIMAX model.

		Test Set1						
ARIMAX (2,1,14)	Period	MAE	RMSE	CORR	Period	MAE	RMSE	r
lag=1	365	0.058	0.144	0.9998	733	0.057	0.143	0.9998
lag=5	365	0.686	1.072	0.9892	733	0.685	1.070	0.9893
lag=10	365	1.787	2.442	0.9446	733	1.799	2.446	0.9443
lag=14	365	2.357	3.162	0.9132	733	2.377	3.174	0.9118
lag=28	365	4.240	5.235	0.7598	733	4.244	5.236	0.7558

Table 4.7: Result Table of ARIMAX with periodical average as exogenous variable, test set1

Table 4.7 shows the forecasting error of ARIMAX (2,1,14) on dataset1 with 365 days historical average ILI rate and 733 days historical average ILI rate as exogenous variables.

		Test Set2						
ARIMAX (2,1,14)	Period	MAE	RMSE	CORR	Period	MAE	RMSE	r
lag=1	365	0.042	0.108	0.9998	733	0.042	0.108	0.9998
lag=5	365	0.501	0.808	0.9898	733	0.504	0.804	0.9899
lag=10	365	1.360	1.874	0.9465	733	1.384	1.856	0.9479
lag=14	365	1.948	2.503	0.9107	733	1.990	2.494	0.9125
lag=28	365	3.460	3.932	0.8347	733	3.511	3.951	0.8348

Table 4.8: Result Table of ARIMAX with periodical average as exogenous variable, test set2

Table 4.8 shows the forecasting error of ARIMAX (2,1,14) on dataset2 with 365 days historical average ILI rate and 733 days historical average ILI rate as exogenous variables.

From Table 4.7 and Table 4.8, the error of forecasting only improves very little from ARIMA model, for example, for lag= 28 dataset2, ARIMA (1,1,14) has MAE=

3.521, RMSE= 4.081 and here we ARIMAX (2,1,14) for lag= 28 dataset2 we have MAE= 3.511, RMSE= 3.951. (see Table 4.3). Therefore, using the history periodical average flu rate as exogenous variable in ARIMAX model is not a good choice for forecasting daily flu rate.

4.2.5.5 ARIMAX using the result of Elastic Net estimation as exogenous variable

	Test Set1			Test Set 2			
ARIMAX (2,1,15)	MAE	RMSE	CORR	ARIMAX (2,1,15)	MAE	RMSE	r
lag=1	0.043	0.137	0.9998	lag=1	0.033	0.107	0.9998
lag=5	0.587	1.022	0.9902	lag=5	0.417	0.796	0.9899
lag=10	1.646	2.328	0.9501	lag=10	1.193	1.851	0.9469
lag=14	2.084	2.967	0.9197	lag=14	1.637	2.425	0.9099
lag=28	3.390	5.028	0.7716	lag=28	2.372	3.440	0.8213

Table 4.9: Result Table of ARIMAX using Elastic Net estimation results as the Exogenous Variable

Table 4.9 shows the forecasting error of ARIMAX (2,1,15) with the estimation result from elastic net models as exogenous variable on dataset1 and dataset2. The value with blue means that the error is the lowest among all the auto-regression models.

From Table 4.9, ARIMAX with elastic net estimation as exogenous variable input gives the best prediction for both datasets among all the experiments of autoregressive models in previous sections.

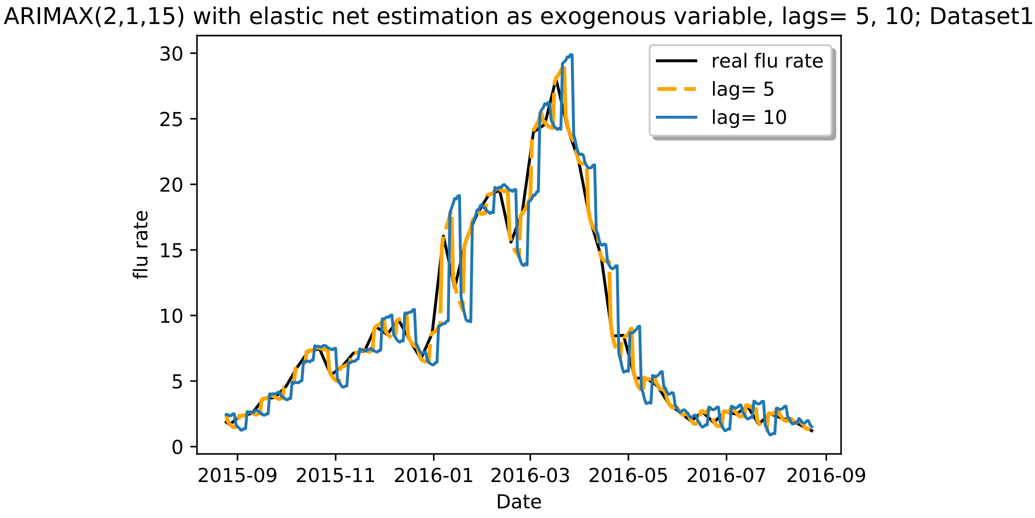

Figure 4.11: ARIMAX (2,1,15) with the estimation from elastic net regression as exogenous variable on dataset1, forecasting 5 and 10 days length in future

Figure 4.11 shows the flu rate forecasting result of 5 days and 10 days in the future by ARIMAX (2,1,15) model on dataset1 with the estimation result from elastic net regression as exogenous variable. As in Figure 4.11, the black line is the real flu rate, the orange dashed line is the prediction of 5 days future flu rate (lag= 5), and the the solid blue line shows the forecasting of 10 day length future (lag= 10).

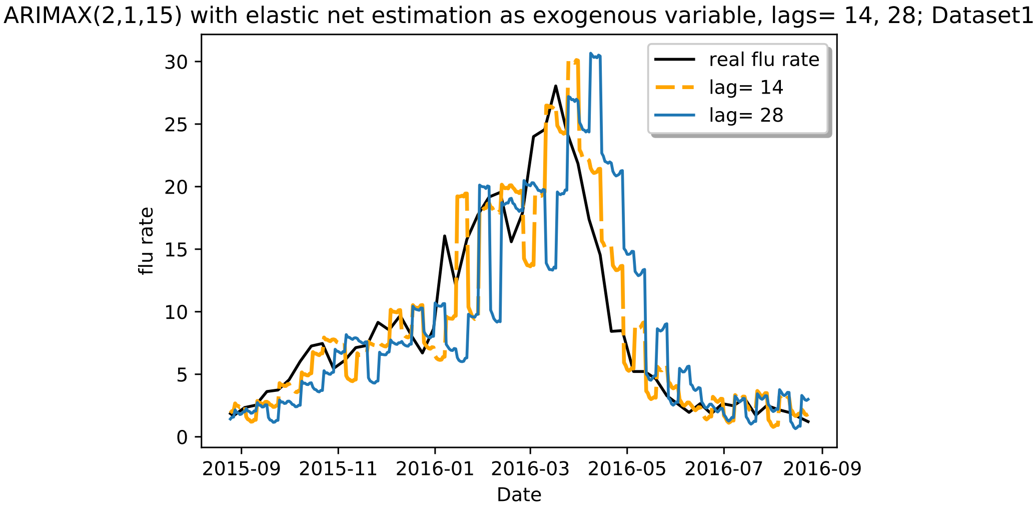

Figure 4.12: ARIMAX (2,1,15) with the estimation from elastic net regression as exogenous variable on dataset1, forecasting 14 and 28 days length in future

Figure 4.12 shows the flu rate forecasting result of 14 days and 28 days in the future by ARIMAX (2,1,15) model on dataset1 with the estimation result from elastic net regression as exogenous variable. As in Figure 4.12, the black line is the real flu rate, the orange dashed line is the prediction of 14 days future flu rate (lag= 14), and the the solid blue line shows the forecasting of 28 day length future (lag= 28).

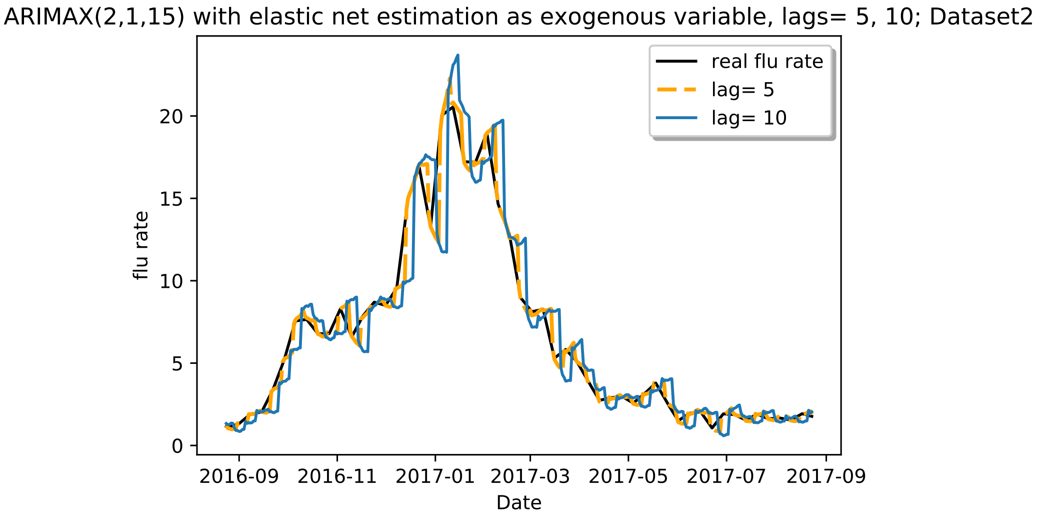

Figure 4.13: ARIMAX (2,1,15) with the estimation from elastic net regression as exogenous variable on dataset2, forecasting 5 and 10 days length in future

Figure 4.13 shows the flu rate forecasting result of 5 days and 10 days in the future by ARIMAX (2,1,15) model on dataset2 with the estimation result from elastic net regression as exogenous variable. As in Figure 4.13, the black line is the real flu rate, the orange dashed line is the prediction of 5 days future flu rate (lag= 5), and the the solid blue line shows the forecasting of 10 day length future (lag= 10).

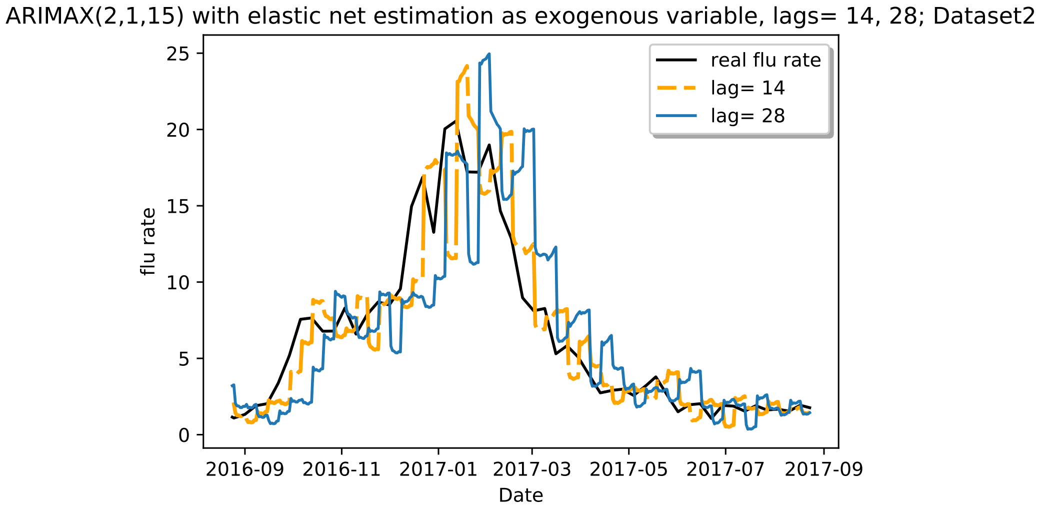

Figure 4.14: ARIMAX (2,1,15) with the estimation from elastic net regression as exogenous variable on dataset2, forecasting 14 and 28 days length in future

Figure 4.14 shows the flu rate forecasting result of 5 days and 10 days in the future by ARIMAX (2,1,15) model on dataset2 with the estimation result from elastic net regression as exogenous variable. As in Figure 4.14, the black line is the real flu rate, the orange dashed line is the prediction of 5 days future flu rate (lag= 5), and the the solid blue line shows the forecasting of 10 day length future (lag= 10). By comparing Figure 4.12 to Figure 4.8 and Figure 4.14 to Figure 4.10 , we can clearly see the ARIMAX (2,1,15) with the estimation results from elastic net regression gives a better forecasting than ARIMA (1,1,14).

4.2.6 SARIMA and SARIMAX Models

	Test Set1			Test Set 2			
SARIMA (1,1,14) (1,0,1,7)	MAE	RMSE	CORR	SARIMA (0,2,14) (1,0,1,7)	MAE	RMSE	r
lag=1	0.047	0.150	0.9998	lag=1	0.037	0.110	0.9998
lag=5	0.630	1.118	0.9882	lag=5	0.469	0.820	0.9895
lag=10	1.850	2.578	0.9369	lag=10	1.407	1.932	0.9431
lag=14	2.456	3.366	0.8971	lag=14	2.044	2.623	0.9008
lag=28	4.423	5.660	0.6835	lag=28	3.601	4.164	0.7932

Table 4.10: Result Table of SARIMA Models

Table 4.10 shows the forecasting error of SARIMA(1,1,14)(1,0,1,7) for dataset1 and SARIMA (0,2,14)(1,0,1,7) for dataset2. The SARIMA model with lowest AIC includes a seasonal term of period 7. This 7 days period does not have strong meaning, because flu is definitely not weekly happen. It is not surprise to observe that the SARIMA model does not improve from ARIMA model (Table 4.3) by adding a seasonal term of 7.

	Test Set1			Test Set 2			
SARIMAX (1,2,14) (1,0,0,7)	MAE	RMSE	r	SARIMAX (1,2,14) (0,0,1,12)	MAE	RMSE	CORR
lag=1	0.055	0.143	0.9998	lag=1	0.039	0.114	0.9998
lag=5	0.605	1.045	0.9897	lag=5	0.428	0.834	0.9891
lag=10	1.692	2.373	0.9476	lag=10	1.234	1.924	0.9442
lag=14	2.176	3.033	0.9146	lag=14	1.734	2.539	0.9050
lag=28	3.641	5.284	0.7435	lag=28	2.721	3.845	0.7962

Table 4.11: Result Table of ARIMAX with first 3 queries as exogenous variable

Table 4.11 shows the forecasting error of SARIMAX (1,2,14)(1,0,0,7) on dataset1 and SARIMAX (1,2,14)(0,0,1,12) on dataset2 with the three queries which are most correlated to flu activity as the input exogenous variables.

From the Table 4.11, including a seasonal period of 7 for dataset1 and a seasonal period of 12 for dataset2 improve the forecasting performance from ARIMA model (Table 4.3), for example, dataset1, lag= 28, ARIMA(1,1,14) has MAE= 4.396, RMSE= 5.574 and here SARIMAX (1,2,14)(1,0,0,7) has MAE= 3.641 and RMSE= 5.284. But a seasonal period of 7 or 12 is impractical for a flu season.

SARIMA and SARIMAX models are designed for short period, for example monthly (period = 12) or quarterly (period = 4). Here our object is to forecast ILI rate which should be a long period (discussed in Section 4.2.1 Analysis of Autocorrelation). Therefore, SARIMA and SARIMAX models are not our focus in this these.

Chapter 5

Conclusion and Discussion

This chapter summarize the results from experiments and state the limitation of the experiments and how to improve.

5.1 Summary

In this work, we first use the linear regression models (elastic net regression, ridge regression and least square regression) with the search query frequencies to estimate the flu rate. This is not forecasting because we know all the search query information of all time. Then we focusing on forecasting flu rate with auto-regressive models (ARIMA, ARIMAX, SARIMA and SARIMA) and find ARIMAX is the most useful model among them. For ARIMAX models we have tried several different exogenous variable inputs and finally find out the ARIMAX model with the estimation results from elastic net regression as exogenous variable input gives the best forecasting result of daily flu rate.

5.2 Drawbacks

For elastic net regression models, we did not find out the optimized hyper-parameters since we fixed L1 ratio α. For ARIMAX model with the estimation results from elastic net regression as exogenous variable. We did not use the optimized estimation results from elastic net regression because using the fixing parameters optimized in one elastic net model for all the elastic net estimations. For ARIMAX models, did not conduct experiments with exogenous variable combinations, for example the first query which is most correlated to flu activity and the estimation results from elastic net regression together as the exogenous variable of the ARIMAX model. SARIMA and SARIMAX models can not handle large periods and small seasonal period is not realistic for ILI season.

5.3 Future Work

Firstly, more experiments with ARIMAX model can be designed. The combinations of different variables (estimation from elastic net regression, query most correlated to flu activity, periodical average flu rate and many others) together as the input exogenous variable of ARIMAX model is valuable to be compared.

Secondly, many other on-line user generated information besides Google search query can be used as input data. In [21], the author provides a new Internet resource (www.uptodate.com) of data which is used by 700,000 clinicians in 158 countries and almost 90% of academic medical centers in the United States. The author compared the multi-variable linear model on Google flu trends (GFT) data and the ”uptodate” website data with the Centers for Disease Control and Prevention (CDC) reported influenza-like illness (ILI) activity data. The author concluded that the ”uptodate” website data fitted better to the reported data and is a good source for ILI research. In addition the Twitter post is good choice as well.

Thirdly, we can use neural network models to forecast flu rate. Recurrent Neural Net (RNN) models are proven good for time series [22]. It is valuable to compare the performance of three basic RNN models (simple RNN, LSTM, GRU) with auto regression models.

Appendix A

Tools and Packages

Google Health Trend API to obtain Google search engine query information.

Anaconda with Jupyter notebook

Python 3 packages:

Numpy, Pandas, Statsmoldel,Keras

The python code can be found in `https://github.com/yangzhouy/Influenza-Like-Illness-ILI-Forecasting-Based-on-Historical-Flu-Rate` All python code are written in Jupyer notebook. In the Jupyter note book files, dataset1 is the dataset2 in this report and dataset2 is the dataset1 in this report. If there are many files with copy number, the largest copy number will be one that been used.

Bibliography

[1] World Health Organization. Up to 650 000 people die of respiratory diseases linked to seasonal flu each year. http://www.who.int/news-room/detail/14-12-2017-up-to-650-000-people-die-of-respiratory-diseases-linked-to-seasonal-flu-each-year, 2017.

[2] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437(7056):209, 2005.

[3] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. Digital disease detection-harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009.

[4] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012, 2009.

[5] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[8] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[9] Richard McCleary, Richard A Hay, Erroll E Meidinger, and David McDowall. *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA, 1980.

[10] Billy Williams. Multivariate vehicular traffic flow prediction: evaluation of arimax modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (1776):194–200, 2001.

[11] Mario Cools, Elke Moons, and Geert Wets. Investigating the variability in daily traffic counts through use of arimax and sarimax models: assessing the effect of holidays on two site locations. *Transportation Research Record: Journal of the Transportation Research Board*, (2136):57–66, 2009.

[12] Michael J Kane, Natalie Price, Matthew Scotch, and Peter Rabinowitz. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, 15(1):276, 2014.

[13] Vasileios Lampos, Bin Zou, and Ingemar Johansson Cox. Enhancing feature selection using word embeddings: The case of flu surveillance. In *Proceedings of the 26th International Conference on World Wide Web*, pages 695–704. International World Wide Web Conferences Steering Committee, 2017.

[14] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[15] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[16] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[17] Hirotugu Akaike. Factor analysis and aic. In *Selected Papers of Hirotugu Akaike*, pages 371–386. Springer, 1987.

[18] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[19] William WS Wei. Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*. 2006.

[20] Niels Dalum Hansen, Kåre Mølbak, Ingemar J. Cox, and Christina Lioma. Seasonal web search query selection for influenza-like illness (ili) estimation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1197–1200, New York, NY, USA, 2017. ACM.

[21] Mauricio Santillana, Elaine O Nsoesie, Sumiko R Mekaru, David Scales, and John S Brownstein. Using clinicians search query data to monitor influenza epidemics. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 59(10):1446, 2014.

[22] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.