博士論文

Study on Predicting Influenza Outbreaks

(インフルエンザにおける流行予測に関する研究)

張捷

Acknowledgments

I am grateful to Prof. Kazumitsu Nawata Dr., my supervisor. His profound knowledge triggered my curiosity for the academic world and his earnest professional shaped me the attitudes towards scientific research.

Besides, I appreciate the constructive instruction and kind help from Prof. Ichiro Sakata Dr., Prof. Nariaki Nishino Dr., Prof. Kenji Tanaka Dr., and Prof. Yoshiko Mizuno Dr.

Furthermore, thank my family: my parents, wife, daughter, and son. Their selfless support encouraged me to complete the Ph.D. study.

Abstract

Every year the worldwide influenza infection places a substantial burden on people's health. Most past studies just focused on a ”single-step” prediction of ”regional” influenza outbreaks, although the infection presents a strong geolocational-temporal correlation. Therefore, we highlight the necessity of developing a geolocational-temporal predictive system to perform a multistep prediction of an annual worldwide influenza outbreak. To achieve this goal, we divided our research into three steps:

1. To find the best model type and the best time lag: we performed a series of experiments and compared six different types of models. We found the Long Short Term Memory (LSTM) with a time lag of 52 weeks achieved the best predictive accuracy.

2. To find the best multistep prediction algorithm: we compared four different multistep predictive algorithms. We implemented these four algorithms in the LSTM with the time lag of 52 weeks and predicted 2-, 3-,..., 13-week-ahead influenza data. We found the algorithm of Multiple Single-Output Prediction (MSOP) achieved the best accuracy.

3. To develop the best system of a geolocational-temporal prediction of influenza: we collected the influenza data of the 152 countries/regions from the FluNet, a database of the World Health Organization. Then we selected 22 countries, the influenza data of which had no ”N/A”s (not available), as features; and select 6 countries, the populations of which were relatively large, as predictive targets. We implemented MSOP in LSTM with a time lag of 52 weeks and predicted 1-, 2-, 3- 4-week-ahead influenza data.

The main results were (a) in the Southern Hemisphere (i.e. Australia and Brazil in this study), the 1-, 2-, 3-, and 4-week-ahead predictive Mean Absolute Percentile Error (MAPE) with feeding the influenza data of other countries into models were higher than those without feeding the influenza data of other countries; (b) in the Northern Hemisphere (China, Japan, United Kingdom, and United States of America in this study), the 2-, 3-, and 4-week-ahead predictive MAPEs with other countries were lower than those without other countries; and (c) in the Northern Hemisphere, the 1-week-ahead predictive MAPEs with other countries were usually higher than the MAPEs without feeding other countries, except for United Kingdom.

Then we conducted further experiments and look for an explanation. We found that since the 22 countries were mostly in the Northern Hemisphere. Feeding the historical influenza data of the 22 countries into the predictive models helped forecast influenza data in the Northern Hemisphere due to the high correlation among the influenza data in the Northern Hemisphere; but exacerbate the predictive accuracy in the Southern Hemisphere since influenza seasons in the Southern Hemisphere usually peaked in June, July, and August, totally unrelated to those in the Northern Hemisphere. Furthermore, the reason that the 1-week-ahead predictive MAPEs without feeding influenza data of other countries were better than those with feeding influenza data of other countries in the Northern Hemisphere was that the spread of influenza among countries need some time and thereby has a time lag. In addition, as for the United Kingdom, the rapidly increasing number of travelers in 2017 and 2018 (around two-thirds of the population of the United Kingdom) disrupted the time lag of flu spread. Thereby, we concluded feeding relevant geolocational-temporal factors in the same hemisphere into a forecast model improves the accuracy of the multistep prediction of the worldwide influenza outbreaks.

Contents

Chapter 1

Introduction

This section gives the introduction of the whole Ph.D. thesis. We divide this section into two parts. In the first half, we explain influenza and the relevant background. In the second half, we illustrate our research approaches.

1.1 Influenza (Flu)

Influenza, or shortly flu, is an infectious disease caused by flu viruses, which infect respiratory system [1, 2]. Flu symptoms begin in a period after one gets an infection of the flu virus. The period can last 2 - 7 days [1]. In the early infection, people can hardly distinguish between the common cold and flu [3]. Many people feel fever (38 - 39 °C), aches and pains of their bodies, backs and legs [4, 5]. Besides, coughing, sneezing, sore throat, headache, and feeling tired also occur [1]. Occasionally, flu causes severe complications [6], which includes viral pneumonia, secondary bacterial pneumonia, sinus infections, etc [2, 7, 8].

1.1.1 Flu Virus

We have already known four types of the flu virus, called Influenza Virus A (IVA), Influenza Virus B (IVB), Influenza Virus C (IVC), and Influenza Virus D (IVD) [2]. Among them, IVA, IVB, and IVC can infect humans [9]. IVA is the most popular flu virus and causes the severest disease. Based on the antibodies that respond to these viruses, we can divide IVA into several serotypes [10]. Among them, H1N1 caused the Spanish Flu in 1918, and Swine Flu in 2009; H2N2 caused Asian Flu in 1957. H3N2 caused Hong Kong Flu in 1968, and H5N1 caused Bird Flu in 2004. IVB is much milder than IVA since IVB mutates much slower than IVA [11]. IVB has only one serotype [10] and only infects humans [10]. Thanks to IVB's limited host range and reduced rate of mutation, pandemics of IVB may never happen [12]. IVC is the least common type of flu virus. IVC has also only one species and only causes mild disease in children [13, 14], dogs and pigs [15, 16]. IVD is the newest flu virus. IVD was identified in 2016 and has the potential to infect people but currently not [17–25].

1.1.2 Flu Transmission

During one person's infection period, he or she is infectious to others [7, 8]. Smoking increases the probability of flu infection and causes more severe symptoms after infection [26, 27]. The duration in which a person might be infectious to another person is called flu virus shedding. In the 2nd day after infection, the flu virus shedding reaches peak [25]. The average duration of flu virus shedding is 5 days [25], and the max duration of flu virus shedding is 9 days [25]. Other animals, such as pigs, horses, and birds, can also be infected [28]. The flu virus spreads in a relatively short distance [29], and can be transmitted in three ways [30, 31]:

(1) direct transmission: direct transmission happens when one person's eyes, nose or mouth is contaminated by an infected person's sneezes.

(2) airborne transmission: an infected person spreads more than half a million virus particles when he or she sneezes or coughs [32]. A single sneeze contains around 40,000 droplets [33]. Droplets from infected people are very small. People can inhale droplets from 0.5 μm to 5 μm in diameter, and inhaling even one droplet may cause flu infection [30].

(3) hand-to-eye, hand-to-nose, or hand-to-mouth transmission, either from contaminated surfaces or from direct personal contact such as a handshake. The flu virus can persist outside of the body. Flu can survive on plastic or metal for 1 - 2 days, on dry paper tissues for around 15 minutes, on the skin for only 5 minutes [34]. Contaminated surfaces such as paper [35] and household items [4] helps transmit flu virus. The flu virus may indirectly fly into mouths or eyes and cause flu infection [7, 8, 29].

1.1.3 Flu Season

Due to the strong infection, flu becomes the most popular infectious disease around the world. Flu infection has a close relation to humidity and temperature. In temperate zones, as temperature decreases or humidity increases, the number of flu instances will increase, and in temperate zones, flu usually has an annual outbreak. In the Northern Hemisphere, the flu season usually takes place from October to May and peaks in February. In Southern Hemisphere, the flu season usually occurs from May to October and peaks in August. In the tropics and subtropics, as temperature increases or humidity increases, the number of flu instances will increase., and in the tropics and subtropics, flu reason lasts whole years [1]. Larger outbreaks (pandemics) are less frequent. Only three flu pandemics occurred, Spanish flu in 1918 (around 50 million deaths), Asian flu in 1957 (two million deaths), and Hong Kong flu in 1968 (one million deaths) [36]. The latest pandemic was in 2009, caused by a new type of H1N1 [37].

1.1.4 Flu Burden

The flu outbreaks place a substantial burden on human beings. Around the world, flu caused approximately 3 to 5 million annual cases of severe illness and 250,000 to 500,000 deaths in 2016 [1]. Clinics and hospitals are overwhelmed during peak illness periods. Flu is one of the costliest epidemics worldwide. Flu has direct and indirect costs. Direct cost is the expense of lost productivity and associated medical treatment. Indirect cost is the spending on preventative measures. Generally, around the world, a 3% sickness rate and a three-week length of illness would decrease gross domestic product by 5%. Additional costs would come from medical treatment of 18 million to 45 million people, and total economic costs would be approximately 700 billion United States (US) dollar [38].

The following contents were the collected direct and indirect cost from flu outbreaks in different countries. In Australia, in 2010, flu was estimated to result in a direct medical cost of over (US)\$ 96 million [39, 40]. In France, in 2010, flu was estimated to result in an economic cost of over (US)\$

3 billion, in which direct medical costs were (US)$ 292 million annually and indirect costs were (US)$3.35 million [39, 41]. In Norway, in 2010, flu was estimated to result in an economic cost of over (US)$ 196 million, in which direct medical costs were (US)$ 19 million annually and indirect costs were (US)$ 215 million [39, 42]. In Spain, in 2010, flu was estimated to result in an economic cost of over (US)$ 1.5 billion, in which direct medical costs were (US)$ 550 million annually and indirect costs were (US)$ 986 million [39, 43]. In Germany, in 2010, flu was estimated to result in an economic cost of over (US)$ 3.9 billion, in which direct medical costs were (US)$ 3.4 billion annually and indirect costs were (US)$ 467 million [39, 44]. In Japan, in 2010, flu was estimated to result in a direct medical cost of over (US)$ 5.6 million [39, 45]. In Hong Kong, in 2010, flu was estimated to result in an economic cost of over (US)$ 24 million, in which direct medical costs were (US)$ 2 million annually and indirect costs were (US)$ 21 million [39, 46]. In Thailand, in 2010, flu was estimated to result in an economic cost of over (US)$ 47 million, in which direct medical costs were (US)$ 26 million annually and indirect costs were (US)$ 21 million [39, 47]. In the US, in 2010, flu was estimated to result in an economic cost of over (US)$ 20 billion, in which direct medical costs were (US)$ 7 billion annually and indirect costs were (US)$ 12 million [39, 48]. In the US, in 2018, flu was estimated to result in an economic cost of over (US)$ 11 billion in average yearly [49], in which, direct medical costs are over (US)$ 3 billion annually and indirect costs were (US)$ 8 billion [49].

1.1.5 Flu Prevention and Treatment

Personal hygiene habits, such as not touching your eyes, nose or mouth [50]; hand washing [51, 52]; covering coughs and sneezes; avoiding close contact with flu patients; wearing face masks [53, 54]; avoiding spitting [55]; and staying home when sick obviously helps people to prevent flu infection during the flu seasons. Besides, flu vaccine is a cost-effectiveness way to prevent flu [8, 56–58]. flu vaccine has been widely evaluated for different groups [59], such as in children [60], and the elderly [61]. World Health Organization and the Centers for Disease Control and Prevention recommended the flu vaccine for high-risk groups, such as children, the elderly, health care workers, and people who had chronic illnesses such as asthma, diabetes, heart disease, or were immuno-compromised among others [62]. The flu vaccine takes about two weeks to become effective [63]. Therefore, it is also possible to get infected just before vaccination and get sick with the strain that the vaccine is supposed to prevent. The results of economic evaluations of flu vaccination have often been found to be dependent on key assumptions [64, 65]. During the 2015-2016 flu season, flu vaccine prevented an estimated 5.1 million illnesses, 2.5 million medical visits, 71000 hospitalizations, and 3000 pneumonia and flu (P&I) deaths [7]. Manufacturing of flu vaccine is a challenging job because flu virus undergoes high mutation rates and frequent genetic re-assortment (combination and rearrangement of genetic material) [66–70]. Therefore, every year, flu vaccines production suffers from a complicated procedure. In Februaries, World Health Organization assesses the strains of flu virus that are most likely to be circulating over the following winter. Then, vaccine manufacturers can only produce flu vaccines in a very limited time [71]. As a result, the first batch of vaccine is usually unavailable for the patients until every September [71]. During flu peak periods, clinics and hospitals are overwhelmed. Beds assignment to flu patients in hospitals is a challenging task due to the limited capacity of hospital beds, time-dependencies of bed request arrivals, and unique treatment requirements of flu patients [72]. Besides, flu seasons vary in timing, severity, and duration from one season to another [71]. Therefore, flu hospitalization also varies by sites and time in each season [73], which makes beds assignment to flu patient more difficult for hospitals.

1.2 Flu Prediction

In this part, we discuss the prediction for the flu. Firstly, we explain the goal of flu prediction. Secondly, we summarize past studies on geolocational-temporal flu prediction.

1.2.1 Geolocational-Temporal Multistep Prediction for Flu

We need a geolocational-temporal multistep prediction of flu outbreaks. For one thing, a multistep prediction of flu outbreaks helps prepare for flu outbreaks on time. As we aforementioned, vaccine manufacturing and dynamic hospitalization need a buffer duration. For another, a geolocational-temporal prediction helps improve predictive accuracy, when one considers the correlation of flu outbreaks among different countries.

(a) Multistep Prediction

A multistep-ahead time series prediction, or a multistep prediction, is an analytical task of predicting a sequence of values in the future by analyzing observed values in the past [74]. By the multistep prediction of flu, we can understand flu spreading trend several weeks ahead and thereby dynamically plan flu vaccine manufacturing and hospital bed assignments. Nonetheless, few past studies have focused on the multistep prediction of flu outbreaks. The probable reason is that a multistep prediction usually results in poor accuracy due to some insuperable problems, such as error accumulation [75, 76]. One compromising solution is that one can aggregate raw data to a larger time unit and then use a single-step prediction to replace a multistep prediction. For example, if raw data is weekly based, we can aggregate weekly values to monthly values and then perform a single-step prediction of the total value of the coming month (roughly around four weeks). The demerit is that the aggregation hinders us from understanding the internal variation during the coming four weeks.

(b) Geolocational-Temporal Prediction

Second, we also need a geolocational prediction instead of a regional prediction. A geolocational prediction leverages flu data from countries all over the world. Since flu is an infectious disease, a model with inputs of flu data all of the countries around the world helps understand the developing trends of flu and improve predictive accuracy. Nevertheless, the past studies focused on regional flu outbreaks prediction [77–80]. From our perspective of views, there are two probable reasons. Firstly, the flu virus shows sensitivity to temperature and humidity, and different locations of one country or one region, to some extent, share similar geolocational characteristics, such as humidity and temperature. As a result, predicting flu infection of one country or one region is considered reasonable and approachable. Secondly, flu virus transmission is believed to occur mostly over relatively short distances. Usually, the flu virus is spread through the air from coughs or sneezes. When an infected person coughs or sneezes, droplets containing viruses (infectious droplets) are dispersed into the air and can spread up to one meter, and infect persons near who breathe these droplets in.

However, one fast-growing risk group, travelers, is neglected from these two overviews above. Several changes resulting from our globalizing world contribute to the growing influence of the traveler group:

(i) the steady increase in total travel volume worldwide,

(ii) the advent of mass tourism, and

(iii) increasing numbers of immune-compromised and elderly travelers.

International sporting events and festivals (such as the 2018 Russian World Cup) as well as traveling by airplane or cruise ship could facilitate flu virus transmission and therefore causes the geolocational-temporal spread of flu [81]. Another previous study shows that flu outbreaks correlate with each other in all the countries around the world [82]. Therefore, we suppose a geolocational-temporal prediction improve predictive accuracy.

1.2.2 Past Studies on Flu Prediction

Table 1.1 summarized the past studies on flu prediction in the past five years. Michiels et al. [83] performed a single step prediction of flu outbreaks in Belgium. Wu et al. [80], Bu et al. [84], Guo et al. [85, 86], Liang et al. [87], and Wang et al. [88], performed a single step prediction of flu outbreaks in China. Chaudhary et al. [89] performed a single step prediction of flu outbreaks in India. Seleznev et al. [90] performed a single step prediction of flu outbreaks in Russia. Fu et al. [91], Tung et al. [92], and Chen et al. [93], performed a single step prediction of flu outbreaks in Taiwan. Murray et al. [94], performed a single step prediction of flu outbreaks in Scotland. Spreco et al. [95, 96], performed a single step prediction of flu outbreaks in Sweden. Alkouz et al. [97] performed a single step prediction of flu outbreaks in the United Arab Emirates. Corbella et al. Alessa et al. [98], Lee et al. [99], Bardak et al. [100], Verma et al. [101], Xue et al. [102], Du et al. [103,104], Belkhiria et al. [105], Lu et al. [106], Paul et al. [107], and Morita et al. [108] performed a single step prediction of flu outbreaks in United States. Thrastarson et al. [109] performed a single step prediction of flu outbreaks in the United States and South Africa. The contents above simply summarized the previous researches on flu prediction.

The past studies on flu prediction have some common characteristics. First, most studies usually performed a single step prediction of flu data in one country or one region. Second, few past studies performed multistep predictions, let alone a complete exploration of algorithms of multistep prediction. Third, few studies performed translocation flu prediction, as Table 1.1 shows. In conclusion, past studies of predicting flu outbreaks have some common drawbacks:

(1) Past studies usually conducted single step predictions. However, hospitals and vaccine manufacturers, and so on. need to prepare for flu outbreaks on time.

(2) Past studies usually aimed at one country or one region. Nevertheless, flu spreads worldwide. A predictive model with inputs of flu data all over the world helps improve accuracy.

(3) Past studies usually chose models without hyperparameter search. Adopting advanced models with hyperparameter selection is supposed to improve accuracy.

Table 1.2 shows the predictive accuracy of the previous studies. Since almost all papers had different research objectives, here, we only present the best accuracy of all the models in the previous studies respectively. The reasons for ”NA” includes six reasons:

(a) the papers presented the accuracy in other metrics instead of mean square error (MSE) or

Table 1.1: Past studies on flu prediction.

Author	Country or Region	Year	Type	Range
Michiels et al.	Belgium	2017	single step	one country
Wu et al.	China	2017	single step	one country
Bu et al.	China	2018	single step	one country
Guo et al.	China	2017	single step	one country
Guo et al.	China	2017	single step	one country
Liang et al.	China	2018	single step	one country
Wang et al.	China	2017	single step	one country
Chaudhary et al.	India	2017	single step	one country
Seleznev et al.	Russia	2018	single step	one country
Fu et al.	Taiwan	2017	single step	one region
Tung et al.	Taiwan	2015	single step	one region
Murray et al.	Scotland	2018	single step	one region
Spreco et al.	Sweden	2017	single step	one country
Spreco et al.	Sweden	2017	single step	one country
Alkouz et al.	United Arab Emirates	2018	single step	one country
Alessa et al.	United States	2018	single step	one country
Lee et al.	United States	2017	single step	one country
Bardak et al.	United States	2017	single step	one country
Verma et al.	United States	2017	single step	one country
Xue et al.	United States	2018	four-step	one country, several regions
Du et al.	United States	2017	single step	one country
Du et al.	United States	2018	single step	one country
Lu et al.	United States	2018	single step	one country
Belkhiria et al.	United States	2018	single step	one country
Paul et al.	United States	2017	single step	one country
Morita et al.	United States	2018	single step	one country
Thrastarson et al.	United States & South Africa	2017	single step	two countries, separately

The column of ”Author” describes the authors of past studies. The column of ”Country or Region” describes the location of the past studies. The column of ”Year” describes the publishing years. The column of ”Type” describes the single-step prediction or multistep prediction. The column of ”Range” describes the range of studies, such as one country, one region, and so on.

mean absolute percentage error (MAPE);

(b) the papers did not publish the numbers of MSE or MAPE;

(c) the papers focused on the correlation between some features and flu index instead of forecasting;

(d) the papers with an invalid URL could not be reached;

(e) the papers have no ”full text” access and therefore could not be reached;

(f) the papers that The University of Tokyo does not have access permission could not be reached.

The ”NA”s in Table 1.2 also tells us not so many previous studies focus on the improvement of predictive accuracy of flu outbreak models. However, it is quite practical to achieve an accurate predictive model for flu outbreaks, since the model can effectively and efficiently help the billions of people. Except for the drawbacks we aforementioned about the previous studies, the width and depth of the previous studies are not enough for landing science and technologies for real human lives. In other words, our research is not designed to improve the current algorithm or technologies but explore a pragmatic approach to a real project all over the world. We believe these types of research is quite necessary since it aims at real-world problems, and recently, more scientific focus is being put on these practice areas, such as the research from Airbnb [110]. This paper emphasized the practical technologies and skills in the real project in Airbnb and achieved the Best Paper of KDD 2018. This tells us a trend of scientific research that part of scientific research should put efforts on real-world problems and practical solutions.

1.3 Our Research Plan

The drawbacks of past studies gave us a piece of inspiration for our research plan. To find an effective and efficient approach to perform a geolocational-temporal multistep prediction of flu outbreaks. Accordingly, we designed our research in three steps:

(1) To find the best model and best hyperparameter of flu prediction. We used the flu data from the United States. We scraped the data from the Centers for Disease Control and Prevention. We compared the predictive accuracy of the six common models in statistical, machine learning, and deep learning. We also explored hyperparameters to find the best combination of models and hyperparameter. Besides, we also discussed some key featuring engineering, such as the number of time lags, metrics, and so on. Many experiences from this work were the basis of the 2nd work and the 3rd work and effectively support our future researches.

(2) To find the best algorithms of multistep prediction for flu outbreaks. We used the flu data from the United States. We scraped the data from the Centers for Disease Control and Prevention, just as the first step did. We compared the four types of algorithms of multistep prediction as well as the different number of layers in the neural networks to find the best algorithm of multistep prediction for flu outbreaks. The best combination of models and hyperparameters as well as some other experience found in the first step was applied in the 2nd research, and the 2nd research was the basis of the 3rd work.

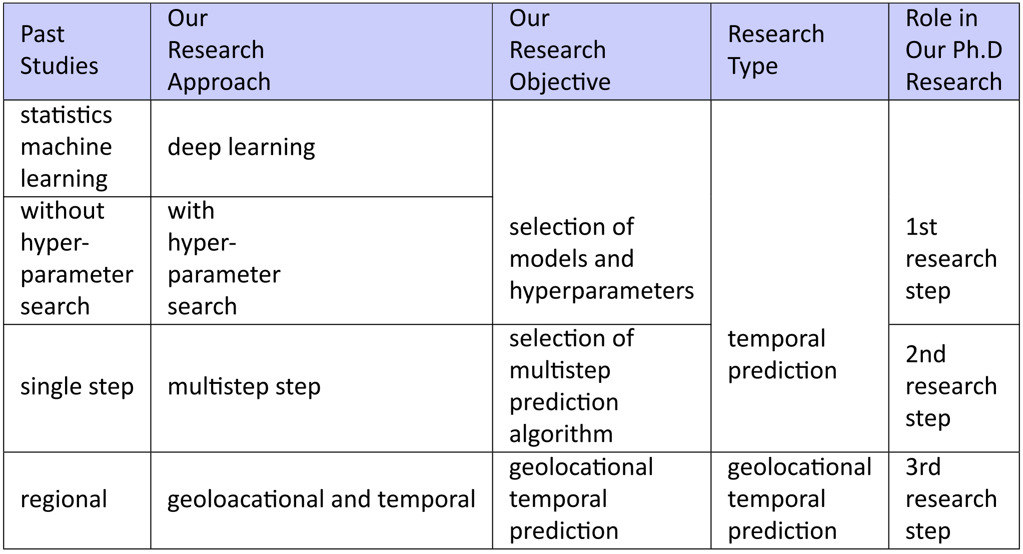
(3) To build up an effective and efficient approach to perform geolocational-temporal multistep of flu. We collected the flu data of all the 152 countries from the World Health Organization. The best combination of models and hyperparameters found in the first step and the best algorithms of

Table 1.2: Predictive Accuracy of Other Previous Studies.

Author	Year	target	best MSE	best MAPE (%)
Michiels et al.	2017	influenza-like illnesses	13.8822	NA
Wu et al.	2017	influenza-like illnesses	NA	4.35
Bu et al.	2018	influenza-like illnesses	1.4333	NA
Guo et al.	2017	the number of flu patients	4393.83	NA
Guo et al.	2017	the number of flu patients	3396.04	NA
Liang et al.	2018	the number of flu patients	42.654	26.197
Wang et al.	2017	influenza-like illnesses	0.014	28.785
Chaudhary et al.	2017	the number of flu patients	NA	NA
Seleznev et al.	2018	the number of flu patients	NA	NA
Fu et al.	2017	influenza-like illnesses	NA	NA
Tung et al.	2015	influenza-like illnesses	NA	NA
Murray et al.	2018	influenza-like illnesses	NA	NA
Spreco et al.	2017	the number of flu patients	NA	0.26
Spreco et al.	2017	the number of flu patients	NA	NA
Alkouz et al.	2018	the number of flu patients	0.0196	NA
Corbella et al.	2018	the number of flu patients	NA	NA
Lee et al.	2017	influenza-like illnesses	NA	NA
Bardak et al.	2017	the number of flu patients	0.061	NA
Verma et al.	2017	influenza-like illnesses	0.07023	NA
Xue et al.	2018	influenza-like illnesses	0.0083	7.3531
Du et al.	2017	the number of flu patients	NA	NA
Du et al.	2018	the number of flu patients	NA	NA
Lu et al.	2018	influenza-like illnesses	0.038	16.3
Belkhiria et al.	2018	the number of flu patients	NA	NA
Paul et al.	2017	influenza-like illnesses	NA	NA
Morita et al.	2018	influenza-like illnesses	NA	NA
Thrastarson et al.	2017	the number of flu patients	NA	NA

The table shows the predictive accuracy of the previous studies in the recent three years. The columns of the ”best MSE” and the ”best MAPE” show the best accuracy of all the models in previous studies, respectively. The ”NA” means ”Not Available” in the columns of ”best MSE” and ”best MAPE”. The reasons of ”NA” includes (a) the papers presented the accuracy in other metrics instead of MSE or MAPE; (b) the papers did not publish the numbers of MSE or MAPE; (c) the papers focused on the correlation between some features and flu index (such as the number of flu patients, or ILI) instead of forecasting; (d) the papers have no an invalid URL and could not be reached; (e) the papers without ”full text” access could not be reached; and (f) the papers that The University of Tokyo does not have access permission could not be reached.

Table 1.3: Comparison of past studies and our research.



Past Studies	Our Research Approach	Our Research Objective	Research Type	Role in Our Ph.D Research
statistics machine learning	deep learning	selection of models and hyperparameters	temporal prediction	1st research step
without hyper-parameter search	with hyper-parameter search			
single step	multistep step	selection of multistep prediction algorithm		2nd research step
regional	geoloacational and temporal	geolocational temporal prediction	geolocational temporal prediction	3rd research step

The column of ”Past Studies” describes the common methods used in past studies for flu prediction. Comparatively, the column of ”Our Research Approach” describes our methods used in our research. Especially, the column of ”Our Research Objective” describes the goal of every step of our researches. Among them, the first research step focused on the selection of models and hyperparameters. The second research performed the selection of multistep prediction algorithms, and the third research built up the whole system of geolocational-temporal multistep prediction. Regarding research types, the first and second steps were temporal predictions; and the third step was a geolocational-temporal prediction.

multistep prediction found in the second research were applied. Table 1.3 compares the approaches of the past studies and our study. The column of ”Past Studies” describes the common methods used in past studies for flu prediction. Comparatively, the column of ”Our Research Approach” describes our methods used in our research. Especially, The column of ”Our Research Objective” describes the goal of every step of our researches. Regarding research types, the first and second steps were temporal prediction; and the third step was the geolocational-temporal prediction. Hopefully, our research could help all countries better prepare for the annual flu outbreak.

Chapter 2

Related Work

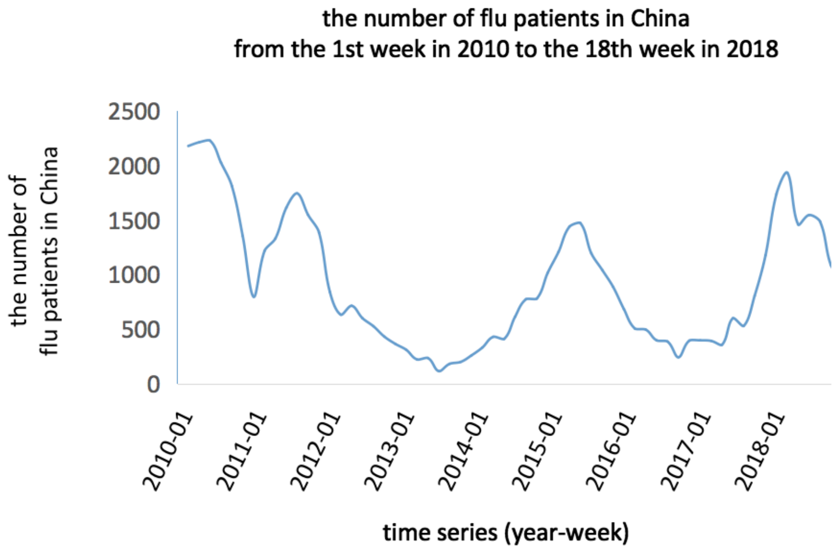


Figure 2.1: The number of flu patients in China.

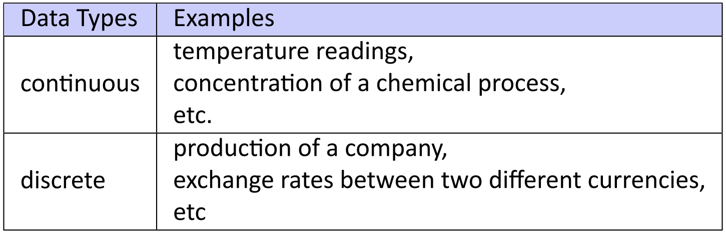
The flu data of China began from the 1st week in 2010 end in the 18th week in 2018.

This section describes previous related research work. In details, firstly, we explain time series and time series data; secondly, we discuss time series analytics that can be applied to time-series predictions.

2.1 Time Series and Time Series Data

Time series is a sequence of time units, such as hour, day, week, month or year. Time series is ordered by equally spaced time intervals. Time series data are a series of data points in a natural temporal ordering. Typical examples are height and time of tiding, temperature, foreign exchange rates, a stock market, and so on. In mathematics, time series data are defined as a set of vectors X(t) where t represents the time elapsed (t = 0, 1, 2, ...) and X represents a vector including all variables at some time spot. The length of X vectors can be 1 or a number larger than 1, in case of the different situation. A single time series data (the length of the X vector is 1) is called univariate, such as the flow rate of a river at some location. A multiple time series data is termed as multivariate. An example could be the price of two relevant stocks, such as Coca-Cola and Pepsi. Time Series are frequently plotted in line charts, such as Figure 2.1. Figure 2.1 plots the number of flu patients in China from the 1st week in 2010 to the 18th week in 2018. Table 2.1 shows, by data types of time series values, time series can be divided into continuous time series or discrete time series. such as the temperature readings, the concentration of a chemical process, and so on. Discrete-time series contains observations at discrete points of time, such as the production of a company, exchange rates between two different currencies, and so on. A continuous time series can be transformed into a discrete one by a sum over a specified time interval, such as aggregating monthly values to yearly values, the process of which is called aggregation.

Table 2.1: Data types of time series values.



Data Types	Examples
continuous	temperature readings, concentration of a chemical process, etc.
discrete	production of a company, exchange rates between two different currencies, etc

2.2 Time Series Analytics and Time Series Prediction

Time series analytics are methods of analyzing time series data to extract meaningful features of the data. Time series analyses have different application. Among them, forecasting is one of the most common objectives of time series analyses. Time series forecasting is applying a model to predicting future values based on previously observed values. We need time series forecasting in many cases: deciding whether to build a hospital in a specified location in the next few years requires forecasts of future population and medical demand; inventory control requires forecasts of future sales; yearly flu vaccine production requires forecasts of regional flu outbreaks. Sometimes, forecasting time series values can be very easy or very hard. For example, astronomers can predict the lunar eclipse precisely. On the other hand, one can hardly forecast tomorrow's stock price because too many social factors could impact the stock market positively and/or negatively. We can apply time series analysis to continuous data, discrete numeric data, and discrete symbolic data. Typical discrete symbolic data are sequences of characters in languages [111].

2.3 Time Series Predictive Models

By using different features, predictive models for time series data can be categorized into 3 types. The first type of model is an autoregressive model, which uses past values as features ("Xs"). Typical examples include the Auto-Regressive Integrated Moving Average (ARIMA) model and the Vector Auto-Regression model (VAR). The second type of model is common regressive models. These models use predictors (such as temperature, humidity, and so on.) instead of past flu data. These models include linear regression, random forest, and so on. The typical example is "Google Flu Trends" [112], which used search engine query data as features and a linear regression model. The third type of model is a combination of the first and second types. It uses the numbers of flu patients in the past as features (as in the first type) and regression models (as in the second type) [78]. We used the third types in the second and third researches.

2.3.1 Auto-Regressive Integrated Moving Average (ARIMA)

An Auto-Regressive Integrated Moving Average (ARIMA) model is one of the most general predictive models of time series values. The theoretical base of ARIMA is that the time series data show autocorrelation. Autocorrelation is also named as a serial correlation. Autocorrelation is a phenomenon that a series of values is correlated with its delayed copy. Many models aim at solving problems of forecasting autocorrelated data, such as autoregressive (AR) models, moving average (MA) models, autoregressive moving average models (ARMA). ARIMA model is a generalization of an ARMA model. ARIMA is composed of three parts: "AR", "I", and "MA". Firstly, the auto-regression (AR) algorithm presents that any future value can be regressed by its own lagged values in the series., as the name indicates “self-regression”. Besides, the moving average (MA) algorithm presents that the error of the regression is an LR of error terms at a variety of the past time spots. Moreover, the integrated (I) algorithm presents that the target values have been replaced with the difference between their values and the previous values, such as first-order difference, second-order difference, and so on. When data show evidence of non-stationarity, a differencing step, perhaps in conjunction with nonlinear transformations such as logging or deflating, can be applied one or more times to eliminate the non-stationarity. That is the difference between ARIMA and ARMA. In non-seasonal ARIMA, there are three parameters, denoted as (p,d,q). The parameters of p, d, and q are non-negative integers. The parameter of ”p” is the number of time lags of the autoregressive model, ”d” is the degree of differencing, and ”q” is the order of the moving-average model. In seasonal ARIMA, there are seven parameters, denoted as (p,d,q)(P, D, Q). The parameter of ”m” is the number of periods in each season, ”P” is autoregressive terms for the seasonal part, ”D” is differencing terms for the seasonal part, ”Q” is moving average terms for the seasonal part. Special cases of ARIMA are as follows:

(1) ARIMA (1,0,0) is AR(1)

(2) ARIMA(0,1,0) is I(1)

(3) ARIMA(0,0,1) is MA(1)

(4) ARIMA(0,1,0) is a random

(5) ARIMA(0,1,0) with a constant is a random walk with drift

(6) ARIMA(0,0,0) is a white noise

(7) ARIMA(0,1,2) is a Damped Holt’s model

(8) ARIMA(0,1,1) without constant is a basic exponential smoothing

(9) ARIMA(0,2,2) is Holt’s linear method with additive errors or double exponential smoothing.

There are many variations of ARIMA. A seasonal ARIMA (SARIMA) is used for time series data with obvious seasonality. A vector ARIMA (VARIMA) model is used for multiple time series. In VARIMA, The dimension of features is larger than one. In other words, a vector of time series features is inputted into the models.

2.3.2 NM Predictive Model

An NM Predictive Model forecasts the cumulative sales quantity for products with a non-linear algorithm. The NM Predictive Models focus same or similar products (usually defined clusters) they have same or similar characteristics. Formula 2.1 shows the equation of the NM Predictive Models.

$$X_M = NM_{(group,N,M)} \times R_N \tag{2.1}$$

where, R_N is cumulative sales quantity until N day(s) / week(s); X_M is prediction of cumulative sales quantity until M day(s) / week(s); $NM_(group, N, M)$ is NM coefficient. There are two steps in NM Predictive Models:

(a) calculate the NM coefficient: NM Predictive Models apply linear algorithms to calculate the NM coefficient, as Figure 2.2 shows. The NM coefficient is actually the slope of the linear regression

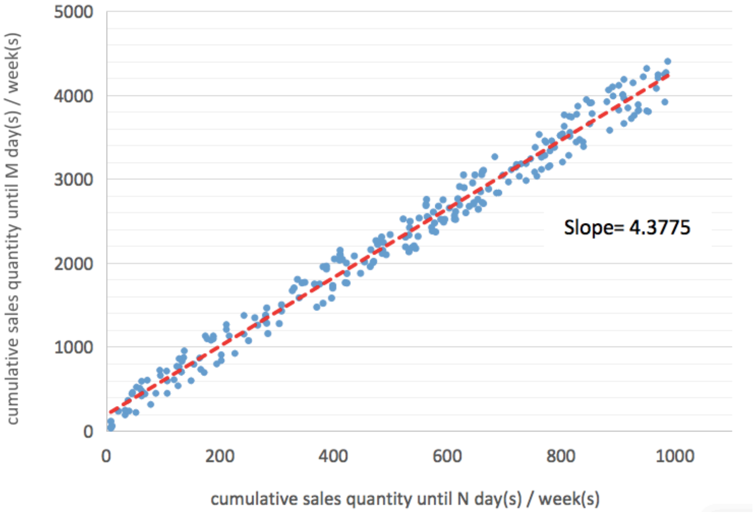


Figure 2.2: NM coefficient calculation.

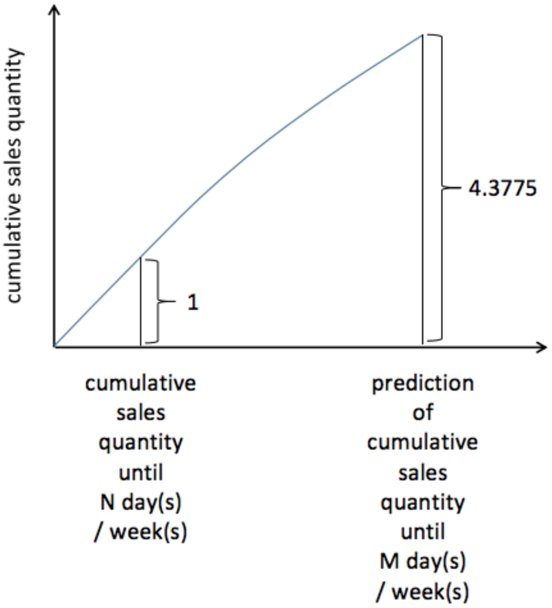

cumulative sales quantity
1
4.3775
cumulative sales quantity until N day(s) / week(s)
prediction of cumulative sales quantity until M day(s) / week(s)

Figure 2.3: Prediction of future sales quantity by NM models.

Prediction of future sales quantity by NM models. The ”4.3775” is calculated in the previous step, i.e. the ”slope”.

between cumulative sales quantity until N day(s) / week(s) and cumulative sales quantity until M day(s) / week(s).

(b) predict future sales quantity: NM Predictive Models predict a future sales quantity by the NM coefficient (i.e. the slope) that has been calculated in the previous step, as Figure 2.3 shows.

2.3.3 Agent-Based Model (ABM)

An Agent-Based Model (ABM) is one of the classic approaches that can be used to model social systems. The model type of ABMs is different from predictive models that aim to regression or classification. An ABM is a progressive model that is designed to simulate a whole system. In detail, an ABM is to create real-world-like complexity by simulating the operations and interactions of multiple agents and complex phenomena of a whole system. Nonetheless, the principle of ABMs is quite simple and unadorned, known as K.I.S.S. (”Keep it simple, stupid”). An ABM consists of rule-based agents that interact with each other step by step. Agents are supposed to apply heuristics or simple decision-making rules and act in what they perceive as their interests, namely ”boundedly rational”. The ”boundedly rational” concepts typically include reproduction, economic benefit, or social status, and so on. [113]. ABMs process inductively in the situation at hand. The modelers/researchers watch phenomena emerge from the agents’ interactions, such as equilibrium, an emergent pattern, an unintelligible mangle, and so on. Figure 2.4 shows an example of the application of ABM. ABMs can be applied to infectious disease transmission by a susceptible-infected-recovered (SIR) framework (Section 2.3.4), which is a traditionally aggregate, compartmental model. An ABM introduces individual heterogeneity and more complex network interactions into SIR and therefore provide further insight into infectious processes [114–116]. By the combination of ABM and SIR, the Centers for Disease Control and Prevention and other government agencies evaluate infection control policies and have thus informed the development of containment strategies [117]. ABMs were used to targeted antiviral prophylaxis and social distancing measures to prevent an H5N1 influenza A (bird flu) pandemic. [114, 118] ABMs were also used to vaccination strategies against flu pandemics, including their impact on health care personnel. [114,119,120], which includes the Models of Infectious Disease Agent Study (MIDAS) [114, 121]. Figure 2.5 shows an example of the application of the population health of the ABM. Individual characteristics include demographics, health behaviors, health conditions, and health service utilization. They are influenced by community characteristics, social ties, and other contacts. Ongoing processes include aging and movement in the environment. Population health emerges from a system that is created by these static and time-varying characteristics at multiple levels and the often bidirectional processes that connect them [114]. Researchers also applied ABMs to obtaining insight into health behaviors. Those behaviors increase the risk of disease, as well as potential interventions to reduce risky behaviors, such as smoking, alcohol consumption, physical inactivity, and unhealthy eating. In spite of the advantages from ABMs, due to the nature of ABM, ABMs also have limitations and challenges. As a principle, modelers follows the KISS principle. However, at the same time, they also prefer to try achieving meaningful results for potential interventions and public health planning by taking advantage of the complexity in ABMs and exploring critical elements of systems [114, 122, 123]. How to balance between the necessary for simplified representations of the real world and the need to include enough complex elements to provide new insights, assumes a pivotal role in researches [114, 123]. Besides, the robustness of the ABM seriously depends on empiric data. Empiric data used in ABM usually comes from observational studies. Thereby, the amount of the data is limited and impact the robust of

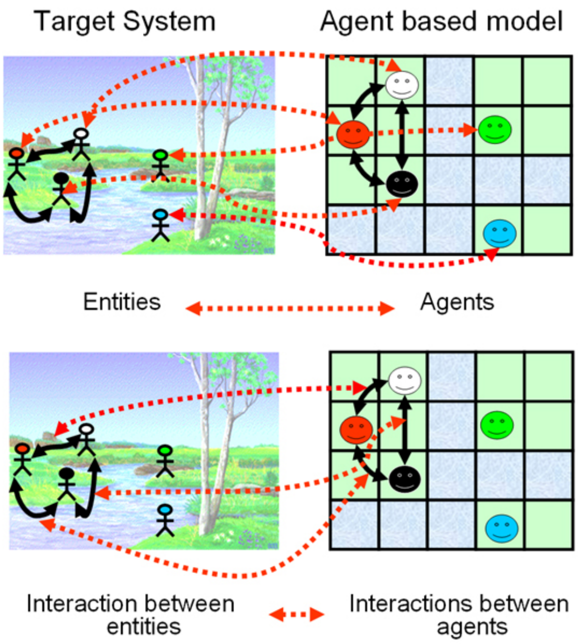


Figure 2.4: An example of Agent-Based Models (ABM).

An example of Agent-Based Models (ABM). The figure is cited from [124].

the ABM negatively because of various distributions due to the different objectives of observational studies. So the bias of ABMs could be high. Furthermore, we usually input every data into the model to make the most of ABM's advantage over more traditional approaches: social network influences and the strength of interactions between units Nonetheless, validation itself is challenging when validation dataset are scarce because, ideally, data used for validation purposes should be independent of those used to build and calibrate the model. This phenomenon is called overfit-ting. So the variance of ABMs could be also very high. The high bias and variance negatively impact the accuracy of almost every application of ABM to infectious studies in public health.

2.3.4 SIR Model

A SIR model is a model, which is used to simplify the mathematical progress of infectious disease [115, 125]. Each person of the population typically progresses from susceptible to infectious to recovered. This can be shown as a flow (Figure 2.6), in which the boxes represent the different compartments and the arrows the transition between compartments. The following derivation was from the explanation of David Smith and Lang Moore [126] and of Peng Feng [127]. By formula derivation of the SIR model, we can also get the conclusion that ”The number of people who may be infected with any epidemic will always decrease.” In detail, the model consists of three compartments. The first set of dependent variables counts people in each of the groups, each as a function of time, as Formula 2.2 show. They are $S(t)$ is the number of susceptible individuals; $I(t)$ is the number of infected individuals; $R(t)$ is the number of recovered individuals. If N is the total population, we have Formula 2.2, as follows.

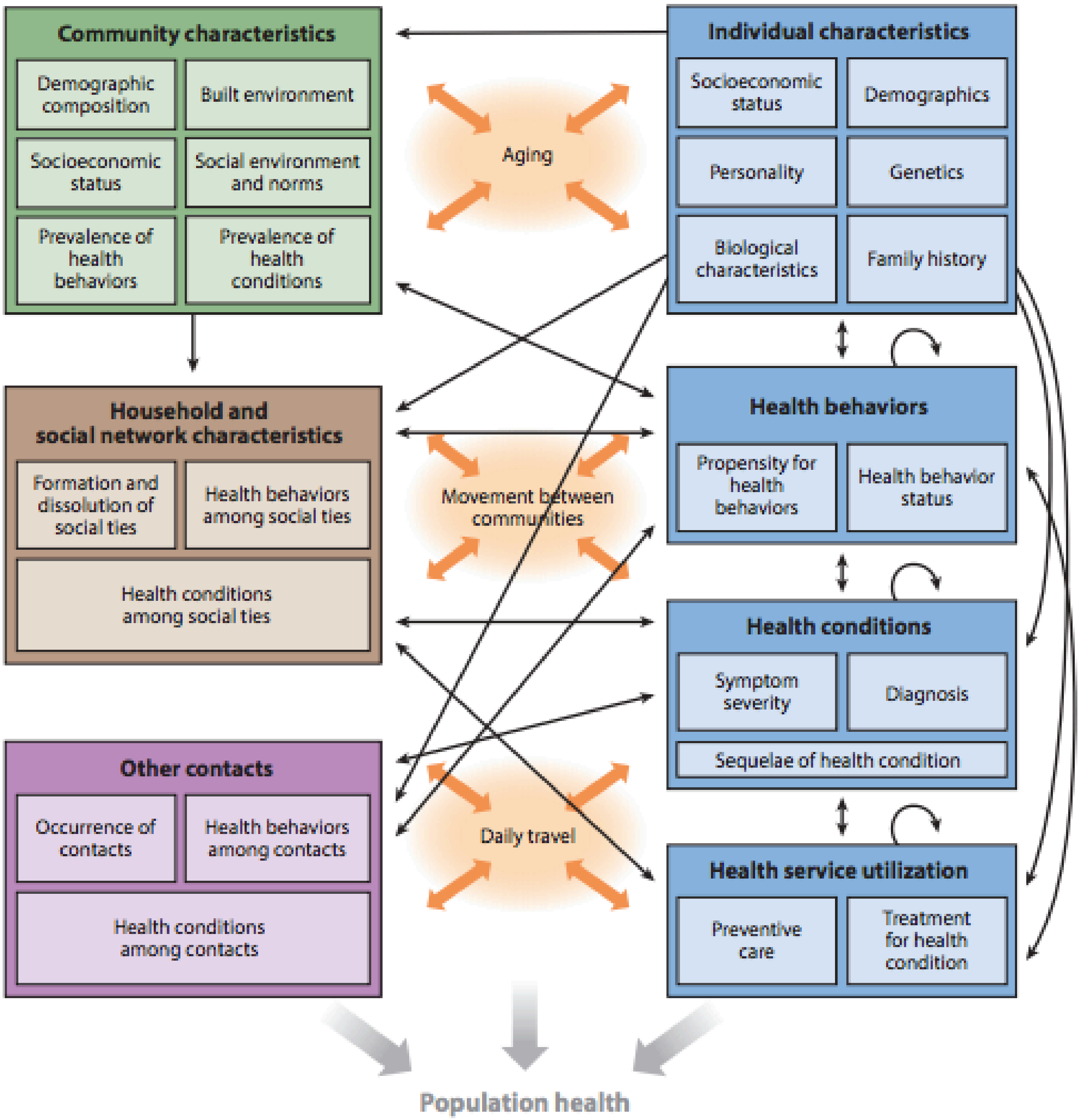

Community characteristics
Demographic composition
Built environment
Socioeconomic status
Social environment and norms
Prevalence of health behaviors
Prevalence of health conditions
Individual characteristics
Socioeconomic status
Demographics
Personality
Genetics
Biological characteristics
Family history
Aging
Household and social network characteristics
Formation and dissolution of social ties
Health behaviors among social ties
Health conditions among social ties
Movement between communities
Health behaviors
Propensity for health behaviors
Health behavior status
Health conditions
Symptom severity
Diagnosis
Sequelae of health condition
Other contacts
Occurrence of contacts
Health behaviors among contacts
Health conditions among contacts
Daily travel
Health service utilization
Preventive care
Treatment for health condition
Population health

Figure 2.5: An illustration of a hypothetical agent-based model.

Individual characteristics such as demographics, health behaviors, health conditions, and health service utilization (blue) influence and are influenced by community characteristics (green), social ties (brown), and other contacts (purple), as well as ongoing processes such as aging and movement through the environment (orange). Taking together, these static and time-varying characteristics at multiple levels and the often bidirectional processes that connect them create a system from which population health emerges. The figure is cited from [114].

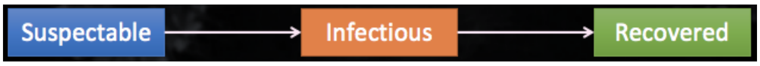


Figure 2.6: Three compartments of infectious disease.

$$\begin{aligned} S &= S(t) \\ I &= I(t) \\ R &= R(t) \\ S(t) + I(t) + R(t) &= N(constant) \end{aligned} \tag{2.2}$$

The second set of dependent variables represents the fraction of the total population in each of the three categories, as Formula 2.3 show. The $s(t)$ is the susceptible fraction of the population. The $i(t)$ is the infected fraction of the population. The $r(t)$ is the recovered fraction of the population. From Formula 2.2, we have Formula 2.3, as follows.

$$\begin{aligned} s &= \frac{S(t)}{N} \\ i &= \frac{I(t)}{N} \\ r &= \frac{R(t)}{N} \\ s(t) + i(t) + r(t) &= 1 \end{aligned} \tag{2.3}$$

We assume that the time-rate of change of $S(t)$ depends on the number already susceptible, the number of individuals already infected, and the amount of contact between susceptible and infected individuals. In particular, suppose that each infected individual has a fixed number, β, of contacts per day that are sufficient to spread the disease.

$$\beta = c\,\chi \tag{2.4}$$

In Formula 2.4, c is the number of contacts in the time unit, and χ is infectiveness with an infective person. The fraction of these contacts that are with susceptible individuals is $s(t)$. Thus, on average, each infected individual generates $\beta\ s(t)$ new infected individuals per day. So we have Formula 2.5, as follows.

$$\begin{aligned} \frac{ds}{dt} &= -\beta\ s(t)\ i(t) \\ \frac{dS}{dt} = -\beta\ s(t)\ I(t) &= -\frac{\beta\ S(t)\ I(t)}{N} \end{aligned} \tag{2.5}$$

We also assume that a fixed fraction γ of the infected group will recover during any given day.

$$\begin{aligned} \frac{dr}{dt} &= \gamma\ i(t) \\ \frac{dR}{dt} &= \gamma\ I(t) \end{aligned} \tag{2.6}$$

τ is the average time spent as an infective, i.e. the average duration of the infection. (Formula 2.7) For example, if the average duration of infection is three days, then, on average, one-third of the currently infected population recovers each day.

$$\gamma = \frac{1}{\tau} \tag{2.7}$$

Since the total population has no change, we get Formula 2.8.

$$\begin{aligned} \frac{ds}{dt} + \frac{di}{dt} + \frac{dr}{dt} = 0 \\ \frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0 \end{aligned} \tag{2.8}$$

From Formulae 2.5 and 2.6 and 2.8, we get Formula 2.9.

$$\begin{aligned} \frac{di}{dt} &= \beta\ s(t)\ i(t) - \gamma\ i(t) \\ \frac{dI}{dt} = \beta\ s(t)\ I(t) - \gamma\ I(t) = \frac{\beta\ S(t)\ I(t)}{N} &- \gamma\ I(t) \end{aligned} \tag{2.9}$$

Now we can discover how $s(t)$, $i(t)$, and $r(t)$ will act, as t goes. From the Formulae 2.2 and 2.6, we have Formula 2.10.

$$\frac{d}{dt}(S(t) + I(t)) = -R(t) = -\gamma\ I(t) < 0 \tag{2.10}$$

Formula 2.10 yields Formula 2.11.

$$S(t) + I(t) \leq N \tag{2.11}$$

From Formula 2.11, we get Formula 2.12.

$$t \to \infty, S(t) \searrow S_\infty \tag{2.12}$$

Besides, Formulae 2.2 and 2.10 yield Formula 2.13.

$$\gamma \int_0^t I(s)d(s) = R(t) = N - S(t) - I(t) \leq N \tag{2.13}$$

From Formula 2.13, we get Formula 2.14.

$$\int_0^t I(s)d(s) < +\infty \tag{2.14}$$

Consequently, we have Formula 2.15.

$$t \to \infty, I(t) \to N - S_\infty - \gamma \int_0^t I(s)d(s) \tag{2.15}$$

Formula 2.15 implies Formula 2.16.

$$t \to \infty, I(t) \to 0 \tag{2.16}$$

Concerning the value of S_∞, which is called the size of the epidemics because is a measure of its strength From Formula 2.5, we have Formula 2.17.

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta\ S(t)\ I(t)}{N} \\ \frac{1}{S(t)}\ dS &= -\frac{\beta}{N}\ I(t)\ dt \\ \int_0^\infty \frac{1}{S(t)}\ dS &= -\frac{\beta}{N}\ \int_0^\infty I(t)\ dt \\ \ln(S_\infty) - \ln(S_0) &= -\frac{\beta}{N}\ \int_0^\infty I(t)\ dt \end{aligned} \tag{2.17}$$

From Formula 2.13, we have Formula 2.18.

$$\int_0^t I(s)d(s) \leq \frac{N}{\gamma} \tag{2.18}$$

From Formulae 2.17 and 2.18, we have Formula 2.19.

$$\begin{aligned} \ln(S_\infty) - \ln(S_0) = -\frac{\beta}{N}\ \int_0^\infty I(t)\ dt \geq -\frac{\beta}{N}\ \frac{N}{\gamma} &= \frac{\beta}{\gamma} \\ e^{\ln(S_\infty)-\ln(S_0)} &\geq e^{\beta/\gamma} \\ \frac{S_\infty}{S_0} &\geq e^{\beta/\gamma} \\ S_\infty &\geq S_0\ e^{\beta/\gamma} \end{aligned} \tag{2.19}$$

Formula 2.19 shows why the number of susceptible people will not be depleted even at the end of the epidemic. From Formula 2.9, we have Formula 2.20.

$$\frac{dI}{dt} = (\frac{\beta}{N}\ S(t) - \gamma)\ I(t) \tag{2.20}$$

Let us define Formula 2.21.

$$\lambda = \frac{\beta}{N} \tag{2.21}$$

From Formulae 2.20 and 2.21, we have Formula 2.22.

$$\frac{dI}{dt} = (\lambda\ S(t) - \gamma)\ I(t) \tag{2.22}$$

Thereby, we have Formula 2.23.

$$\begin{aligned} (if)\ \lambda\ S_0 \leq \gamma,\ (then)\ \frac{dI}{dt} &\leq 0; \\ (if)\ \lambda\ S_0 > \gamma,\ (when)\ t < t^*,\ (then)\ \frac{dI}{dt} &> 0 \\ (if)\ \lambda\ S_0 > \gamma,\ (when)\ t \geq t^*,\ (then)\ \frac{dI}{dt} &\leq 0 \end{aligned} \tag{2.23}$$

In Formula 2.23, t^* is defined by Formula 2.24.

$$\lambda\ S(t^*) = \gamma \tag{2.24}$$

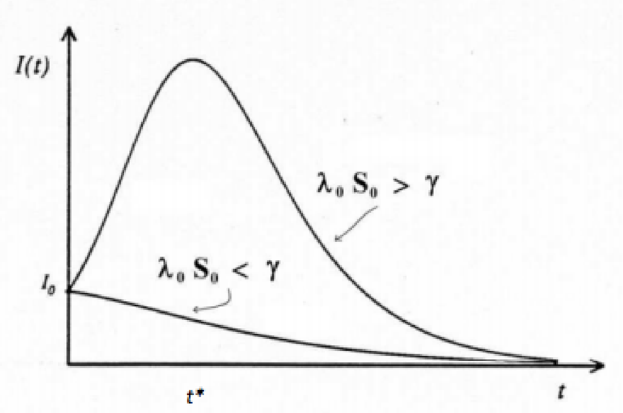


Figure 2.7: The two possible curves of the number of the infected people

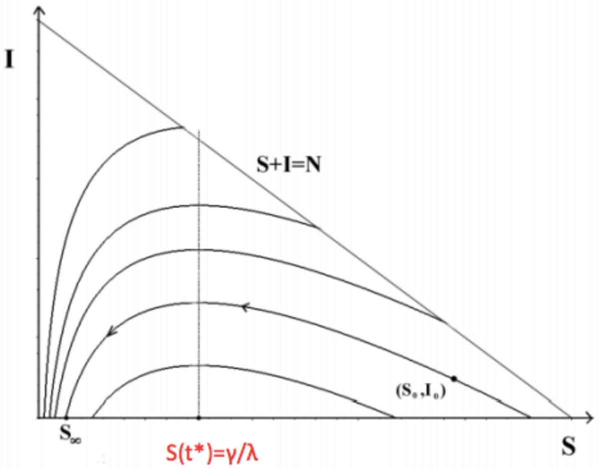


Figure 2.8: The (S, I) trajectories.

The two curves in Figure 2.7 correspond to Formulae 2.20 and 2.24. Figure 2.8 shows the (S, I) trajectories. Actually all trajectories are represented by Formula 2.25.

$$S + I - \frac{\gamma}{\beta}\,N\ \ln(S) = constant \tag{2.25}$$

Therefore, theoretically, we can conclude that the number of people who may be infected with any epidemic will always decrease.

2.3.5 Support Vector Regression (SVR)

A Support Vector Regression (SVR) model can also be used as a predictive model. SVR is a regression model, coming from the Support Vector Machine (SVM), which is a supervised learning model of classification. A trained SVM assigns new instances to one category or the other. An SVM model is a representation of generalized hyperplane, by which the separate categories are divided by a clear gap. The gap is as wide as possible. New examples are mapped into all spaces and predicted to a category based on which side of the gap they fall. SVMs can efficiently perform a linear and non-linear classification by a trick called Kernel. Kernel maps their inputs into high-dimensional feature spaces. Usually, the original feature spaces are limited in a finite dimensional space, while the response might not be linearly regressed of classified in that space. As a result, mapping the

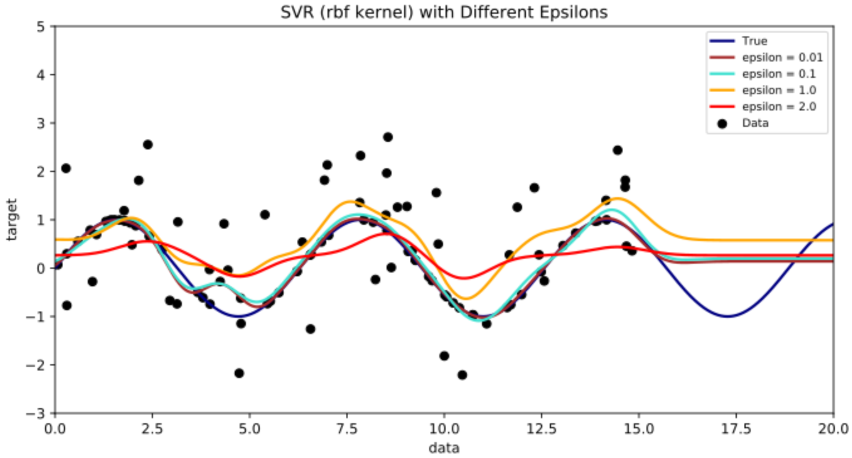


Figure 2.9: Support Vector Machine of Radial Basis Function with Different Epsilons.

The data were generated by normally perturbing a sine curve. The plot was prepared using Scikit-Learn. The figure is cited from [128].

original finite-dimensional space into a much higher-dimensional space presumably helps improve the regression accuracy. One trick of designing kernel is that dot products of pairs input data vectors can be computed easily by defining them in terms of a kernel function, expressed as K(x, x_i), selected to suit the problem. The kernel function is one of the core concepts of SVR. Different types of SVR kernels keep the computational load reasonable and efficiently perform a linear or non-linear regression by implicitly mapping their inputs into high-dimensional feature spaces. Another core concept of SVR is a margin of tolerance (ϵ). The main purpose is to minimize error and maximizes the margin. Instances that fall within the margin do not incur any loss cost, that is why SVR models refer to the loss as "ϵ-insensitive". Figure 2.9 shows different SVR models with different Epsilons. The data was generated by normally perturbing a sine curve. The plot was prepared using Scikit-Learn. The figure is cited from [128]. SVR models depend only on a subset of the training data because the cost function for building the model ignores any training data close to the model prediction. In other words, SVR models are to find a function, f(x), with at most ϵ-deviation (which is called soft margin in SVR and SVM) from the response (y).

2.3.6 Random Forest (RF)

A Random Forest (RF) model is an ensemble learning method. RF can also be used as a predictive model. The basic building block of an RF model is a decision tree. A decision tree comes from observations (called nods in decision trees)of an instance to the target (called leaves in decision trees) of the instance. Figure 2.10 shows an example of a regressive decision tree. The case in Figure 2.10 estimates the probability of kyphosis after surgery. The features are the age of the patient and the vertebra where surgery was started. From left to right, the same tree is shown in three different ways. In the left sub-figures, the leaves show the kyphosis's probability (colorful number), and the percentage of patients (percentage number). The middle sub-figures illustrate a perspective plot of the tree. The right sub-figure presents an aerial view of the middle plot. The probability of kyphosis is higher in darker areas. The example is cited from [129]. Decision trees

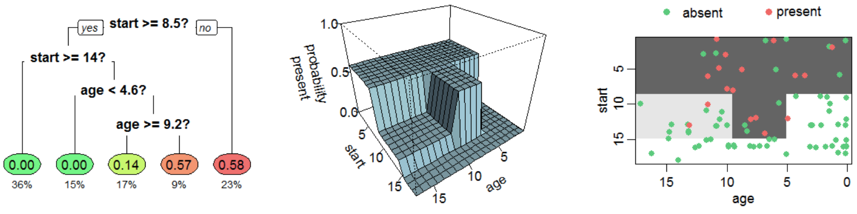


Figure 2.10: An example of the regressive decision tree.

This example of the regressive decision tree estimates the probability of kyphosis after surgery, given the age of the patient and where at which surgery was started. From left to right, the same tree is shown in three different ways. Left: the leaves show the kyphosis's probability (colorful number), and the percentage of patients (percentage number). Middle: a perspective plot of the tree. Right: presents an aerial view of the middle plot. The probability of kyphosis is higher in darker areas. The example is cited from [129].

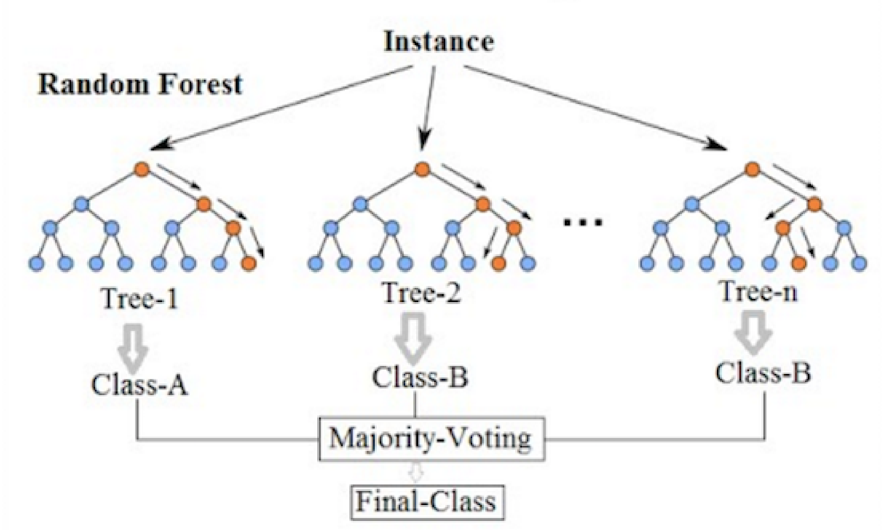


Figure 2.11: The Algorithm of Random Forest.

The classification of random forest trains a great many trees. Then the classification random forest feeds instances into those trees, performs majority voting, and outputs the final class. The figure is cited from [132].

have some limitations, as follows:

(1) Trees can be very non-robust. In other words, trees have low bias and high variance. The ”high variance” means that a small error in the training data can result in a large error in output [130].

(2) Decision-tree learners can create over-complex trees that do not generalize well from the training data. In other words, if decision trees are too deep, decision trees overfit training sets [131]. Overfit is the output of analysis corresponds too closely or exactly to a particular set of data (usually called a training dataset), and may, therefore, fail to fit other data (usually called a testing dataset) reliably [130].

RF removes the limitation of decision trees [133]. The first algorithm of random forests was created by Tin Kam Ho [134]. An extension of the algorithm was developed by Leo Breiman [135]. Figure 2.11 shows the algorithm of RF. To get a more accurate and stable prediction, RF trains a great many decision trees and merges them by averaging in the case of regression or voting in the case of classification. By averaging or voting, RF models improve the testing accuracy at the expense of a small increase in the bias and some loss of interpretability. RF models train on different parts of the same training set. RF samples a specified fraction of instances with replacement. The process of sampling is called ”bagging” and the specified fraction is called out of the bag (OOB). By OOB, RF achieves a running unbiased estimate of the error. After all trees are built, all instances run down one part of all trees, and proximities are calculated for all instances. At the end of the run, the proximities are normalized by dividing by the number of trees. The OOB algorithm can also help get estimates of variable importance.

2.3.7 Gradient Boosting (GB)

Gradient Boosting (GB) is also an ensemble learning method. GB can also be used as a predictive model, which belong to the second type. The characteristics of GB is that GB accumulates weak ”learners” into a strong “learner” in an iterative fashion by additive learning. Take the least-squares regression as an example, where the goal is to ”teach” a model $f(x)$ to predict values of the form $y = f(x)$ by minimizing the mean squared error (MSE). At each stage $m(1 <= m <= M)$ of GB, the model of GB might be still imperfect, i.e. a very weak learners. GB improves f_m by training a new model $h(x)$. A perfect h would imply Formula 2.26.

$$\begin{aligned} f_{m+1}(x) &= f_m(x) + h(x) \\ f_{m+1}(x) &= f_m(x) + h(x) = y \\ h(x) &= y - f_m(x) \end{aligned} \tag{2.26}$$

In other words, GB will fit h to the Formula 2.26. In another explanation, GB trains each f_{m+1}(x) to correct the errors of its predecessor f_m(x). That is why GB is a gradient descent algorithm, and generalizing it entails ”plugging in” a different loss and its gradient. The gradient boosting method assumes a real-valued y and seeks an approximation in the form of a weighted sum of functions, called base (or weak) learners. GB starts with a model, consisting of a constant function f_0(x), and incrementally expands it greedily. Different from RF, GB could overfit training data. Several regularization techniques reduce this overfit effect by constraining the fitting procedure, as follows:

(a) Gradient Boosting Iterations

When the base learner is a decision tree, the gradient boosting iterations M is the number of trees. Increasing M reduces the error on the training set, but setting it too high may lead to over-fitting. An optimal value of M is often selected by monitoring prediction error on a separate validation data set.

(b) Shrinkage

Shrinkage modifies the update rule as Formula 2.27 shows.

$$f_{m+1}(x) = f_m(x) + \nu \cdot \gamma_m \cdot h_m(x) \tag{2.27}$$

In Formula 2.27, parameter ν is called the ”learning rate”. However, it comes at the price of increased computational time both during training and querying: lower learning rate requires more iterations.

(c) Stochastic gradient boosting

At each iteration of the algorithm, a base learner of GB could be fit on a subsample of the training set drawn at random without replacement.

(d) Penalize Complexity of Tree

To perform regularization of models, one usually adds a penalty term to the loss function. The model complexity can be defined as the proportional number of leaves in the trained trees. The joint optimization of predictive loss and model complexity corresponds to a post-pruning algorithm to remove branches that fail to reduce the loss by a threshold.

2.3.8 Artificial Neural Network (ANN)

The concept of an Artificial Neural Network (ANN) model is based on the biological neural network (BNN) in brains. There are approximately 100 billion neurons in the human brain. Electro-chemical signals communicate neurons. When the sum of the signals surpasses a threshold, a response goes through the axon. The ANN performs as the computational mirror of the BNN. However, ANN is not comparable to BNN since the number and complexity of neurons and the used in a BNN is many times more than those in an ANN. The artificial neurons in ANN are called ”nodes”. These nodes are connected, and the strength of their connections to one another is assigned a value (called weights) based on their strength. Nodes are organized in layers. Layers are made up of many interconnected nodes which contain an activation function. The input layer takes in original information (i.e. features). This information pass throughout the network. Based on the weights, the information is passed from node to node. Each of the nodes sums the received information and then perform an activation function. The information flows through the network, through hidden layers, until it reaches the output nodes. The difference between the predicted value and the actual value (i.e. error) will be propagated backward to each node’s weights, which is called backpropagation. A large number of epochs usually help ANN determine the best solution. Most learning rules have built-in mathematical terms to assist in this process which control the ’speed’ (Beta-coefficient) and the ’momentum’ of the learning. The speed of learning is the rate of convergence between the current solution and the global minimum. Once an ANN achieves convergence, it could be used as a predictive tool. The ANN model only works in forward propagation mode only. Figure 2.12 shows a typical structure of ANN. An artificial neural network is interconnected nodes. The structure is similar to the vast network of neurons in a brain. In Figure 2.12, each node represents an artificial neuron and an arrow represents a connection from the input to the outputs. ANN can also be used as a predictive model, which belongs to the second type.

2.3.9 Long Short Term Memory (LSTM)

A recurrent neural network (RNN), and in particular the long-short term memory unit (LSTM) performs the state-of-the-art in time series prediction [137, 138]. LSTM networks are good at classifying, processing and making predictions based on time series data [139]. The efficiency of these networks can be explained by the recurrent connections that allow the network to access the entire history of previous time series values [139]. Figure 2.13 shows the internal structure of LSTM. The core components of a common LSTM unit are LSTM cells and three gates, i.e. an input gate, an output gate, and a forget gate. The LSTM cell passes values over time intervals. The LSTMs remove or add information to the cell state by regulating by structures called gates. The three gates

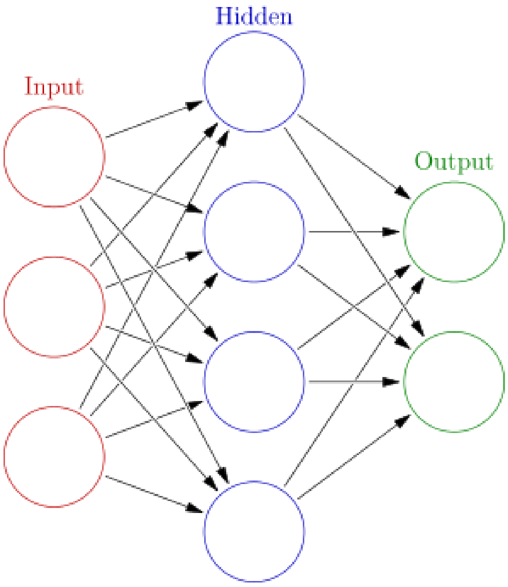


Figure 2.12: A typical structure of artificial neural network.

Each node represents an artificial neuron and an arrow represents a connection from the input to the outputs. The figure is cited from [136].

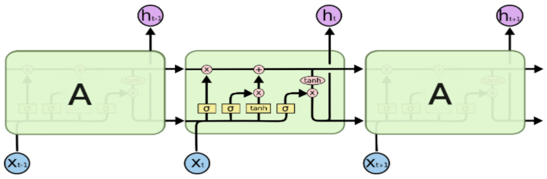


Figure 2.13: The repeating module in an LSTM contains four interacting layers.

The core components of a common LSTM unit are LSTM cells and three gates, i.e. an input gate, an output gate, and a forget gate. The LSTM cell passes values over time intervals. The LSTMs remove or add information to the cell state by regulating by structures called gates. The figure is cited from [139].

adjust the flow of information into and out of the cell. Gates optionally let information through [139]. Connections between nodes form a directed graph along a sequence. LSTM's elaborate structure (multilayers and gated cells) enables LSTM to learn simulate nonlinear function, long-term dependencies [137], and refine time-series prediction [140]. LSTM is applied to predict future values in many fields [141], such as financial services [142, 143], and so on. Other applications are primarily from major technology companies. Google applied LSTM to speech recognition on the smartphone [144, 145], for the smart assistant Allo [146] and for Google Translate [147, 148]. Apple applied LSTM for the "Quicktype" function on the iPhone [149, 150] and for Siri [151]. Amazon uses LSTM for Amazon Alexa [152]. Microsoft reported reaching 95.1% recognition accuracy on the Switchboard corpus, incorporating a vocabulary of 165,000 words. The approach used "dialog session-based long-short-term memory" [153].

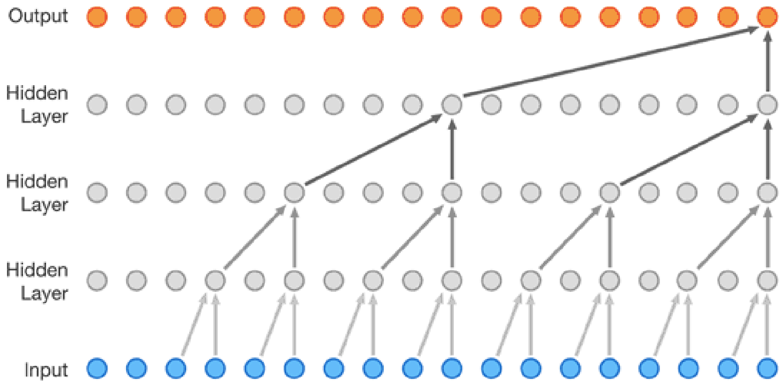


Figure 2.14: A typical structure of artificial neural network.

Each node represents an artificial neuron and an arrow represents a connection from the input to the outputs. The figure is cited from [158].

2.3.10 Convolution Neural Network (CNN)

Convolution Neural Network (CNN) is s a biologically-inspired type of neural network (emulating the response of an individual neuron to visual stimuli). In a CNN model, there are convolutional layers, pooling layers, fully connected layers, and normalization layers. Convolutional layers use sliding a filter (mathematically a weight matrix) over inputs or neurons of previous layers and compute the dot product between inputs or neurons of previous layers and the sliding filter at each data point. A CNN model consists of a sequence of convolutional layers, the output of which is connected only to local regions in inputs or neurons of previous layers. This structure enables filters in CNNs to recognize specific patterns in inputs or neurons of previous layers local regions. That is why data scientists apply CNN to analyzing visual imagery [154, 155]. We can also apply CNN to time series prediction. Similar to the situation of image analytics, filters in CNN can learn and extract specific repeating patterns in the series in local regions. Google researchers published PixelRNN [156] and PixelCNN model [157]. The results showed that generating complex natural images not only one pixel at a time, but one color-channel at a time is also feasible, which inspired Google researchers to adapt our two-dimensional PixelNets to a one-dimensional WaveNet [158]. WaveNet is a neural network that generates raw audio [159]. Figure 2.14 shows the structure of WaveNet [158]. It is a fully convolutional neural network, where the convolutional layers have various dilation factors that allow its receptive field to grow exponentially with depth and cover thousands of time steps [158]. We can regard generating raw audio as a process of predicting future raw audio that mimics real raw audio as much as possible. Time series prediction is also a process of predicting future values that mimics real values as much as possible. In this way, many successive researches followed WaveNet and studied on how to apply CNN to time series classification [160] and time series regression [161–163]. One might employ a CNN with multiple layers of dilated convolutions [164]. The algorithm of dilated convolutions is to apply filter by skipping certain elements in the input, allow for the receptive field of the network to grow exponentially. Hereby, the dilated convolutions allowing the whole network to access a broad range of historical values.

2.3.11 Attention Mechanism Applied in Time Series Prediction

The attention mechanism is an adjustment that equips a neural network with the ability to focus on a subset of its inputs (or features). In other words, attention algorithm filters inputs by applying weights to all of the inputs [165]. In mathematics, let $\chi \in \mathrm{R}^d$ be an input vector, z $\in \mathrm{R}^k$ be a feature vector, $\mathrm{f}_\theta(\chi)$ be a forward neural network with parameters θ, a $\in$ [0,1]^k be an attention vector, g $\in \mathrm{R}^k$ be an attention glimpse, and $\mathrm{f}_\phi(\chi)$ be an attention network with parameters ϕ. Usually, attention is implemented as Formula 2.28 [165].

$$\begin{aligned} a &= f_\phi(\chi) \\ z &= f_\theta(\chi) \\ g &= a \odot z \end{aligned} \tag{2.28}$$

There are two types of attention: soft attention, and hard attention. Soft attention multiplies features with a (soft) mask of values between zero and one. Hard attention multiplies features with a (hard) mask of values, which are exactly zero or one, namely a $\in$ {0,1}^{k}. In the latter case, we can use the hard attention mask to directly index the feature vector: $\tilde{g}$ =z[a] (in Matlab notation), which changes its dimensionality and now $\tilde{g} \in \mathrm{R}^m$ with m$\leq$k [165]. A neural net is a series of matrix multiplications and element-wise non-linearities, where elements of the input or feature vectors interact with each other only by addition. Comparatively, attention mechanisms compute a mask which is used to multiply features, by which the space of functions that can be well approximated by a neural net is vastly expanded [165].

Chapter 3

Method

This section describes the research methods that we adopted in all the research steps. In the first subsection, we describe, in the first research, how we scraped the source data and process them and how we adjusted the six analytical models (ARIMA, SVR, RF, GB, ANN, and LSTM) to perform single-step prediction (such as the structure, the hyperparameter, the programming language, and the metrics of the models). In the second subsection, we explain the source data and pretreatment of the source data in the second research and four algorithms of multistep prediction. Besides, we also illustrate the structure selections and the hyperparameter adjustment of the LSTM used to implement the four multistep predicting algorithms. In the third subsection, we illustrate the methodology of the 152 countries' flu data scrape and geolocational-temporal multistep prediction. We also performed the RF and SVR and used their results as baselines to compare the results. we provide all the details of the model structure, the model hyperparameter, the programming language, and the analytical metrics.

3.1 Methods of Comparative Study on Models

This is the first subsection of the method. We explain all the methods we used for the first research related to data and models.

3.1.1 Experiment Data

This part describes the data and the data pretreatment for the first, including data source, the method to tackle “not available” (N/A) values, the process of response (y), the historical plot of the data, the split of data into training and testing, and the process of features (Xs) for the models.

(a) Data Source

In the first research step, we collected all the U.S. flu season data from the “FluView” Portal of the website for the Centers for Disease Control and Prevention.

(b) Tackling N/As

The data are posted “weekly” with “not available” (N/A) values from the 21st week to the 39th week of 1998, 1999, 2000, 2001, and 2002. One could not find an official explanation of the missing data from the “FluView” Portal. One possible explanation is that there was no flu patient in the U.S in these weeks. In other words, the numbers were ”zeros”. These weeks (from the 21st to the 39th) were usually from the end of May to the end of September and were near or in summers. The flu seasons occurred in winters and early springs in temperature zones where the U.S. locates, and the historical records sometimes omitted the ”zeros”. However, if we simply fill ”zeros”, there could be three problems. First, filling ”zeros” conflicts the analytical metrics. We adopt the mean absolute percentage error in the first research (also in the second and the third) due to some realistic reason (please refer to the description of metrics). We cannot use ”zeros” to calculate the mean absolute percentage errors since one cannot use ”zeros” as denominators. Second, even in the weeks of the 21st to the 39th from 2003 to 2017 (totally 15 years), more or less, there were still some flu patients. The flu is becoming increasingly serious and flu patients existed even in summers. Filling ”zeros” in the summers of the first five years (1998, 1999, 2000, 2001, and 2002) makes no sense and would disrupt the future prediction since it gives models false appearance that the flu data could be ”zeros”, to some extent. Finally, this explanation is just our inference, we cannot guarantee our explanation is really correct. Based on these three reasons, we did not simply fill these N/As with ”zeros”. We might have two probable solutions to these N/As. One is to interpolate the missing data with some special analytical methods, such as Kalman Filter or just simple linear interpolation. If we adopt this method, there could be problems. First, these weeks locate in troughs (the low area between two peaks). Simply supposing the lowest troughs locates the midpoints in the two neighbor peaks brings huge errors. The flu seasons vary year by year, which made the troughs move left and right year by year. The unjust supposing influence not only the points of the lowest troughs but also the neighbor tens of points. To make the matter worse, the unjust interpolation makes the predicting models believe the flu seasons vary limitedly year by year. As a result, the predicting models produce more predicting errors, which we are trying our best to avoid in principle. The other solution is to give up the data before the 40th week of 2002. To keep the data's originality, we adopted this method and gave up any rendering method. We only used the U.S. Flu Season Data from the 40th week of 2002 to the 30th week of 2017.

(c) Response(y)

In predicting models, we have Xs, which are also called features, input, independent variables, or sometimes just variables, and y, which are also called response, target, output, dependent variable. For time series predicting models, one can directly use flu data as response. However, directly using flu data barely takes into population fluctuation into account, such as immigration, emigration, baby booms, aging society, and so on. We prefer to reflect the severity of flu seasons in a percentage, called Influenza-Like Illness (ILI) rate, more precisely. The ILI rates are calculated by a formula, where the number of ILI is divided by the total number of illness, as the formula 3.1 shows. The variation of the ILI rates reflect the relative severity of every flu season and removes other irrelevant factors such as population fluctuation.

$$ILIrate = \frac{the\_number\_of\_ILI}{total\_number\_of\_illness} \tag{3.1}$$

Since we have chosen weekly ILI rates as responses, we can complete the historical plot. Figure 3.1 presents a historical plot of the U.S. ILI rates from the 40th week of 2002 to the 30th week of 2017. In Figure 3.1, the Y-axis represents the weekly ILI rate, and the X-axis represents the time series. We can easily find flu data's seasonality. Besides, flu seasons vary in timing, severity, and duration. Flu seasons also have unprecedented pandemic break such as in 2009 when swine flu occurred. The 2009 flu pandemic was the second of two pandemics involving the H1N1 flu virus (the first was the 1918 flu pandemic), albeit a new variety.

(d) Split of Training and Testing

We split the data into two parts: the first 2/3 was the training set and the last 1/3 was the testing set, as shown in Figure 3.2. The training set is from the 40th week in 2002 to the 52nd week in 2012, and the testing set is from the 1st week in 2013 to the 30th week in 2017.

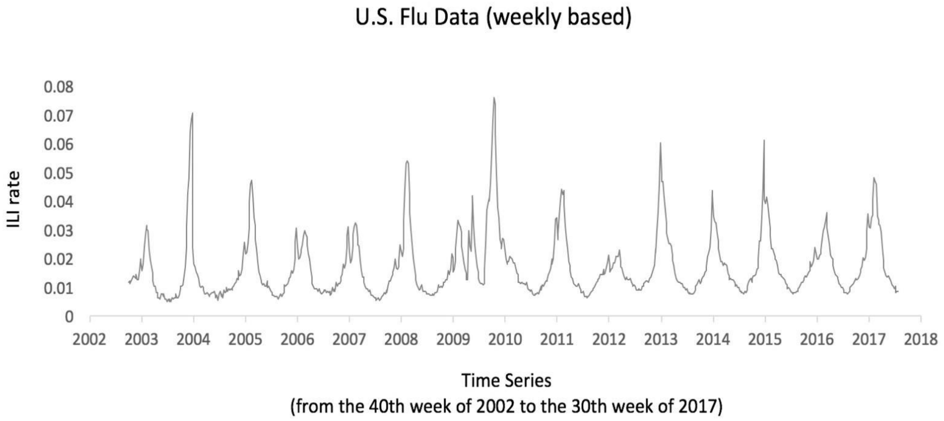


Figure 3.1: The U.S. flu season data.

The Y-axis represents the weekly ILI rate, and the X-axis represents the time series. The data were from the 40th week of 2002 to the 30th week of 2017.

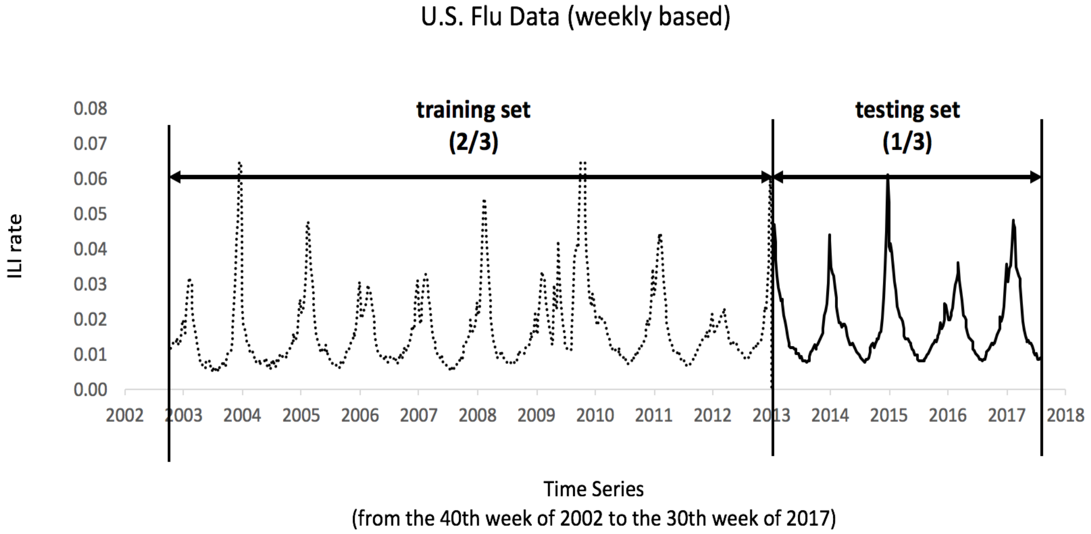


Figure 3.2: Split of training and testing set of U.S. flu data.

The dashed line is the first 2/3 used for the training set, and the solid line is the last 1/3 used for the testing set. The Y-axis represents the weekly ILI rate, and the X-axis represents the time series. The training set is from the 40th week in 2002 to the 52nd week in 2012, and the testing set is from the 1st week in 2013 to the 30th week in 2017.

(e) Feature Space

We have two types of features as feature space for the predicting models:

(a) the historical flu data by applying time lags; and

(b) first-order differences.

In time series predicting models, one usually predict the future values by building a model based on the past few values. We call the number of the past few values the "time lag". Different time lags are supposed to result in different levels of accuracy. For one thing, usually, the more time lag, the better the predicting accuracy will be if we have unlimited historical data. However, the predicting accuracy may not be improved furthermore after we increase the time lag to some extent. For another, the more time lag is, the fewer the training data will be. When we perform "time lag" (which sometimes we call "look back"), we have to sacrifice the first few data since their "time lag" data are "N/As". The fewer the training data, the worse the predicting accuracy will be. As a result, one usually needs to select the best number of time lag to bring the best accuracy. In other words, the selection of time lags can be essential to improve the predicting accuracy. However, all past studies simply adopted a time lag for models without comparing or selecting the appropriate number of time lags, which could make the model misunderstand past outbreak patterns. In this study, since flu seasonality is an annually recurring time period characterized by the prevalence of outbreaks of flu. Therefore, in this study, we reviewed a maximum of 52 weeks (approximately 1 year), we tried the time lags of 2 weeks (around half a month), 4 weeks (approximately 1 month), 9 weeks (approximately 2 months), 13 weeks (approximately 3 months), 26 weeks (around half a year), and 52 weeks (approximately 1 year) for model training and compared the results. We suppose 104 weeks hardly contribute to predicting accuracy since the second 52 weeks seem to be a simple repetition of the first 52 weeks.

In theory, the first order difference helps to predict model understand the ascending or descending trend of the time series data. In practice, some previous studies also found that first-order differences helped improve the results of the prediction models for flu data [166]. On the other side of the coin, adding more features in machine learning models or deep learning models barely decrease the predicting accuracy. Therefore, we also included the first-order differences as a part of the feature spaces. The formula 3.2 shows how to calculate the first order differences. In Formula 3.2, the Value$_{(t)}$ means the the value at the "t" moment; Value$_{(t-1)}$ means the the value at the "t-1" moment; Value$_{(t-2)}$ means the the value at the "t-2" moment; ... ; Value$_{(t-3)}$ means the the value at the "t-3" moment; ... ; Value$_{(t-52)}$ means the the value at the "t-52" moment; first_order_difference$_{(t-1)}$ means the first order difference at the "t-1" moment; first_order_difference$_{(t-2)}$ means the first order difference at the "t-2" moment; first_order_difference$_{(t-3)}$ means the first order difference at the "t-3" moment; ... ; first_order_difference$_{(t-52)}$ means the first order difference at the "t-52" moment.

$$\begin{aligned}
first\_order\_difference_{(t-1)} &= Value_{(t)} - Value_{(t-1)} \\
first\_order\_difference_{(t-2)} &= Value_{(t)} - Value_{(t-2)} \\
first\_order\_difference_{(t-3)} &= Value_{(t)} - Value_{(t-3)} \\
&\dots \\
first\_order\_difference_{(t-52)} &= Value_{(t)} - Value_{(t-52)}
\end{aligned} \tag{3.2}$$

In the case of the time lag of 52 weeks, we used (a) the ILI rate of the current week, (b) the ILI rates of the past 52 weeks, and (c) the 52 first-order differences. In total, we have 105 predictors (a + b + c) for use as feature spaces. Figure 3.3 illustrates the pretreatment of the source data when we look back 52 weeks. In the head of the table, since we were unable to look back and to calculate the first-order differences for the first 52 rows, we removed the first 52 rows. However, it doe not mean the first 52 rows are useless. The first 52 rows were indirectly used as historical features and for calculation of first-order differences of the flu data from the 53rd row to 104th row. In the tail of the table, and we could not have future data as a response (y) of the last row. In other words, the flu data of the last row can be only used as a response (y) of the last second row. Alike, in the case of the time lag of 2, 4, 9, 13, 26 weeks, we had 5, 9, 19, 27, 53 predictors and had to dropped the first 2, 4, 9, 13, 26 rows (the first 2, 4, 9, 13, 26 weeks) since we are unable to calculate the first-order differences for the first 2, 4, 9, 13, 26 rows (the first 2, 4, 9, 13, 26 weeks). We compared the predicting accuracy of the models of different time lags. However, the models with fewer time lags could have more training data, and more training data usually help improve predicting accuracy. Therefore, more training data was considered unfair. To fairly compare the predicting accuracy of adopting different time lags, we uniformly removed the first 52 rows (the first 52 weeks) from the training set of all the models.

3.1.2 Predictive Models

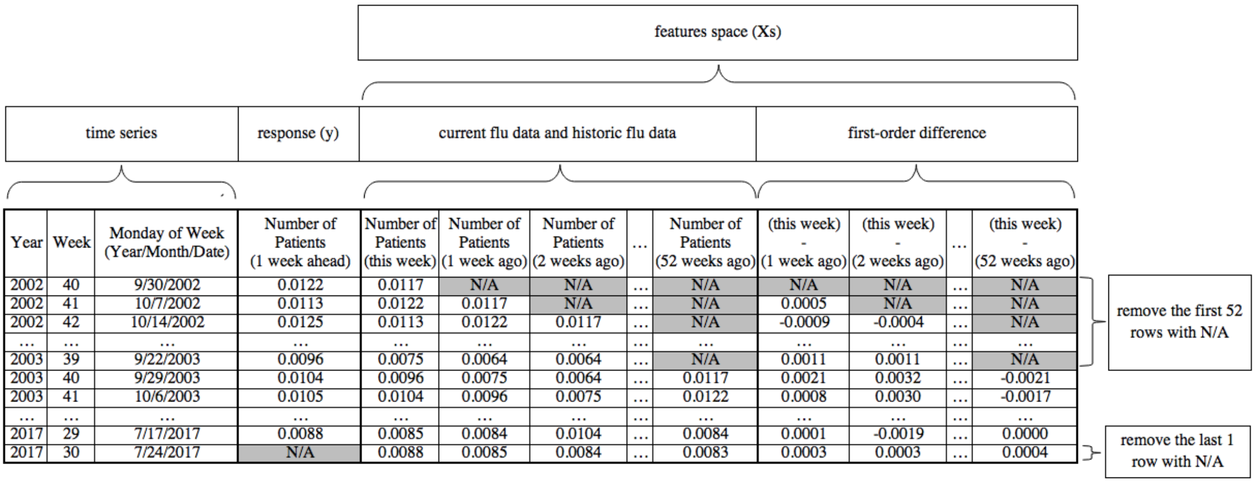
This section describes the models, including programming languages for models, the types of models, the structure of deep learning models, the hyperparameters of all the machine learning and deep learning models, and the analytical metrics that we used to compare predicting accuracy.

(a) Programming Languages

Python and R are the two languages we used for the first research step. We did not uniform programming languages because we prefer using R (version 3.4.1) for ARIMA and SVR sue to some special reasons.

(b) Structure of Models

As to our ARIMA models, the integrated (“I”) algorithm in ARIMA presents that the target values have been replaced with the difference between their values and the previous values, such as first-order difference, second-order difference, and so on. The integration is repeated until the processed data achieve a stable status. The popular library of ARIMA in R programming language (”Forecast” Package, Version 8.1) has a module called ”auto.arima”, which helps stabilize data automatically while the ARIMA library in Python needs a manual process to achieve a stable status before modeling. Regarding our SVR models, according to our experience, the library of SVR in Python usually takes too much time to train an SVR model. We applied the ”Caret” Package (Version 6.0-76) to decrease training time and used cross-validation algorithm to stabilize model and improve predicting accuracy. For our RF models, we used the Scikit Learn Package (Version 0.18.1) in Python (Version 3.6.0). We applied the algorithm of the grid search and the cross-validation to find the best combination of the number of trees (i.e. the hyperparameter of ”n.estimator”), the number of features selected in every single tree(i.e. the hyperparameter of ”max_features”), and the depth of every single tree (i.e. the hyperparameter of ”max_depth”). For our GB models, we also used the



features space (Xs)
time series
response (y)
current flu data and historic flu data
first-order difference
remove the first 52 rows with N/A
remove the last 1 row with N/A

Year	Week	Monday of Week (Year/Month/Date)	Number of Patients (1 week ahead)	Number of Patients (this week)	Number of Patients (1 week ago)	Number of Patients (2 weeks ago)	…	Number of Patients (52 weeks ago)	(this week) - (1 week ago)	(this week) - (2 weeks ago)	…	(this week) - (52 weeks ago)
2002	40	9/30/2002	0.0122	0.0117	N/A	N/A	…	N/A	N/A	N/A	…	N/A
2002	41	10/7/2002	0.0113	0.0122	0.0117	N/A	…	N/A	0.0005	N/A	…	N/A
2002	42	10/14/2002	0.0125	0.0113	0.0122	0.0117	…	N/A	-0.0009	-0.0004	…	N/A
…	…	…	…	…	…	…	…	…	…	…	…	…
2003	39	9/22/2003	0.0096	0.0075	0.0064	0.0064	…	N/A	0.0011	0.0011	…	N/A
2003	40	9/29/2003	0.0104	0.0096	0.0075	0.0064	…	0.0117	0.0021	0.0032	…	-0.0021
2003	41	10/6/2003	0.0105	0.0104	0.0096	0.0075	…	0.0122	0.0008	0.0030	…	-0.0017
…	…	…	…	…	…	…	…	…	…	…	…	…
2017	29	7/17/2017	0.0088	0.0085	0.0084	0.0104	…	0.0084	0.0001	-0.0019	…	0.0000
2017	30	7/24/2017	N/A	0.0088	0.0085	0.0084	…	0.0083	0.0003	0.0003	…	0.0004

Figure 3.3: Response and feature space in the first research

This table shows the response and feature space in the case of the time lag of 52 weeks. The table can be divided into 4 parts by columns. The four parts are (1) time series, (2) response, (3) current and historical data, and (4) first order difference. In the part of the ”(1) time series”, the first and second columns are the sequence of years and weeks. The third column is the dates of the Mondays of the weeks. In the part of the ”(2) response”, the column is the ILI rate of the one week ahead, which we used as a response (y) in the single step prediction. In the tail of ”(2) response”, we have an N/A in the last row. That is because we are predicting the ILI rate of one week ahead. For the last row, we could not have the ILI rate of the one week ahead of the last row. In other words, the last row lacks future data as ground truth (i.e. response) to predict. Therefore, we removed the last row. Nevertheless, it does not mean the last row is useless. We used the flu data, i.e. ”ILI Rates (this week”), of the last row as the response (y) of the last second row. The parts of ”(3) current and historical data” and ”(4) first-order difference” include all the feature space (Xs). In the head the part of ”(3) current and historical data” and ”(4) first-order difference”, since we look back 52 weeks, we were unable to have flu data of the past 52 weeks and thereby calculate the first-order differences for the first 52 rows. As a result, we have 1, 2, 3, ..., and 52 N/As in the column of ”ILI Rates (1 week ago)”, ”ILI Rates (2 weeks ago)”, ”ILI Rates (3 weeks ago)”, ... and ”ILI Rates (52 weeks ago)”, respectively. Similarly, we have 1, 2, 3, ..., and 52 N/As in the column of ”(this week)-(1 week ago)”, ”(this week)-(2 weeks ago)”, ”(this week)-(3 week ago)”, ... and ”(this week)-(52 week ago)”, respectively. The flu data of first 52 rows of the column of ”ILI Rates (this week”) were used as historical features and for calculation of first-order differences of the flu data from the 53rd row to 104th row. We also removed the first 52 weeks from the training set to keep all the feature space complete with no N/As.

Table 3.1: The models, programming languages, libraries, and hyperparameter adjustments in the first research.

Models	Programming Languages	Programming Libraries	Hyperparameter Adjustment
ARIMA	R (Version 3.4.1)	Forecast (Version 8.1)	# auto.arima
SVR	R (Version 3.4.1)	Caret (Version 6.0-76)	# cross validation
RF	Python (Version 3.6.0)	Scikit Learn (Version 0.18.1)	# cross validation # grid search # n_estimators # max_features # max_depth
GB	Python (Version 3.6.0)	Scikit Learn (Version 0.18.1)	# cross validation # grid search # learning rate # subsample # n_estimators # max_features # max_depth
ANN	Python (Version 3.6.0)	Keras (Version 2.0.4) Tensorflow (Version 1.1.0)	# different layers (up to 5 layers) # with/without dropout # with/without regularization # with/without batch normalization
LSTM	Python (Version 3.6.0)	Keras (Version 2.0.4) Tensorflow (Version 1.1.0)	# different layers (up to 10 layers) # with/without dropout # with/without regularization # with/without batch normalization

Scikit Learn Package (Version 0.18.1) in Python (Version 3.6.0). We applied the algorithm of the grid search and the cross-validation to find the best combination of the learning rate, the number of instance used for every single tree (i.e. the hyperparameter of ”subsample”), the number of trees (i.e. the hyperparameter of ”n.estimator”), the number of features selected in every single tree(i.e. the hyperparameter of ”max_features”), and the depth of every single tree (i.e. the hyperparameter of ”max_depth”). As to our ANN models, we used Python (Version 3.6.0) and the Keras package (Version 2.0.4) based on Tensorflow (Version 1.1.0). We tried 3-layer (input layer + fully connected layer + output layer), 4-layer (input layer + fully connected layer × 2 + output layer), and 5-layer (input layer + fully connected layer × 3 + output layer) to compare the results. We also added regularization or the pair of dropout and batch normalization or nothing to each layer to compare the results. we adopted an “early-stopping” algorithm with a “patience” of 100 epochs (for a total of 1000 epochs). Regarding our LSTM models, we also used Python (Version 3.6.0) and the Keras package (Version 2.0.4) based on Tensorflow (Version 1.1.0). We tried 3-layer (input layer + LSTM layer + output layer), 4-layer (input layer + LSTM layer × 2 + output layer), 5-layer (input layer + LSTM layer × 3 + output layer), 6-layer (input layer + LSTM layer × 4 + output layer), and 10-layer (input layer + LSTM layer × 8 + output layer) to compare the results. We also added regularization or the pair of dropout and batch normalization or nothing to each layer to compare the results. We adopted an “early-stopping” algorithm with a “patience” of 100 epochs (for a total of 1000 epochs). Table 3.1 illustrates the predicting models, programming languages, libraries (i.e. packages), and hyperparameter adjustments we used in this study.

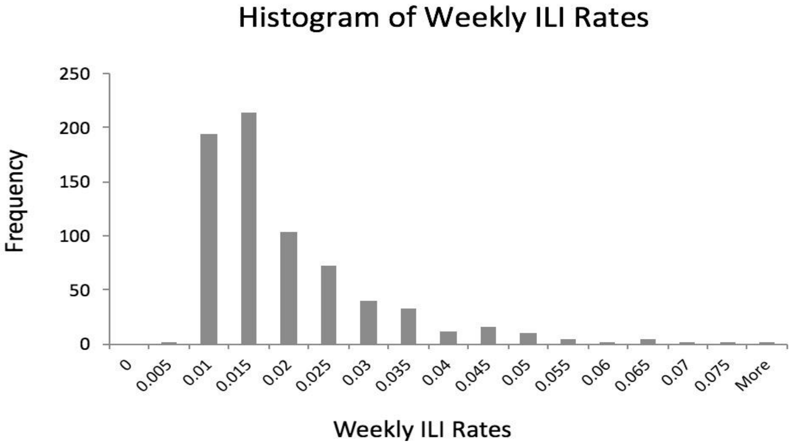


Figure 3.4: The histogram of weekly ILI Rates from U.S. flu data.

The histogram is right skewed. The distribution is a non-normal distribution.

(c) Metrics

One usually use metrics to compare the performance of predicting models. We compared different models and different time lags using the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) as Key Performance Indicators (KPIs).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|\frac{F_t - A_t}{A_t}| \times 100\% \tag{3.3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(F_t - A_t)^2} \tag{3.4}$$

Predicting models usually use mean square error (MSE) or root mean square error (RMSE) as metrics, since predicting models are regression models. However, we prefer to mean absolute percentage error (MAPE) in the first research step. The Formula 3.3 and 3.4 illustrate the calculation of MAPE and RMSE, respectively. Comparing models by MSE or RMSE mainly reflects the difference of mean since more values appear around mean. Comparing models by MAPEs mainly reflects the difference of the median since more values appear around the median. If the source data follows Gaussian Distribution, the mean value is equal to the median. As a result, comparing by MSE / RMSE and MAPE have the same effectiveness. However, when the source data does not follow Gaussian Distribution, comparing by MAPEs reflects the performance better than comparing by mean. Figure 3.4 illustrates the histogram of the weekly ILI rates of the U.S. flu data. The histogram is right-skewed. Besides, when we examine statistical test, the result (p-value < 0.001) of the Kolmogorov–Smirnov Test shows the distribution is a non-normal distribution. Therefore, we consider taking MAPE rather than RMSE as metrics. In practice, in the first research step, we regard the MAPE as the first KPI and the RMSE as an assistant KPI.

3.2 Methods of Comparative Study on Algorithms of Multistep Prediction

This is the second subsection of this chapter. In this subsection, we describe all the methods we used for the second research. Some methods are inherited from the first research since the result of the first research found they were effective, such as the type of the model, and so on. Different from the first research, the second research focused on the multistep prediction, which also prepares for the third research.

3.2.1 Experiment Data

This part explains the data and the preprocess of the source data for the second research.

(a) Data Source

In the second search step, we collected the same U.S. flu data with the first research step from the ”FluView” Portal of the CDC.

(b) Tackling N/As

Due to the same reasons that we proposed in the part of ”Tackling N/As” in the first research step, we dropped the periods with N/As used the flu data from the 40th week of 2002 to the 30th week of 2017.

(c) Response(y)

Due to the same reasons that we proposed in the part of ”Response(y)” in the first research step, we also adopted the ILI rates (Formula 3.1) as the response (y) of the predicting models. In the second research, we performed a multistep prediction. We forecast the 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, and 13-step-ahead ILI rates. When we train or test the model, we need ground truth as a response (y). Figure 3.3 illustrates the pretreatment of the source data when we look back 52 weeks and look forward 1 week. In Figure 3.3, we removed the last row since the last row does not have response (y) when performing the single step prediction. Alike, we need to remove the last 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 rows when we predicted the 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, and 13-step-ahead ILI rates, respectively.

(d) Split of Training and Testing

We used the first 2/3 of the data for the training set and the last 1/3 of the flu data for the testing set. In detail, the training set is from the 40th week in 2002 to the 52nd week in 2012, and the testing set is from the 1st week in 2013 to the 30th week in 2017. (Figure 3.2) The split was the same as the one in the first research.

(e) Feature Space

In the result part of the first research, we compared the predicting accuracy of the models with the time lags of 2 weeks (around half a month), 4 weeks (approximately 1 month), 9 weeks (approximately 2 months), 13 weeks (approximately 3 months), 26 weeks (around half a year), and 52 weeks (approximately 1 year). We found, in LSTM models, the time lag of 52 weeks brought the best predicting accuracy. Therefore, in the second research, we took the experience from the first research and only adopted 52 weeks as the time lag. As feature spaces, we used the ILI rate of the current week, the ILI rates of the past 52 weeks, and the 52 first-order differences. Totally, we have 105 features, which is also the same as the number of features in the first research.

3.2.2 Predictive Models

This section describes the details of the models, such as programming languages, the structure, the hyperparameters, the metrics that we used in the second research.

(a) Programming Languages

Since we only used the LSTM structure in the second research, (we talked about the reason why we only used the LSTM structure in the coming part, i.e. ”Structure of Models”), we just used Python (Version 3.6.0).

(b) Structure of Models

In the second research, we leveraged the LSTM. Our selection of LSTM was based on theoretical and practical consideration. In theory, LSTM is a special kind of RNN. Its elaborate structure (multilayers and gated cells) enables LSTM to learn simulate nonlinear function, long-term dependencies [137], and refine time-series prediction [140]. In practice, LSTM achieved the best accuracy in all the six models when we performed a single-step prediction for the same U.S flu data in the first research [167]. We adopted an “early-stopping” algorithm with a “patience” of 100 epochs (for a total of 1000 epochs) and compared the accuracy. We tried 3-layer (input layer + LSTM layer + output layer), 4-layer (input layer + LSTM layer × 2 + output layer), 5-layer (input layer + LSTM layer × 3 + output layer), 6-layer (input layer + LSTM layer × 4 + output layer), and 10-layer (input layer + LSTM layer × 8 + output layer).

(c) Metrics

Due to the same reasons that we proposed in the part of ”Metrics” in the first research step, we also adopted MAPE as the metrics in the second research.

3.2.3 Multistep Prediction Algorithms

In this part, we discuss the algorithms on multistep prediction. There are mainly two types of methods to perform multistep prediction. We can name them (1) “recursive” prediction and (2) “jumping” prediction. Generally, the method of “recursive” predicts step-by-step: using predicted value to predict further values. For example, when performing the two-step-ahead prediction, ”recursive” method firstly predict 1-week-ahead value (i.e. a single step prediction) and then uses the predicted 1-week-ahead value to predict the 2-week-ahead value, and when predicting the 3-week-ahead value, the model will also use the predicted 1-week-ahead value and 2-week-ahead value. By recursive predicting, the models predict some-step-ahead values.

(a) Multi-Stage Prediction (MSP)

The Multi-Stage Prediction (MSP) is a classic “recursive” prediction. MSP usually can use any regression model to train a single-step predictive model. To achieve a multiple-step prediction, MSP applied this single-step predictive model recursively by feeding its previous output [168].

$$\begin{aligned}
X_{t+1}(pred) &= MSP\_MODEL\_1[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)] \\
X_{t+2}(pred) &= MSP\_MODEL\_1[X_{t+1}(obs), X_t(obs), X_{t-1}(obs), \ldots, X_{t-51}(obs)] \\
X_{t+3}(pred) &= MSP\_MODEL\_1[X_{t+2}(obs), X_{t+1}(obs), X_t(obs), \ldots, X_{t-50}(obs)] \\
&\ldots \\
X_{t+13}(pred) &= MSP\_MODEL\_1[X_{t+12}(obs), X_{t+11}(obs), X_{t+10}(obs), \ldots, X_{t-40}(obs)]
\end{aligned} \tag{3.5}$$

where ”X” means values at different time steps; ”pred” means prediction, and ”obs” means observation. Formula 3.5 illustrates the algorithm of MSP. Take a 13-step-ahead prediction by the values of the past 52 weeks (just like the second research step) as an example. Firstly, the MSP model uses the values of the current week (denoted as X_t) and the past 52 weeks (denoted as X_{t-1}, X_{t-2}, … and X_{t-52}) to predict 1-week-ahead value (denoted as X_{t+1}). Then, MSP uses the predicted value (X_{t+1}) and the value of the current week (X_t) and the values of the past 51 weeks (X_{t-1}, X_{t-2}, … and X_{t-51}) to predict 2-week-ahead value (denoted as X_{t+2}). Alike, Then, MSP uses the predicted value (X_{t+12}, X_{t+11}, X_{t+10}, … , and X_{t+1}) and the value of the current week (X_t) and the values of the past 40 weeks (X_{t-1}, X_{t-2}, … and X_{t-40}) to predict 13-week-ahead value (denoted as X_{t+13}). Nonetheless, when MSP performs the 1-week-ahead prediction, the predicted value has, more or less, predicting error. Feeding the predicted values with error into models will negatively pull down the accuracy of the further-step-ahead prediction. Just like an avalanche, MSP accumulates increasing predicting errors as MSP performs further-step-ahead prediction.

(b) Adjusted Multi-Stage Prediction (AMSP)

To limit the accumulation of predicting errors, by some methods, we can adjust further-step-ahead prediction. Adjusted Multi-Stage Prediction (AMSP) is an adjusted version of MSP. AMSP assume the predicting errors from the previous steps follow some distribution that can be learned by regression models. Therefore, take the 2-week-ahead prediction as an example, AMSP trained another model instead of applying the same model that is used to perform the 1-week-ahead prediction. Sometimes, we call the new model the 2-week-ahead model. Actually, the 2-week-ahead model adjusts the error produced from the 1-week-ahead model and performs the 2-week-ahead prediction. When AMSP preforms 3-week-ahead prediction, it trains and applies a 3-week-ahead prediction. Such a modification helps suppress error accumulation [76, 169]. Formula 3.6 illustrates the algorithm of AMSP. Actually, the formulas are quite similar to those of MSP, except for the number of models trained and applied.

$$
\begin{aligned}
X_{t+1}(pred) &= AMSP\_MODEL\_1[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)] \\
X_{t+2}(pred) &= AMSP\_MODEL\_2[X_{t+1}(obs), X_t(obs), X_{t-1}(obs), \ldots, X_{t-51}(obs)] \\
X_{t+3}(pred) &= AMSP\_MODEL\_3[X_{t+2}(obs), X_{t+1}(obs), X_t(obs), \ldots, X_{t-50}(obs)] \\
&\ldots \\
X_{t+13}(pred) &= AMSP\_MODEL\_13[X_{t+12}(obs), X_{t+11}(obs), X_{t+10}(obs), \ldots, X_{t-40}(obs)]
\end{aligned}
\tag{3.6}
$$

where "X" means values at different time steps; "pred" means prediction, and "obs" means observation.

(c) Multiple Single-Output Prediction (MSOP)

Both MSP and AMSP are "recursive" prediction. The following two are "jumping" prediction. The method of "jumping" prediction only uses current and past values to predict some-step-ahead value directly instead of predicting step-by-step. Multiple Single-Output Prediction (MSOP) is a typical "jumping" prediction. Formula 3.7 shows the algorithms of MSOP. Take a 13-step-ahead prediction by the values of the past 52 weeks as an example. Firstly, the MSOP model uses the values of the current week (denoted as X$_t$) and the past 52 weeks (denoted as X$_{t-1}$, X$_{t-2}$, ... and X$_{t-52}$) to predict 1-week-ahead value (denoted as X$_{t+1}$). Then, the MSOP model also uses the values of the current week (denoted as X$_t$) and the past 52 weeks (denoted as X$_{t-1}$, X$_{t-2}$, ... and X$_{t-52}$) to predict 1-week-ahead value (denoted as X$_{t+2}$). Alike, the MSOP model also uses the values of the current week (denoted as X$_t$) and the past 52 weeks (denoted as X$_{t-1}$, X$_{t-2}$, ... and X$_{t-52}$) to predict 1-week-ahead value (denoted as X$_{t+13}$). One can easily find that, when performing 13-week-ahead prediction, MSOP never cares about any possible changes in 1-, 2-, ..., or 12-week-ahead values. MSOP just "jumps" prediction of 1-, 2-, ..., or 12-week-ahead values and directly performs 13-week-ahead prediction.

$$
\begin{aligned}
X_{t+1}(pred) &= MSOP\_MODEL\_1[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)] \\
X_{t+2}(pred) &= MSOP\_MODEL\_2[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)] \\
X_{t+3}(pred) &= MSOP\_MODEL\_3[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)] \\
&\ldots \\
X_{t+13}(pred) &= MSOP\_MODEL\_13[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)]
\end{aligned}
\tag{3.7}
$$

where "X" means values at different time steps; "pred" means prediction, and "obs" means observation.

(d) Multiple-Output Prediction (MOP)

Multiple-Output Prediction (MOP) can be regarded as a merged version of MSOP. MOP takes advantage of some models and uses only one model to predict all-step-ahead values. Some models can produce several outputs. This characteristic helps to merge all models of predicting all-step-ahead values into one model. LSTM models are typical models that can be used for MOP. Others include multiple SVR, which was leveraged in some previous researches [75, 168, 170]. Formula 3.8 outlines the algorithm of MOP.

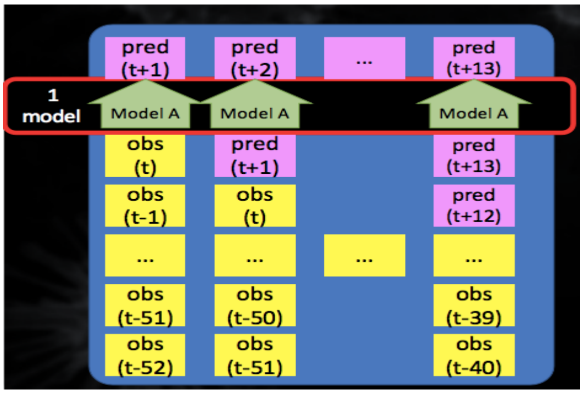


Figure 3.5: Algorithms of Multi-Stage Prediction (MSP)

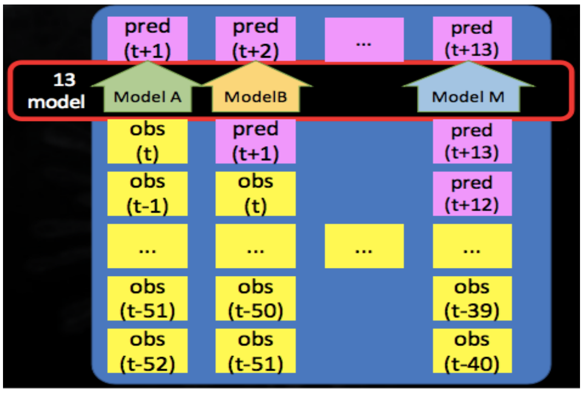


Figure 3.6: Algorithms of Adjusted Multi-Stage Prediction (AMSP)

$$\begin{gathered}X_{t+1}(pred), X_{t+2}(pred), \ldots, X_{t+13}(pred) = \\ LSTM\_MOP\_1[X_t(obs), X_{t-1}(obs), X_{t-2}(obs), \ldots, X_{t-52}(obs)]\end{gathered} \tag{3.8}$$

where ”X” means values at different time steps; ”pred” means prediction, and ”obs” means observation.

Comparison of Algorithms of MSP, AMSP, MSOP, and MOP

Figure 3.5, 3.6, 3.7, and 3.8 compare the difference among algorithms of MSP, AMSP, MSOP, and MOP. MSP and AMSP have same feature space but different model training. MSOP and MOP have also the same feature space but the different number of models. AMSP and MSOP have the same number of models but different feature space. MSP and MOP also have the same number of models but different feature space. In theory, AMSP should bring better accuracy than MSP does since AMSP has an adjustment. MSOP should produce better accuracy than MOP does since MSOP does not need to share ”weights” in the neural network while MOP needs. MSOP should bring better accuracy than AMSP does since MSOP solve the problem of error accumulation while AMSP only adjusts or limits the problem of error accumulation.

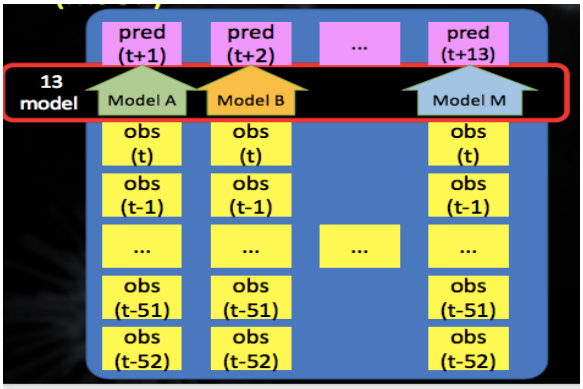


Figure 3.7: Algorithms of Multiple Single-Output Prediction (MSOP)

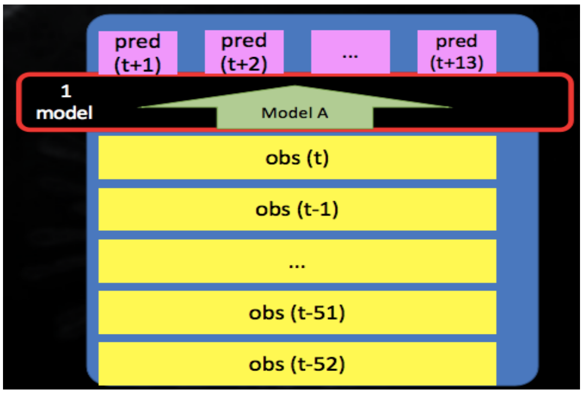


Figure 3.8: Algorithms of Multiple-Output Prediction (MOP)

3.3 Methods of Multistep Spatio-Temporal Prediction of World-wide Flu Outbreaks

We talked about the single step and multistep prediction of flu in only one country. Since flu is an infectious disease, flu spread in all countries is related to each other around the world. Global prediction is needed when considering the geolocational factors of flu infection. In the third research step, we study global prediction. Although the study is more on geolocational prediction, we still inherited some methods that we explored in the previous two pieces of research, such as the number of time lags, the type of the model, and so on. This part describes all the methods we used for the geolocational-temporal multistep prediction of flu around the world.

3.3.1 Experiment Data

This part describes the source data and preprocessing of the source data in the geolocational-temporal multistep prediction of flu.

(a) Data Source

The “FluView” Portal of Centers for Disease Control and Prevention only provides flu data of U.S. For the geolocational-temporal multistep prediction, we changed the data source. FluNet is a global web-based tool for flu virologic surveillance. The data at the country level are available and updated weekly. We scraped flu data of all the 152 countries from the FluNet [171]. The 152 countries were Afghanistan, Albania, Algeria, Angola, Anguilla, Argentina, Armenia, Aruba, Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Bermuda, Bhutan, Bolivia (Plurinational State of), Bosnia and Herzegovina, Brazil, Bulgaria, Burkina Faso, Cambodia, Cameroon, Canada, Cayman Islands, Central African Republic, Chile, China, Colombia, Congo, Costa Rica, Croatia, Cuba, Czechia, Democratic Republic of the Congo, Denmark, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Ethiopia, Fiji, Finland, France, French Guiana, Georgia, Germany, Ghana, Greece, Guadeloupe, Guatemala, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran (Islamic Republic of), Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kyrgyzstan, Latvia, Lebanon, Lithuania, Luxembourg, Madagascar, Malaysia, Maldives, Mali, Malta, Martinique, Mauritania, Mauritius, Mexico, Mongolia, Montenegro, Montserrat, Morocco, Mozambique, Myanmar, Nepal, Netherlands, New Caledonia, New Zealand, Nicaragua, Niger, Nigeria, Norway, Oman, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Republic of Korea, Republic of Moldova, Romania, Russian Federation, Rwanda, Saint Barthelemy, Saint Kitts and Nevis, Saint Lucia, Saint Martin, Saint Vincent and the Grenadines, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, Spain, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Syrian Arab Republic, Thailand, The former Yugoslav Republic of Macedonia, Togo, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania, United States of America, Uruguay, Uzbekistan, Venezuela (Bolivarian Republic of), Viet Nam, and Zambia.

(b) Tackling N/As

In the method of the first research, we discussed how to tackle N/As. We concluded that to keep data's originality, we gave up any rendering method and just selected data without N/As. Similarly, in the third research, in flu data of all the 152 countries, there were only 22 countries that have no N/As from the 1st week of 2009 to the 18th week of 2018. Therefore, we selected these 22 countries as features space. The 22 countries are Australia (AUS), Brazil (BRA), Cambodia , China (CHN), Egypt, French Guiana, Ghana, Indonesia, Iran, Iraq, Ireland, Japan (JPN), Netherlands, Nicaragua, Niger, Norway, Panama, Poland, Republic of Korea, Russia, United Kingdom of Great Britain and Northern Ireland (UK), US.

(c) Response(y)

Different from the previous two pieces of research, we used the raw flu data as the response in the third research since we could hardly find and scrape the number of the weekly patients to calculate ILI rates. We did not predict the flu data of all the 22 countries. We only selected AUS, BRA, CHN, JPN, UK, and USA to perform prediction. These countries have a large number of flu patients every week, which is partly because these countries have a large population. Countries with small populations, such as French Guiana, Niger, and so on., might have only a few flu patients weekly, and in most weeks, the number of flu patients in countries with small populations were almost zeros. The fact that the number of flu patients was almost zeros makes the calculation of metrics (i.e. MAPE) too problematic. The metrics (i.e. MAPE) could be extremely high since the denominator (the number of flu patients) is too small, and the updates of weights in neural nets (also called backpropagation) relies on the metrics (MAPE) produced by the previous forward propagation. The extremely high MAPEs at and near data points with values of zeros makes the backpropagation emphasize the importance of the predicting accuracy at and near data points with values of zeros and therefore update weights in neural nets to better predict values at the data points with values of zeros. However, the target of flu prediction is to understand the trend, possible fluctuation, outbreaks, and so on. around peaks in flu seasons. Emphasizing the importance of the predictive accuracy near data points with values of zeros makes no sense. That is why we selected countries with great many flu patients to perform prediction.

(d) Split of Training and Testing

We split the data into two parts: the last 52 rows (around one year) were used for testing set and the other rows were used for the training set.

(e) Geolocational Features

To include geolocational factors, when predicting the flu data of one country, we feed the historical flu data of all the 22 countries into models.

(f) Temporal Features

Usually, flu infection has an obvious cycle. In temperate climates, flu outbreaks occur mainly during winter; while in the tropical regions, flu outbreaks occur throughout the year. Regarding the time lag, taking the experience of the first research, we adopted 52 weeks as the time lag and used the historical flu data of the past 52 weeks. Regarding the time difference, we took the second order difference in addition to the first order difference. Moreover, in the third research, we extended the feature space by introducing a series of new feature processing, i.e. rolling windows. We adopted the rolling windows of 1, 2, 3, 4, 9, 13, 26, 52 weeks. In every rolling window, we took the mean, median, standard deviation, maximum, and minimum.

3.3.2 Predictive Models

This section describes the details of the models in the third research.

(a) Programming Languages

We just used Python (Version 3.6.0) to scrape flu data and train models.

(b) Structure of Models

Alike to the second research, we selected the LSTM based on theoretical and practical consideration. In theory, the elaborate structure of LSTM helps learn simulate nonlinear function, long-term dependencies [137], and refine time-series prediction [140]. In practice, we found LSTM achieved the best accuracy in all the six models in the first research [167]. We also adopted an “early-stopping” algorithm with a “patience” of 100 epochs (for a total of 1000 epochs) and compared the accuracy. Based on the experience of the second research, We put three layers of LSTM after input layers, and we added three fully connected layers after LSTMs.

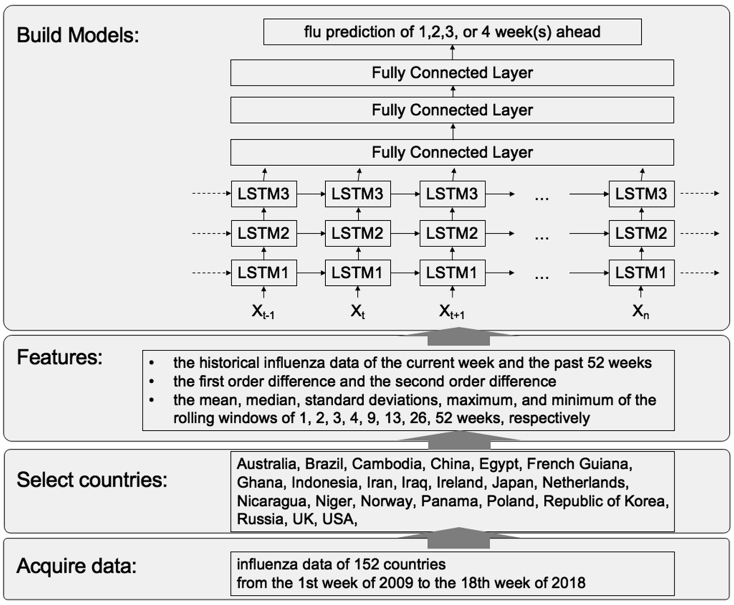

Build Models:
flu prediction of 1,2,3, or 4 week(s) ahead
Fully Connected Layer
Fully Connected Layer
Fully Connected Layer
LSTM3
LSTM2
LSTM1
...
X_{t-1}
X_t
X_{t+1}
X_n
Features:
• the historical influenza data of the current week and the past 52 weeks
• the first order difference and the second order difference
• the mean, median, standard deviations, maximum, and minimum of the rolling windows of 1, 2, 3, 4, 9, 13, 26, 52 weeks, respectively
Select countries:
Australia, Brazil, Cambodia, China, Egypt, French Guiana, Ghana, Indonesia, Iran, Iraq, Ireland, Japan, Netherlands, Nicaragua, Niger, Norway, Panama, Poland, Republic of Korea, Russia, UK, USA,
Acquire data:
influenza data of 152 countries
from the 1st week of 2009 to the 18th week of 2018

Figure 3.9: The flow chart of the spatio-temporal system.

(c) Metrics

Due to the same reasons that we proposed in the part of ”Metrics” in the first research step, we also adopted MAPE as the metrics in the third research.

3.3.3 Multistep Prediction

As we found in the second research, MSOP achieved the best accuracy when performing multistep prediction. In the third research, we adopt the MSOP.

3.3.4 Data Process Flow

Figure 3.9 illustrates the whole process of the data flow. In brief, firstly, We scraped flu data from the FluNet. We selected 22 countries as features. Then we extracted geolocational-temporal factors based on historical data and rolling windows. We conducted MSOP in LSTM. That was we fed the extracted features into four LSTM model combined with 3 LSTM layers and 3 fully connected layers. Finally, the models output flu data of the 1st, 2nd, 3rd, and 4th week ahead, respectively.

3.3.5 Models as Baselines

To compare the predicting accuracy, we performed two types of baselines. First, we applied the models of RF and SVR to present the predictive difference in model types. Second, we still used LSTM but excluded the flu data of the other 22 countries when predicting one country’s flu data, to present the difference between including and excluding geolocational features.

Chapter 4

Results

This section illustrates the results of the three research steps. We compared different models or different structures by comparing predicting accuracy. We used many tables and figures to present differences.

4.1 Results of Comparative Study on Models

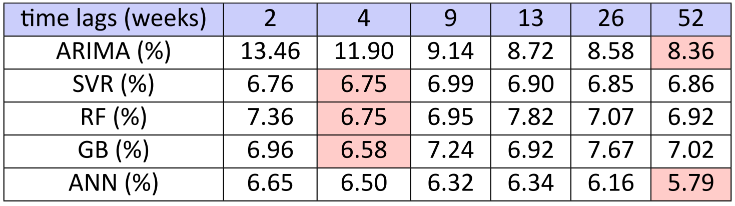
This section shows the predictive accuracy (MAPEs and RMSE) of the first research. As we aforementioned in the method of the first research, we used MAPEs as the primary metrics. Here, we also presented RMSE as the secondary metrics.

4.1.1 Predictive Accuracy of Models of ARIMA, SVR, RF, GB, and ANN

Table 4.1 shows the MAPEs of ARIMA, SVR, RF, GB, ANN in the testing set. The MAPEs of ARIMA with a time lag of 2, 4, 9, 13, 26, 52 was 13.46%, 11.90%, 9.14%, 8.72%, 8.58%, and 8.36%, respectively. The MAPEs of SVR with a time lag of 2, 4, 9, 13, 26, 52 was 6.76%, 6.75%, 6.99%, 6.90%, 6.85%, 6.86%, respectively. The MAPEs of RF with a time lag of 2, 4, 9, 13, 26, 52 was 7.36%, 6.75%, 6.95%, 7.82%, 7.07%, 6.92%, respectively. The MAPEs of GB with a time lag of 2, 4, 9, 13, 26, 52 was 6.96%, 6.58%, 7.24%, 6.92%, 7.67%, 7.02%, respectively. The MAPEs of ANN with a time lag of 2, 4, 9, 13, 26, 52 was 6.65%, 6.50%, 6.32%, 6.34%, 6.16%, 5.79%, respectively. Table 4.2 shows the RMSEs of models of ARIMA, SVR, RF, GB, ANN in the testing set. The RMSEs of ARIMA with a time lag of 2, 4, 9, 13, 26, 52 was 0.00444, 0.00410, 0.00367, 0.00328, 0.00343, 0.00364, respectively. The RMSEs of SVR with a time lag of 2, 4, 9, 13, 26, 52 was 0.00256, 0.00255, 0.00254, 0.00253, 0.00251, 0.00227, respectively. The RMSEs of RF with a time lag of 2, 4, 9, 13, 26, 52 was 0.00242, 0.00252, 0.00258, 0.00269, 0.00255, 0.00259, respectively. The RMSEs of GB with a time lag of 2, 4, 9, 13, 26, 52 was 0.00238, 0.00235, 0.00265, 0.00259, 0.00273, 0.00251, respectively. The RMSEs of ANN with a time lag of 2, 4, 9, 13, 26, 52 was 0.00259, 0.00255, 0.00257, 0.00255, 0.00252, 0.00241, respectively.

In ARIMA models, when we increased the time lag, we found an obvious decrease in MAPEs. We achieved the lowest MAPE (8.36%) when we used the time lag of 52 weeks. In SVR models, when we increased the time lag, we could not find an obvious decrease in MAPEs. We achieved the lowest MAPE (6.75%) when we used the time lag of 4 weeks. In RF models, when we increased the time lag, we could not find an obvious decrease in MAPEs. We also achieved the lowest MAPE (6.75%) when we used the time lag of 4 weeks. In GB models, when we increased the time lag, we could not find an obvious decrease in MAPEs. We achieved the lowest MAPE (6.58%) when we used the time lag of 4 weeks. In ANN models, when we increased the time lag, we also found a decrease in MAPEs. We achieved the lowest MAPE (5.79%) when we used the time lag of 52 weeks. All the lowest values in each type of models are highlighted in Tables 4.1 and 4.2 Figure 4.1 presents the actual and predicted values in the testing set (from the 1st week of 2013 to the 30th week of 2017) in the best ARIMA model (time lag was 52 weeks; MAPE was 8.36%). Figure 4.2 presents the actual and predicted values in the testing set (from the 1st week of 2013 to the 30th week of 2017) in the best SVR model (time lag was 4 weeks; MAPE was 6.75%). Figure 4.3 presents the actual and predicted values in the testing set (from the 1st week of 2013 to the 30th week of 2017) in the best RF model (time lag was 4 weeks; MAPE was 6.75%). Figure 4.4 presents the actual and predicted values in the testing set (from the 1st week of 2013 to the 30th week of 2017) in the best GB model (time lag was 4 weeks; MAPE was 6.58%). Figure 4.5 presents the

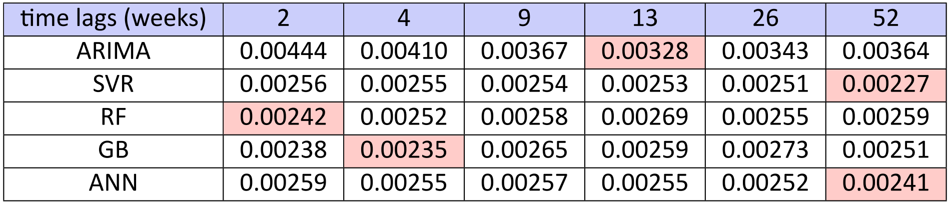
Table 4.1: The MAPEs of ARIMA, SVR, RF, GB, ANN in the testing set.



time lags (weeks)	2	4	9	13	26	52
ARIMA (%)	13.46	11.90	9.14	8.72	8.58	8.36
SVR (%)	6.76	6.75	6.99	6.90	6.85	6.86
RF (%)	7.36	6.75	6.95	7.82	7.07	6.92
GB (%)	6.96	6.58	7.24	6.92	7.67	7.02
ANN (%)	6.65	6.50	6.32	6.34	6.16	5.79

The highlighted cells represent the lowest MAPE in each type of models, respectively. ARIMA achieved the lowest MAPE (8.36%) when using the time lag of 52 weeks. SVR achieved the lowest MAPE (6.75%) when using the time lag of 4 weeks. RF achieved the lowest MAPE (6.75%) when using the time lag of 4 weeks. GB achieved the lowest MAPE (6.58%) when using the time lag of 4 weeks. ANN achieved the lowest MAPE (5.79%) when using the time lag of 52 weeks.

Table 4.2: The RMSEs of ARIMA, SVR, RF, GB, ANN in the testing set.



time lags (weeks)	2	4	9	13	26	52
ARIMA	0.00444	0.00410	0.00367	0.00328	0.00343	0.00364
SVR	0.00256	0.00255	0.00254	0.00253	0.00251	0.00227
RF	0.00242	0.00252	0.00258	0.00269	0.00255	0.00259
GB	0.00238	0.00235	0.00265	0.00259	0.00273	0.00251
ANN	0.00259	0.00255	0.00257	0.00255	0.00252	0.00241

The columns represent different time lags in weeks. The rows represent different models. The highlighted cells represent the lowest RMSEs in each type of models, respectively.

actual and predicted values in the testing set (from the 1st week of 2013 to the 30th week of 2017) in the best ANN model (time lag was 52 weeks; MAPE was 5.79%).

4.1.2 Predictive Accuracy of Models of LSTM

Table 4.3 shows the MAPEs of 3-layer LSTM, 4-layer LSTM, 4-layer LSTM with dropout, 4-layer LSTM with regularization, 5-layer LSTM, 5-layer LSTM with regularization, 6-layer LSTM with regularization, and 10-layer LSTM with regularization. The columns are different time lags in week. The rows represent the models of different number of layers with or without regularization or dropout. The last column is the average MAPE of LSTM models with same structure and hyperparameters but different time lags. The last row is the average MAPE of LSTM models with different structures and hyperparameters but same time lags. The highlighted cells was the lowest MAPE in each type of models. Table 4.4 shows the MAPEs of 3-layer LSTM, 4-layer LSTM, 4-layer LSTM with dropout, 4-layer LSTM with regularization, 5-layer LSTM, 5-layer LSTM with regularization, 6-layer LSTM with regularization, and 10-layer LSTM with regularization. In the model of 3-layer LSTM, the RMSE was 0.00253, 0.00254, 0.00250, 0.00249, 0.00241, 0.00210, respectively. In the model of 4-layer LSTM, the RMSE was 0.00261, 0.00258, 0.00257, 0.00257, 0.00252, 0.00243, respectively. In the model of 4-layer LSTM with dropout, the RMSE was 0.00253, 0.00252, 0.00252, 0.00250, 0.00246, and 0.00235, respectively. In the model of 4-layer LSTM with regularization, the RMSE was 0.00262, 0.00256, 0.00251, 0.00250, 0.00263, and 0.00256, respectively. In the model of 5-layer LSTM, the RMSE was 0.00250, 0.00249, 0.00246, 0.00243, 0.00241, and 0.00210, respectively. In the model of 5-layer LSTM with regularization, the RMSE was 0.00266, 0.00255, 0.00256, 0.00244, 0.00259, and 0.00257, respectively. In the model of 6-layer LSTM with regularization, the RMSE was 0.00259, 0.00256, 0.00250, 0.00248, 0.00252, and 0.00244, respectively.

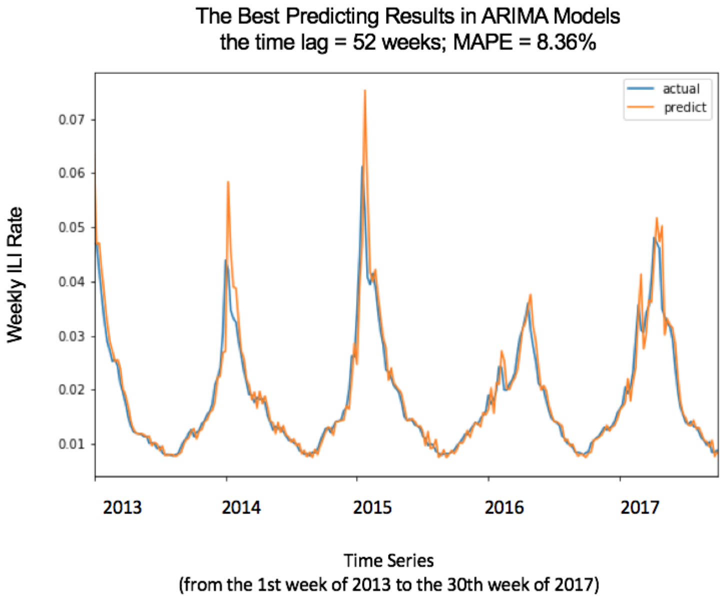


Figure 4.1: The best predicting results of the ARIMA models.

When adopting the time lag of 52 weeks, the ARIMA model achieved its best results. The X-axis represents time, and the Y-axis represents the U.S. weekly ILI rates.

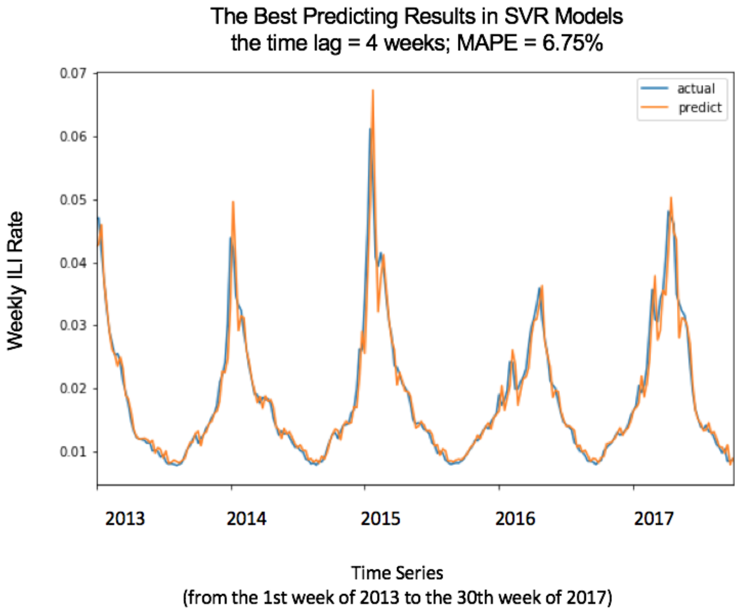


Figure 4.2: The best predicting results of the SVR models.

When adopting the time lag of 4 weeks, the SVR model achieved its best results. The X-axis represents time, and the Y-axis represents the U.S. weekly ILI rates.

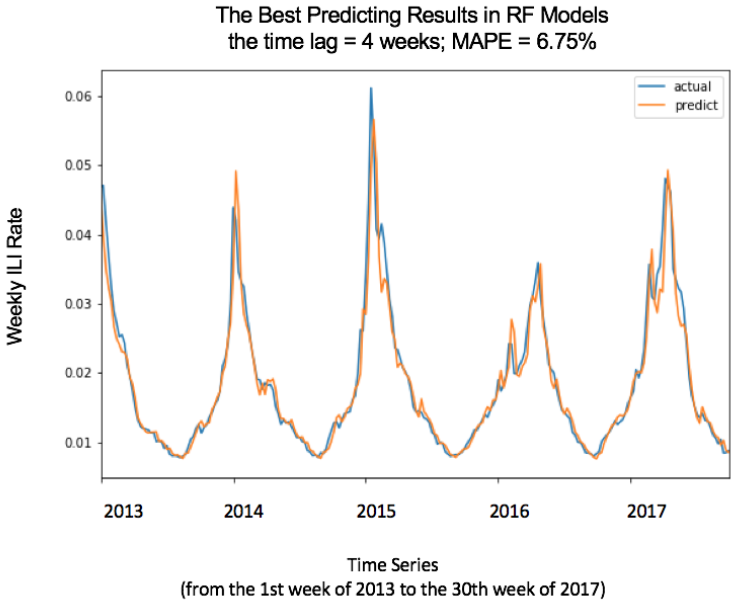


Figure 4.3: The best predicting results of the RF models.

When adopting the time lag of 4 weeks, the RF model achieved its best results. The X-axis represents time, and the Y-axis represents the U.S. weekly ILI rates.

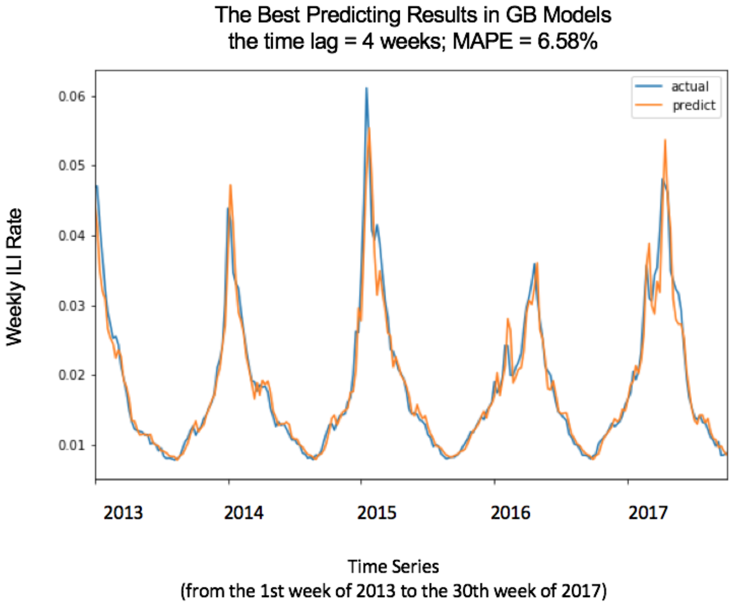


Figure 4.4: The best predicting results of the GB models.

When adopting the time lag of 4 weeks, the GB model achieved its best results. The X-axis represents time, and the Y-axis represents the U.S. weekly ILI rates.

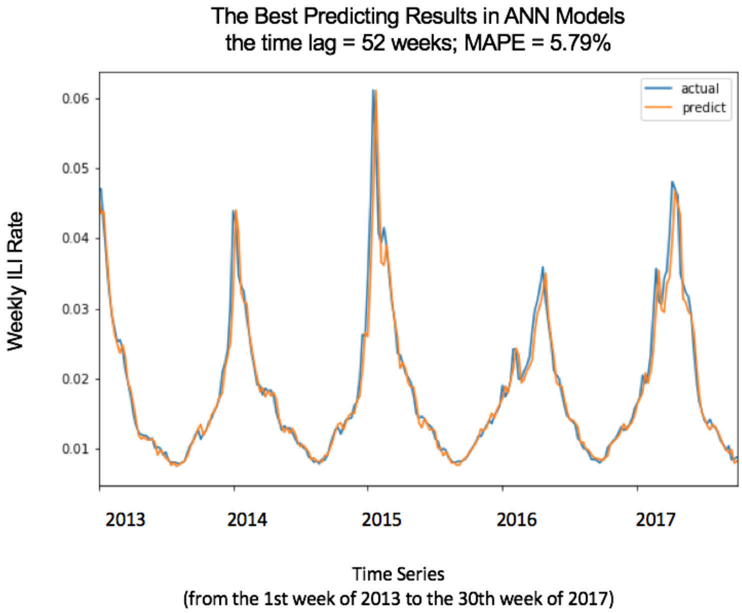


Figure 4.5: The best predicting results of the ANN models.

When adopting the time lag of 52 weeks, the ANN model achieved its best results. The X-axis represents time, and the Y-axis represents the U.S. weekly ILI rates.

In the model of 10-layer LSTM with regularization, the RMSE was 0.00246, 0.00239, 0.00231, 0.00233, 0.00230, and 0.00227, respectively.

4.2 Results of Comparative Study on Algorithms of Multistep Prediction

This part shows the predicting accuracy (MAPEs and RMSE) of four algorithms of multistep prediction in the second research.

4.2.1 Results of MSP

Table 4.5 and Figure 4.7 show the MAPEs of LSTM with MSP algorithm. When predicting the ILI rates of the coming 2nd. 3rd, ... , 13th weeks, the 3-Layer LSTM achieved the predicting MAPEs of 19.76%, 32.97%, 49.10%, 67.87%, 89.89%, 114.98%, 144.00%, 177.72%, 217.72%, 267.33%, 332.58%, and 426.83%, respectively; the 4-Layer LSTM achieved the predicting MAPEs of 9.57%, 11.96%, 13.26%, 13.68%, 13.79%, 14.08%, 14.09%, 14.10%, 13.93%, 13.79%, 13.76%, and 13.77%, respectively; the 5-Layer LSTM achieved the predicting MAPEs of 9.60%, 11.91%, 13.22%, 13.93%, 14.39%, 14.90%, 15.10%, 15.17%, 15.20%, 15.14%, 15.27%, and 15.32%, respectively; the 6-Layer LSTM achieved the predicting MAPEs of 9.02%, 11.48%, 13.11%, 14.55%, 16.17%, 17.19%, 17.44%, 16.98%, 16.69%, 16.89%, 17.13%, and 17.75%, respectively; the 10-Layer LSTM achieved the predicting MAPEs of 9.35%, 11.78%, 13.51%, 14.90%, 16.27%, 17.35%, 18.30%, 18.89%, 19.41%, 19.77%, 19.88%, and 19.93%, respectively. In the results of the 3-Layer LSTM with MSP algorithm, the MAPEs increased by nearly 22 times as the multisteps increased. Comparatively, the MAPEs of the 4-layer LSTM with MSP algorithm increased limitedly, from 9.57% to 13.77% with some

Table 4.3: The MAPEs of 3-layer LSTM, 4-layer LSTM, 4-layer LSTM with dropout, 4-layer LSTM with regularization, 5-layer LSTM, 5-layer LSTM with regularization, 6-layer LSTM with regularization, and 10-layer LSTM with regularization.

time lags (weeks)	2	4	9	13	26	52	mean MAPE of different LSTM (%)
3-layer LSTM (%)	6.80	7.00	7.00	6.87	6.93	6.71	6.89
4-layer LSTM (%)	6.69	6.42	6.28	6.17	6.06	5.44	6.18
4-layer LSTM with dropout (%)	7.62	7.17	7.26	7.18	6.56	6.27	7.01
4-layer LSTM with regularization (%)	6.74	6.32	6.22	6.09	6.07	5.45	6.15
5-layer LSTM (%)	6.85	6.61	7.20	6.64	6.53	6.28	6.69
5-layer LSTM with regularization (%)	6.56	6.38	6.11	6.01	5.91	5.53	6.08
6-layer LSTM with regularization (%)	6.61	6.52	6.20	6.12	5.91	5.46	6.14
10-layer LSTM with regularization (%)	6.46	6.42	5.98	5.90	5.75	5.72	6.04
mean MAPE of different time lags (%)	6.79	6.61	6.53	6.37	6.22	5.86	

The columns represent different time lags in weeks. The rows represent the models of the different number of layers with or without regularization or dropout. The last column represents the average MAPE of LSTM models with same structure and hyperparameters but different time lags. The last row represents the average MAPE of LSTM models with different structures and hyperparameters but same time lags. The highlighted cells represent the lowest MAPE in each type of models. When adopting a time lag of 52 weeks, the LSTM of all structures achieved the lowest MAPEs.

Table 4.4: The RMSEs of 3-layer LSTM, 4-layer LSTM, 4-layer LSTM with dropout, 4-layer LSTM with regularization, 5-layer LSTM, 5-layer LSTM with regularization, 6-layer LSTM with regularization, and 10-layer LSTM with regularization.

time lags (weeks)	2	4	9	13	26	52	mean RMSE of different LSTM
3-layer LSTM	0.00253	0.00254	0.00250	0.00249	0.00241	0.00210	0.00243
4-layer LSTM	0.00261	0.00258	0.00257	0.00257	0.00252	0.00243	0.00255
4-layer LSTM with dropout	0.00253	0.00252	0.00252	0.00250	0.00246	0.00235	0.00248
4-layer LSTM with regularization	0.00262	0.00256	0.00251	0.00250	0.00263	0.00256	0.00256
5-layer LSTM	0.00250	0.00249	0.00246	0.00243	0.00241	0.00210	0.00240
5-layer LSTM with regularization	0.00266	0.00255	0.00256	0.00244	0.00259	0.00257	0.00256
6-layer LSTM with regularization	0.00259	0.00256	0.00250	0.00248	0.00252	0.00244	0.00252
10-layer LSTM with regularization	0.00246	0.00239	0.00231	0.00233	0.00230	0.00227	0.00234
mean RMSE of different time lags	0.00256	0.00252	0.00249	0.00247	0.00248	0.00235	

The columns represent different time lags in weeks. The rows represent the models of the different number of layers with or without regularization or dropout. The last column represents the average RMSE of LSTM models with same structure and hyperparameters but different time lags. The last row represents the average RMSE of LSTM models with different structures and hyperparameters but same time lags. The highlighted cells represent the lowest MAPE in each type of models.

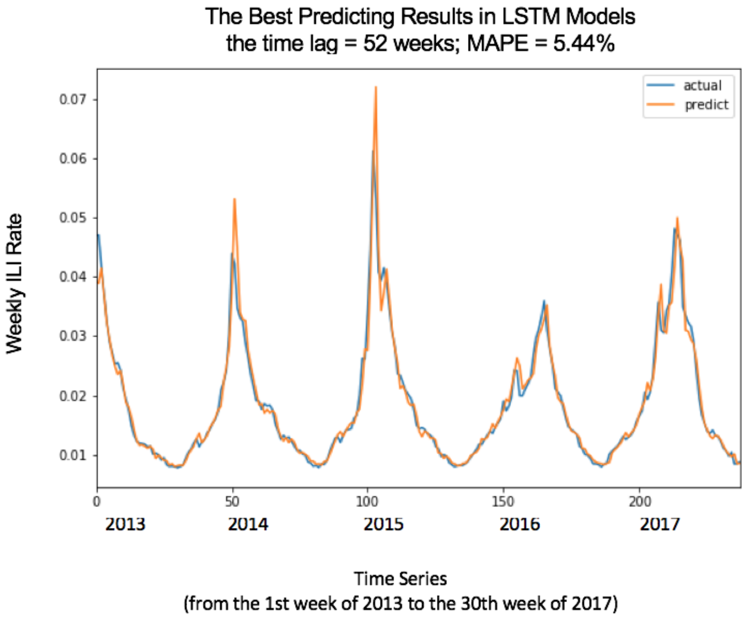


Figure 4.6: The best predicting results of the LSTM models.

When adopting the time lag of 52 weeks, the LSTM model achieved its best results. The X-axis represents time, and the Y-axis represents the U.S. weekly ILI rates.

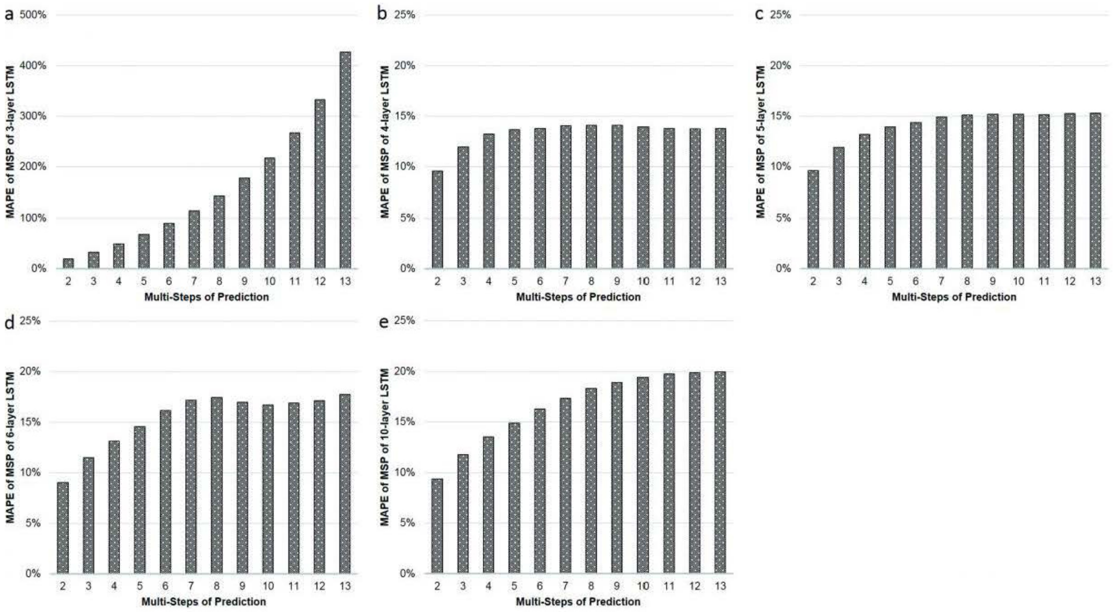


Figure 4.7: The MAPEs of LSTM with MSP.

The Y-axis represents the MAPE of predictions and the X-axis represents the multistep of predictions. The (a) (b) (c) (d) and (e) illustrate the MAPEs with the MSP algorithm of 3-, 4-, 5-, 6-, 10-layer LSTM, respectively. The MAPEs increased by nearly 22 times as the multistep increased in 3-layer LSTM MSP. Comparatively, the MAPEs increased limitedly from 9.57% to 13.77% with some slight setbacks in 10-, 11-, and 12-step prediction in 4-layer LSTM MSP. The MAPEs increased limitedly from 9.60% to 14.11% with a slight setback in 11-step prediction in 5-layer LSTM MSP. The MAPEs increased limitedly from 9.02% to 15.37% with some slight setbacks in 9- and 10-step prediction in 6-layer LSTM MSP. The MAPEs increased limitedly from 9.35% to 16.61% with no setbacks in 10-layer LSTM MSP.

Table 4.5: The MAPEs of LSTM with the multistep predicting algorithms of MSP.

the numbers of multisteps	MSP of 3-layer of LSTM (%)	MSP of 4-layer of LSTM (%)	MSP of 5-layer of LSTM (%)	MSP of 6-layer of LSTM (%)	MSP of 10-layer of LSTM (%)	average MAPE of MSP of 3-, 4-, 5-, 6-, 10-layer LSTM (%)
2	19.76	9.57	9.60	9.02	9.35	11.46
3	32.97	11.96	11.91	11.48	11.78	16.02
4	49.10	13.26	13.22	13.11	13.51	20.44
5	67.87	13.68	13.93	14.55	14.90	24.98
6	89.89	13.79	14.39	16.17	16.27	30.10
7	114.98	14.08	14.90	17.19	17.35	35.70
8	144.00	14.09	15.10	17.44	18.30	41.79
9	177.72	14.10	15.17	16.98	18.89	48.57
10	217.72	13.93	15.20	16.69	19.41	56.59
11	267.33	13.79	15.14	16.89	19.77	66.58
12	332.58	13.76	15.27	17.13	19.88	79.73
13	426.83	13.77	15.32	17.75	19.93	98.72
average MAPE of MSP of all multisteps (%)	161.73	13.31	14.10	15.37	16.61	

The rows represent different time lags in weeks. The columns represent the models of the different number of layers with or without regularization or dropout. The last row represents the average MAPE of LSTM models with same structure and hyperparameters but different time lags. The last column represents the average MAPE of LSTM models with different structures and hyperparameters but same time lags. The highlighted cells represent the lowest MAPE in each type of models.

Table 4.6: The MAPEs of LSTM with the multistep predicting algorithms of AMSP.

the numbers of multisteps	AMSP of 3-Layer of LSTM (%)	AMSP of 4-Layer of LSTM (%)	AMSP of 5-Layer of LSTM (%)	AMSP of 6-Layer of LSTM (%)	AMSP of 10-Layer of LSTM (%)	average MAPE of AMSP of 3-, 4-, 5-, 6-, 10-layer LSTM (%)
2	8.88	9.29	9.45	9.36	8.87	9.17
3	11.33	11.35	11.31	11.38	11.05	11.28
4	12.65	12.91	13.07	12.97	12.38	12.80
5	13.26	13.57	14.08	13.77	13.48	13.63
6	13.90	13.98	14.56	13.25	13.92	13.92
7	14.75	13.68	14.68	14.32	14.28	14.34
8	15.21	14.55	14.11	14.47	15.39	14.75
9	14.63	13.77	15.50	14.41	15.26	14.71
10	14.91	14.62	14.00	13.70	15.21	14.49
11	14.58	15.14	15.31	14.80	17.35	15.44
12	14.72	15.11	15.25	13.62	15.46	14.83
13	14.70	14.57	15.16	14.85	14.29	14.71
average MAPE of AMSP of all multisteps (%)	13.63	13.54	13.87	13.41	13.91	

The rows represent different time lags in weeks. The columns represent the models of the different number of layers with or without regularization or dropout. The last row represents the average MAPE of LSTM models with same structure and hyperparameters but different time lags. The last column represents the average MAPE of LSTM models with different structures and hyperparameters but same time lags. The highlighted cells represent the lowest MAPE in each type of models.

slight setbacks in 10-, 11-, and 12-step prediction. The MAPEs of LSTM of 5, 6, and 10 layers with MSP also increased limitedly.

4.2.2 Results of AMSP

Table 4.6 and 4.8 show the MAPEs of LSTM with AMSP algorithm. When predicting the ILI rates of the coming 2nd. 3rd, ... , 13th weeks, the 3-Layer LSTM achieved the predicting MAPEs of 8.88%, 11.33%, 12.65%, 13.26%, 13.90%, 14.75%, 15.21%, 14.63%, 14.91%, 14.58%, 14.72%, 14.70 %, respectively; the 4-Layer LSTM achieved the predicting MAPEs of 9.29%, 11.35%, 12.91%, 13.57%, 13.98%, 13.68%, 14.55%, 13.77%, 14.62%, 15.14%, 15.11%, 14.57 %, respectively; the 5-Layer LSTM achieved the predicting MAPEs of 9.45%, 11.31%, 13.07%, 14.08%, 14.56%, 14.68%, 14.11%, 15.50%, 14.00%, 15.31%, 15.25%, 15.16 %, respectively; the 6-Layer LSTM achieved the predicting MAPEs of 9.36%, 11.38%, 12.97%, 13.77%, 13.25%, 14.32%, 14.47%, 14.41%, 13.70%, 14.80%, 13.62%, 14.85 %, respectively; the 10-Layer LSTM achieved the predicting MAPEs of 8.87%, 11.05%, 12.38%, 13.48%, 13.92%, 14.28%, 15.39%, 15.26%, 15.21%, 17.35%, 15.46%, 14.29 %, respectively. In AMSP, the average MAPE increased from 9.17% to 14.72% as the multisteps increased from 2 to 13, and varied from 13.63% to 13.92% as the number of layers of LSTM increased from 3 to 10 layers.

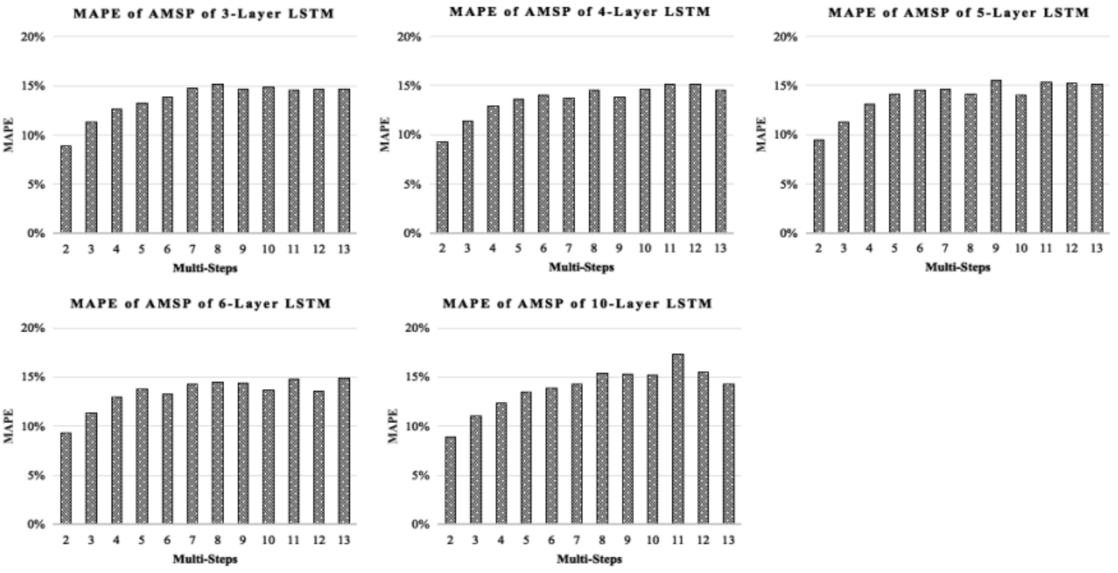


Figure 4.8: The MAPEs of LSTM with AMSP.

The Y-axis represents the MAPE of predictions and the X-axis represents the multistep of predictions. The (a) (b) (c) (d) and (e) illustrate the MAPEs with the AMSP algorithm of 3-, 4-, 5-, 6-, 10-layer LSTM, respectively.

4.2.3 Results of MSOP

Table 4.7 and Figure 4.9 show the MAPEs of LSTM with MSOP algorithm. When predicting the ILI rates of the coming 2nd. 3rd, ... , 13th weeks, the 3-Layer LSTM achieved the predicting MAPEs of 8.76%, 10.78%, 12.03%, 12.99%, 13.41%, 14.08%, 14.56%, 14.53%, 14.40%, 14.73%, 13.98%, 14.42 %, respectively; the 4-Layer LSTM achieved the predicting MAPEs of 8.84%, 10.49%, 12.20%, 12.78%, 13.31%, 14.21%, 13.72%, 13.73%, 13.83%, 13.97%, 13.81%, 14.33 %, respectively; the 5-Layer LSTM achieved the predicting MAPEs of 8.96%, 10.16%, 12.37%, 13.10%, 13.50%, 14.09%, 14.55%, 14.32%, 14.22%, 14.29%, 14.58%, 13.45 %, respectively; the 6-Layer LSTM achieved the predicting MAPEs of 9.11%, 10.46%, 12.11%, 12.98%, 12.87%, 13.78%, 14.19%, 14.09%, 14.56%, 13.30%, 13.67%, 14.04 %, respectively; the 10-Layer LSTM achieved the predicting MAPEs of 8.88%, 10.71%, 11.95%, 13.14%, 13.69%, 13.99%, 15.10%, 14.67%, 15.67%, 14.50%, 14.26%, 14.09 %, respectively. In MSOP, the average MAPE increased from 8.91% to 14.06% as the multisteps increased from 2 to 13, and varied from 12.94% to 13.39% as the number of layers of LSTM increased from 3 to 10 layers.

4.2.4 Results of MOP

Table 4.8 and Figure 4.10 shows the MAPEs of LSTM with MOP algorithm. When predicting the ILI rates of the coming 2nd. 3rd, ... , 13th weeks, the 3-Layer LSTM achieved the predicting MAPEs of 11.77%, 13.04%, 18.16%, 19.73%, 19.25%, 21.22%, 20.35%, 24.04%, 22.49%, 22.68%, 23.09%, 24.53 %, respectively; the 4-Layer LSTM achieved the predicting MAPEs of 10.16%, 13.61%, 16.14%, 17.44%, 18.24%, 20.62%, 20.56%, 24.87%, 24.20%, 19.28%, 24.25%, 24.57 %, respectively; the 5-Layer LSTM achieved the predicting MAPEs of 9.88%, 11.73%, 13.53%, 14.98%, 14.88%, 16.83%, 17.93%, 20.44%, 21.25%, 21.05%, 22.87%, 29.55 %, respectively; the 6-Layer LSTM achieved the predicting MAPEs of 10.35%, 12.45%, 15.25%, 19.17%, 18.51%, 18.69%, 20.81%, 18.01%, 21.71%,

Table 4.7: The MAPEs of LSTM with the multistep predicting algorithms of MSOP.

the numbers of multisteps	MSOP of 3-Layer of LSTM (%)	MSOP of 4-Layer of LSTM (%)	MSOP of 5-Layer of LSTM (%)	MSOP of 6-Layer of LSTM (%)	MSOP of 10-Layer of LSTM (%)	average MAPE of MSOP of 3-, 4-, 5-, 6-, 10-layer LSTM (%)
2	8.76	8.84	8.96	9.11	8.88	8.91
3	10.78	10.49	10.16	10.46	10.71	10.52
4	12.03	12.20	12.37	12.11	11.95	12.13
5	12.99	12.78	13.10	12.98	13.14	13.00
6	13.41	13.31	13.50	12.87	13.69	13.36
7	14.08	14.21	14.09	13.78	13.99	14.03
8	14.56	13.72	14.55	14.19	15.10	14.42
9	14.53	13.73	14.32	14.09	14.67	14.27
10	14.40	13.83	14.22	14.56	15.67	14.54
11	14.73	13.97	14.29	13.30	14.50	14.16
12	13.98	13.81	14.58	13.67	14.26	14.06
13	14.42	14.33	13.45	14.04	14.09	14.06
average MAPE of MSOP of all multisteps (%)	13.22	12.93	13.13	12.93	13.39	

The rows represent different time lags in weeks. The columns represent the models of the different number of layers with or without regularization or dropout. The last row represents the average MAPE of LSTM models with same structure and hyperparameters but different time lags. The last column represents the average MAPE of LSTM models with different structures and hyperparameters but same time lags. The highlighted cells represent the lowest MAPE in each type of models.

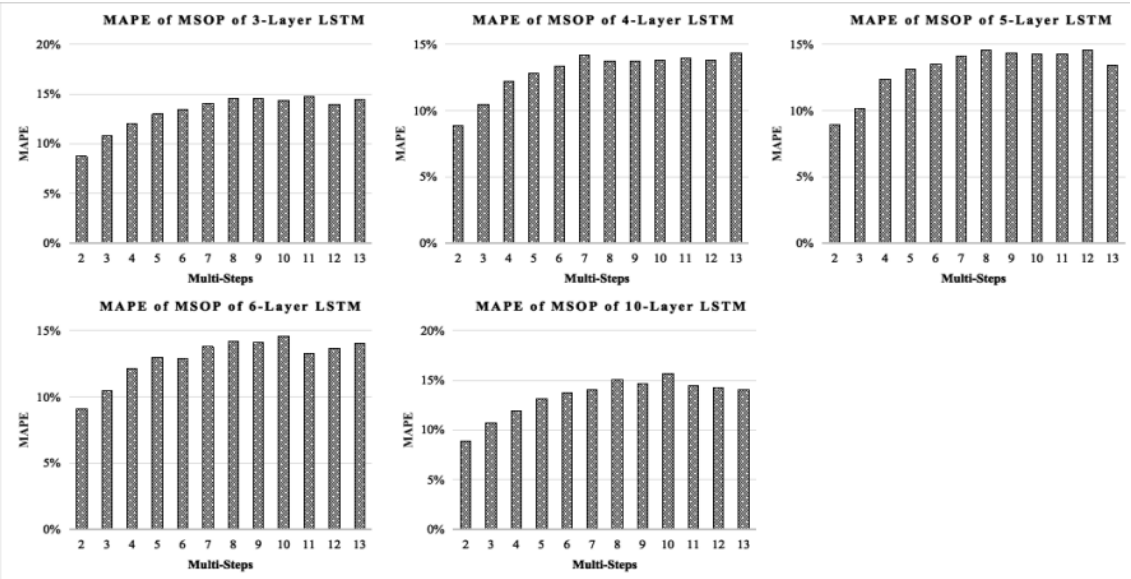


Figure 4.9: The MAPEs of LSTM with MSOP.

The Y-axis represents the MAPE of predictions and the X-axis represents multisteps of predictions. The (a) (b) (c) (d) and (e) illustrate the MAPEs with the MSOP algorithm of 3-, 4-, 5-, 6-, 10-layer LSTM, respectively.

Table 4.8: The MAPEs of LSTM with the multistep predicting algorithms of MOP.

the numbers of multisteps	MOP of 3-Layer of LSTM (%)	MOP of 4-Layer of LSTM (%)	MOP of 5-Layer of LSTM (%)	MOP of 6-Layer of LSTM (%)	MOP of 10-Layer of LSTM (%)	average MAPE of MOP of 3-, 4-, 5-, 6-, 10-layer LSTM (%)
2	11.77	10.16	9.88	10.35	11.58	10.75
3	13.04	13.61	11.73	12.45	12.87	12.74
4	18.16	16.14	13.53	15.25	16.66	15.95
5	19.73	17.44	14.98	19.17	15.29	17.32
6	19.25	18.24	14.88	18.51	24.39	19.06
7	21.22	20.62	16.83	18.69	20.15	19.50
8	20.35	20.56	17.93	20.81	20.86	20.10
9	24.04	24.87	20.44	18.01	23.79	22.23
10	22.49	24.20	21.25	21.71	20.63	22.05
11	22.68	19.28	21.05	22.34	27.19	22.51
12	23.09	24.25	22.87	23.71	26.79	24.14
13	24.53	24.57	29.55	25.31	27.98	26.39
average MAPE of MOP of all multisteps (%)	20.03	19.50	17.91	18.86	20.68	

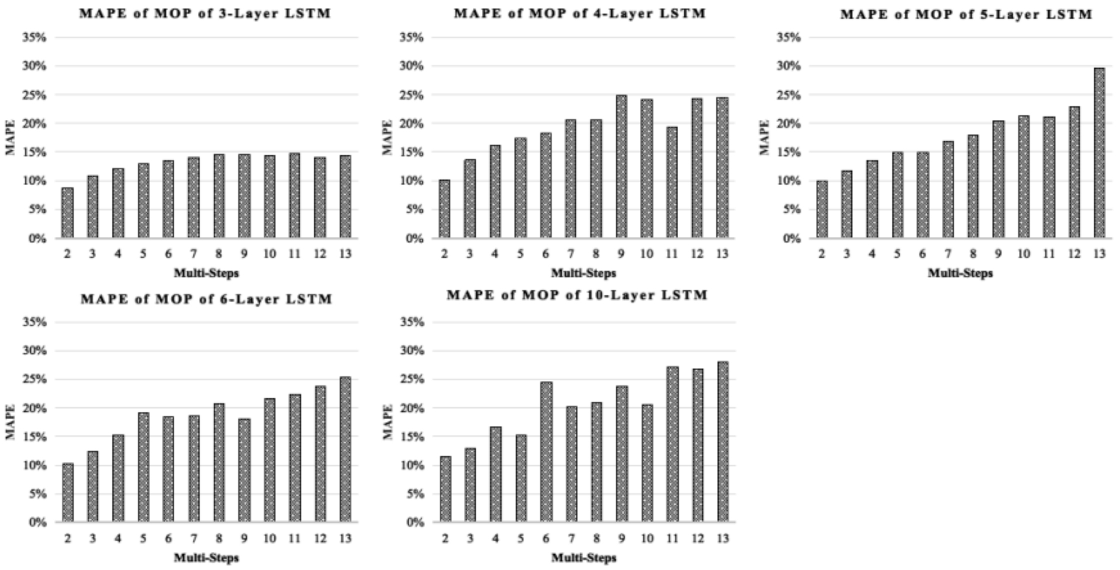


Figure 4.10: The MAPEs of LSTM with MOP.

The Y-axis represents the MAPE of predictions and the X-axis represents multisteps of predictions. The (a) (b) (c) (d) and (e) illustrate the MAPEs with the MSOP algorithm of 3-, 4-, 5-, 6-, 10-layer LSTM, respectively.

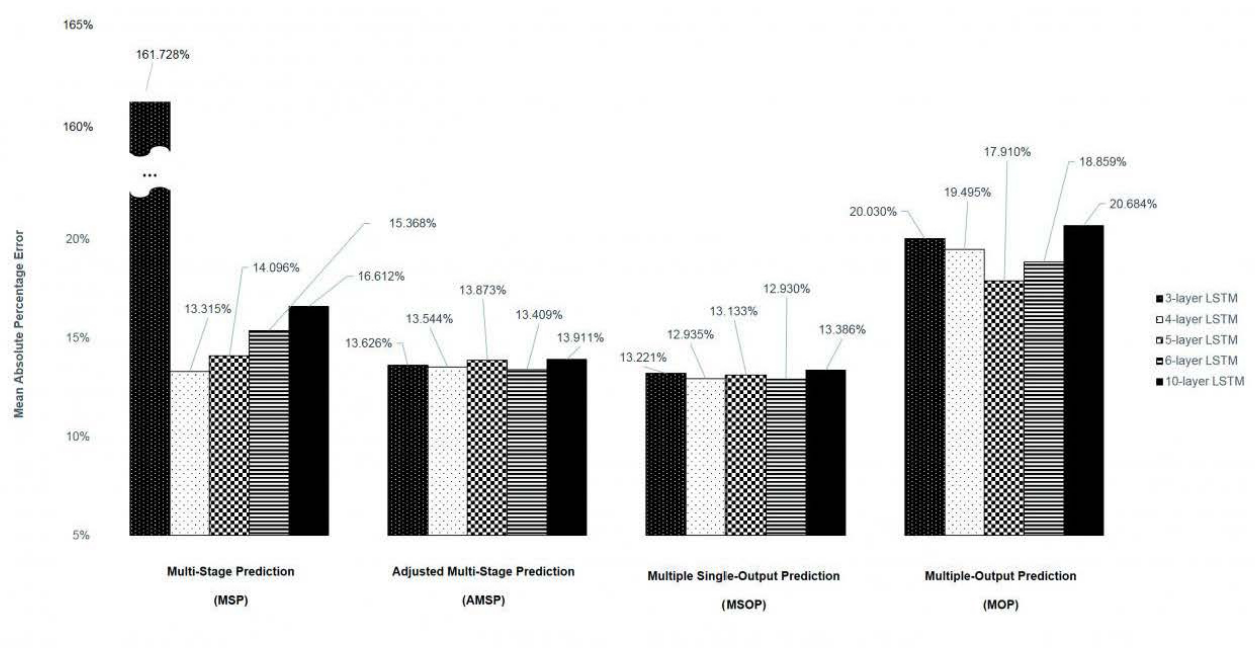


Figure 4.11: The average MAPEs of LSTM of the different number of layers with MSP, AMSP, MSOP, and MOP.

The X-axis represents different structures and different algorithms of multistep prediction. The Y-axis represents the predictive MAPEs.

22.34%, 23.71%, 25.31 %, respectively; the 10-Layer LSTM achieved the predicting MAPEs of 11.58%, 12.87%, 16.66%, 15.29%, 24.39%, 20.15%, 20.86%, 23.79%, 20.63%, 27.19%, 26.79%, 27.98 %, respectively. In MOP, the average MAPE increased from 10.75% to 26.39% as the multisteps increased from 2 to 13, and varied from 17.91% to 20.68% as the number of layers of LSTM increased from 3 to 10 layers.

Table 4.9: The average MAPE of LSTM of the different number of layers with MSP, AMSP, MSOP, and MOP.

	MSP (%)	AMSP (%)	MSOP (%)	MOP (%)
3-layer LSTM	137.628	13.626	13.221	20.030
4-layer LSTM	13.315	13.544	12.935	19.495
5-layer LSTM	14.096	13.873	13.133	17.910
6-layer LSTM	15.368	13.409	12.930	18.859
10-layer LSTM	16.612	13.911	13.386	20.684

The columns represent different algorithms of different algorithms. The rows represent different structures. The highlighted cell represents the lowest MAPE.

4.2.5 Comparison of MSP, AMSP, MSOP, and MOP

Table 4.9 and Figure 4.11 compares the average MAPE of LSTM with MSP, AMSP, MSOP, and MOP. The different numbers of the layers impacted the predicting accuracy tremendously in MSP (from 13.315% to 161.728%); slightly in MOP (from 17.910% to 20.684%), and barely in AMSP (from 13.626% to 13.911%) and MSOP (from 12.930% to 13.386%). In sharp contrast to MSP, the accuracy of AMSP, MSOP, and MOP had little improvement when more layers were used. Finally,

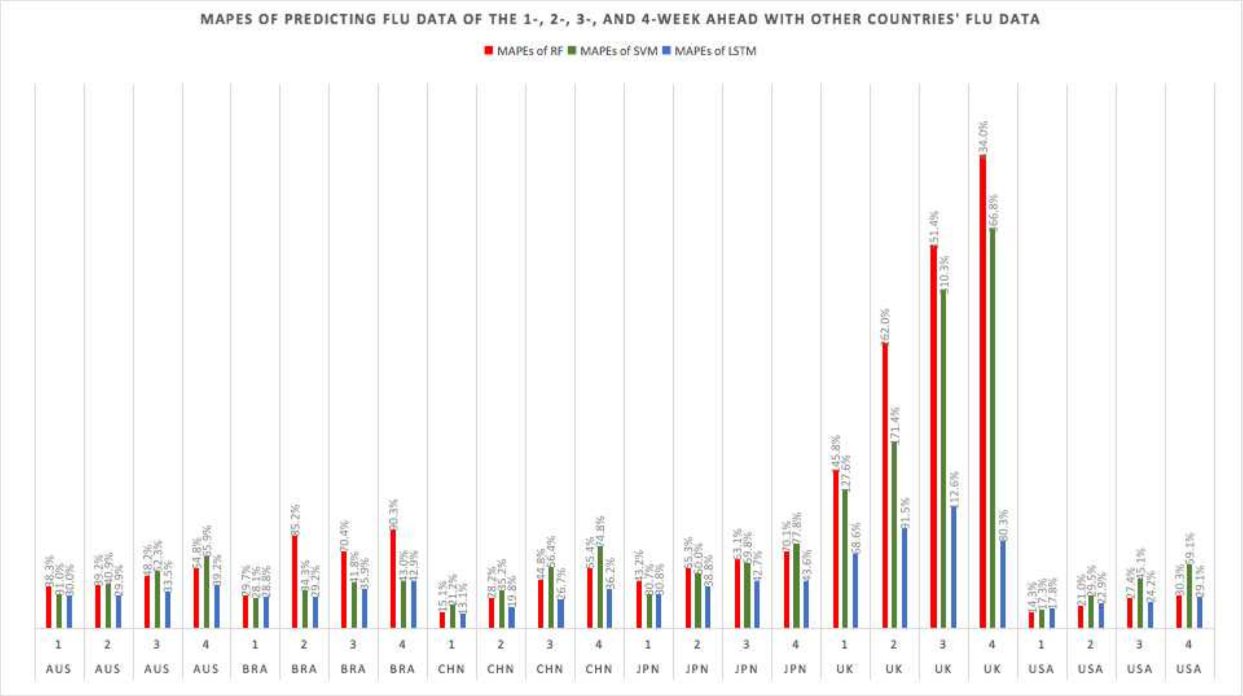


Figure 4.12: The MAPEs with considering geolocational-temporal features.

The X-axis represents different countries and different multistep prediction. Different colors represent different model types. The Y-axis represents the MPAEs of prediction.

implementing MSOP in the 6-layer LSTM structure achieved the best accuracy in this study, and when we implemented MSOP in the 6-layer LSTM structure, all the MAPEs from the 2-step-ahead to the 13-step-ahead prediction for the U.S. ILI rates were all less than 15%, averagely 12.930%, as Table 4.7 and Figure 4.9 showed.

4.2.6 Results of Multistep Geolocational Temporal Prediction of Worldwide Flu Outbreaks

This section shows the predicting results of geolocational-temporal multistep prediction in the third research.

4.2.7 Comparison of the Geolocational-Temporal Multistep Prediction of RF, SVR, and LSTM

Table 4.10, Figure 4.12, and Figure 4.12 compare the geolocational-temporal multistep prediction of RF, SVR, and LSTM. In each model, we also performed prediction with and without feature space of other countries. For example, the 1-, 2-, 3-, and 4-week-ahead MAPEs of the LSTM models with other countries' flu data were 13.1%, 19.8%, 26.7%, 36.2%; while the MAPEs of the LSTM models of predicting without other countries' flu data were 12.5%, 20.2%, 29.0%, and 36.7%. Figure 4.12 compares the MAPEs of SVM, RF, LSTM models of predicting flu data of the 1-, 2-, 3-, and 4-week-ahead with other countries' flu data. Figure 4.13 compares the MAPEs of SVM, RF, LSTM models of predicting flu data of the 1-, 2-, 3-, and 4-week-ahead without other countries' flu data. In almost all cases, the LSTM models achieved the lowest MAPEs.

Table 4.10: The multistep flu outbreak prediction considering the geolocational-temporal features.

hemisphere	country	steps ahead	SVM		RF		LSTM	
			with other countries	without other countries	with other countries	without other countries	with other countries	without other countries
Southern	AUS	1	0.310	0.318	0.383	0.374	0.300	0.235
		2	0.409	0.413	0.392	0.385	0.299	0.247
		3	0.523	0.524	0.482	0.459	0.335	0.299
		4	0.659	0.661	0.548	0.527	0.392	0.320
	BRA	1	0.281	0.285	0.297	0.330	0.288	0.228
		2	0.343	0.327	0.852	0.309	0.292	0.256
		3	0.418	0.358	0.704	0.298	0.359	0.308
		4	0.430	0.420	0.903	0.321	0.429	0.355
Northern	CHN	1	0.212	0.193	0.151	0.199	0.131	0.125
		2	0.352	0.352	0.282	0.287	0.198	0.202
		3	0.564	0.558	0.448	0.426	0.267	0.290
		4	0.748	0.741	0.554	0.507	0.362	0.367
	JPN	1	0.307	0.291	0.432	0.411	0.308	0.279
		2	0.500	0.479	0.553	0.477	0.388	0.395
		3	0.598	0.568	0.631	0.493	0.427	0.428
		4	0.778	0.736	0.701	0.658	0.436	0.540
	UK	1	1.276	1.305	1.458	1.221	0.686	0.861
		2	1.714	1.656	2.620	2.641	0.915	0.954
		3	3.103	2.888	3.514	3.084	1.126	1.169
		4	3.668	3.274	4.340	5.098	0.803	1.187
	USA	1	0.173	0.176	0.143	0.139	0.178	0.147
		2	0.295	0.293	0.210	0.209	0.229	0.245
		3	0.451	0.427	0.274	0.244	0.242	0.292
		4	0.591	0.586	0.303	0.294	0.291	0.302

The table presents the MAPEs of RF, SVM, and LSTM models with and without other countries' flu data. The rows represent different hemispheres, countries, and multistep prediction. The columns represent different models with or without considering flu data of other countries.

4.2.8 Comparison of the Predicting Accuracy of LSTM Models with and without Feature Space of Other Countries

Figure 4.14 compares the MAPEs of the LSTM models with and without feature space of other countries. As for countries in the Southern Hemisphere, i.e. Australia and Brazil, the MAPEs of predicting flu data of the 1-, 2-, 3-, and 4-week-ahead with other countries were slightly higher than those of predicting without other countries. As for countries in the Northern Hemisphere, i.e. China, Japan, the UK, and the US, the MAPEs of predicting flu data of the 2-, 3-, and 4-week-ahead with other countries were lower than those of predicting without other countries. Interestingly, when predicting flu data of the 1st week ahead in the Northern Hemisphere, the MAPEs of predicting with other countries were usually higher than those of predicting without other countries, except for the UK.

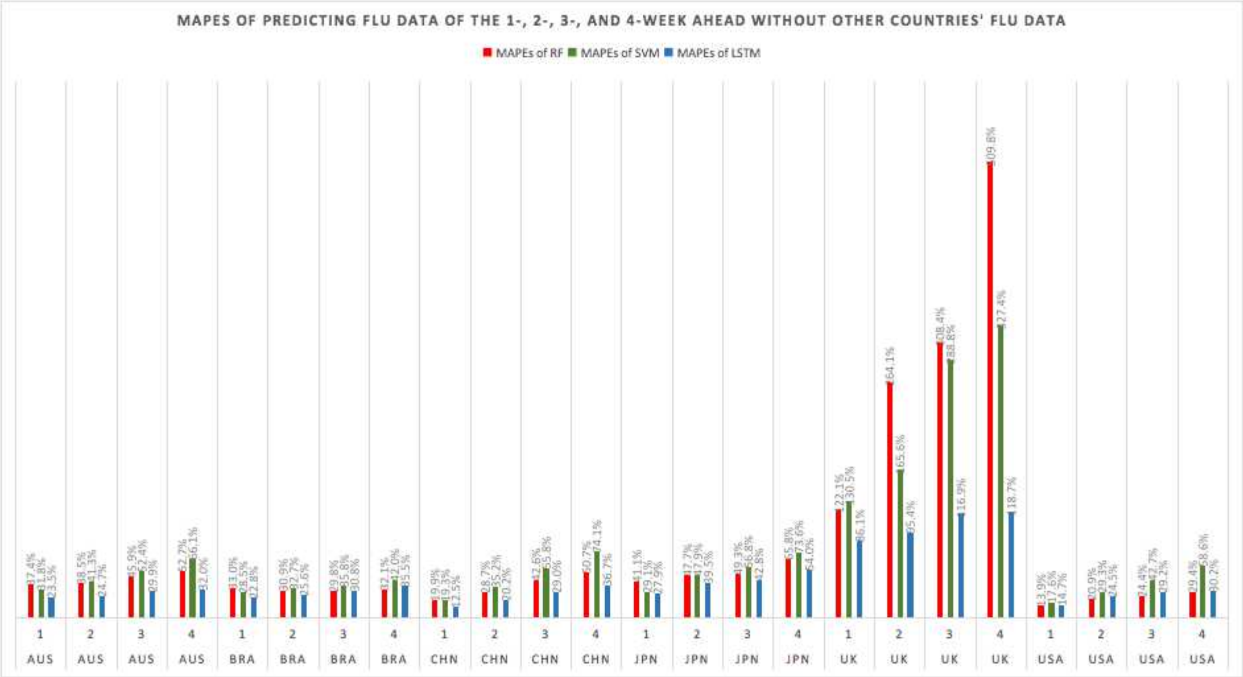


Figure 4.13: The MAPEs without considering geolocational-temporal features.

The X-axis represents different countries and different multistep prediction. Different colors represent different model types. The Y-axis represents the MPAEs of prediction.

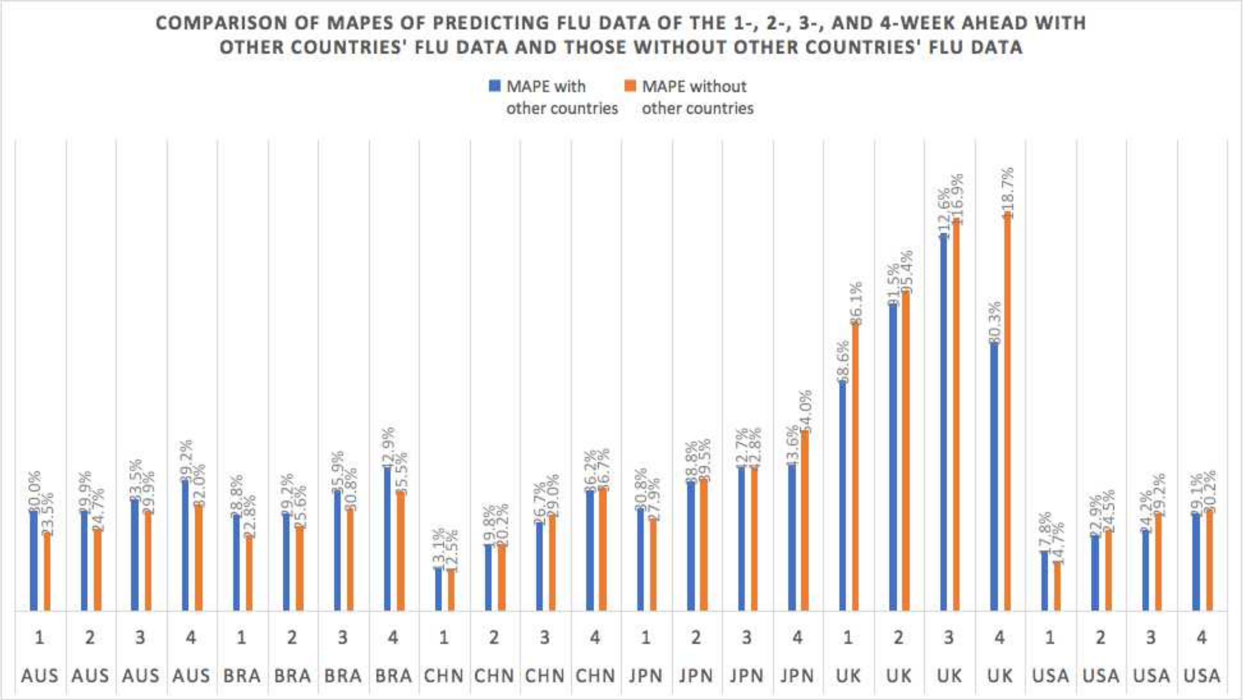


Figure 4.14: The result of predicting flu outbreaks using LSTM.

The X-axis represents different countries and different multistep prediction. Different colors represent models with or without considering flu data from other countries. The Y-axis represents the MPAEs of prediction.

Chapter 5

Discussion

This section discusses all three pieces of research. They are (1) to find the best model and best hyperparameter of flu prediction; (2) to find the best algorithms of multistep prediction for flu outbreaks; and (3) to build up an effective and efficient approach to perform geolocational-temporal multistep of flu.

5.1 Discussion of Comparative Study on Models

This section describes the discussion of the first research. It is (1) to find the best model and best hyperparameter of flu prediction.

5.1.1 Time Lag

Regarding the machine learning models (SVR, RF, and GB), the MAPEs were always approximately 7%, with almost no changes as the time lags increased. That is likely because the machine learning models usually cannot learn the seasonality but can learn the trend of a series of data by inputting the first-order differences into the features. The MAPEs of the ARIMA model decreased from 13.46% to 8.36% when the time lags increased from 2 weeks to 52 weeks (Figure 5.1). The probable explanation for this phenomenon is that ARIMA is an autoregressive model that focus on seasonality. The closer the feature spaces to a complete seasonality, the lower the MAPE will be. In other words, when training ARIMAs for time-series prediction, we need a complete duration. Similar to those of the ARIMA models, the MAPEs of the ANN models also decreased from 13.46% to 8.36% when the time lag increased from 2 weeks to 52 weeks (Figure 5.1). Why not adopt more time lags such as 104 weeks (around 2 years) or more? Firstly, the models with the time lag of 52 weeks (around 1 year) have brought an accuracy of about 95% (i.e. 1 - MAPE), which appeared good enough. Secondly, if we adopt a time lag of 104 weeks or more, we have to drop more training data (the first 104 rows). For one thing, a longer time lag might help improve accuracy. For another thing, the less training data would also setback the accuracy. We suppose that whether the accuracy that would be better or worse should depend on different data. However, a time lag including a complete periodicity is recommended for ARIMA, ANN, and LSTM.

5.1.2 Feature Space

We found that the MAPE of ARIMA > MAPEs of SVR, RF, and GBM > MAPEs of ANN and LSTM. Although the different models have different algorithms, the increasing feature space and the increasing model parameters are important factors that impact the models' accuracy. ARIMA has a very limited feature space. The number of features is equal to the lag times, i.e., 3, 5, 10, 14, 27, or 53. In the ML models (SVR, RF, and GB), we added the first-order differences, and totally, ML models have 105 features. DL models used 255 neurons in every LSTM layer. In brief, a more complicated feature space brings a more accurate prediction.

5.1.3 Regularization

We calculated the standard deviations of the MAPEs of the LSTM models of 3, 4, and 5 layers without regularization and 4, 5, 6, and 10 layers with regularization when using the time lags of 2, 4, 9,13, 26, and 52 weeks (Table 5.1). We found the standard deviations of the MAPEs of the LSTM

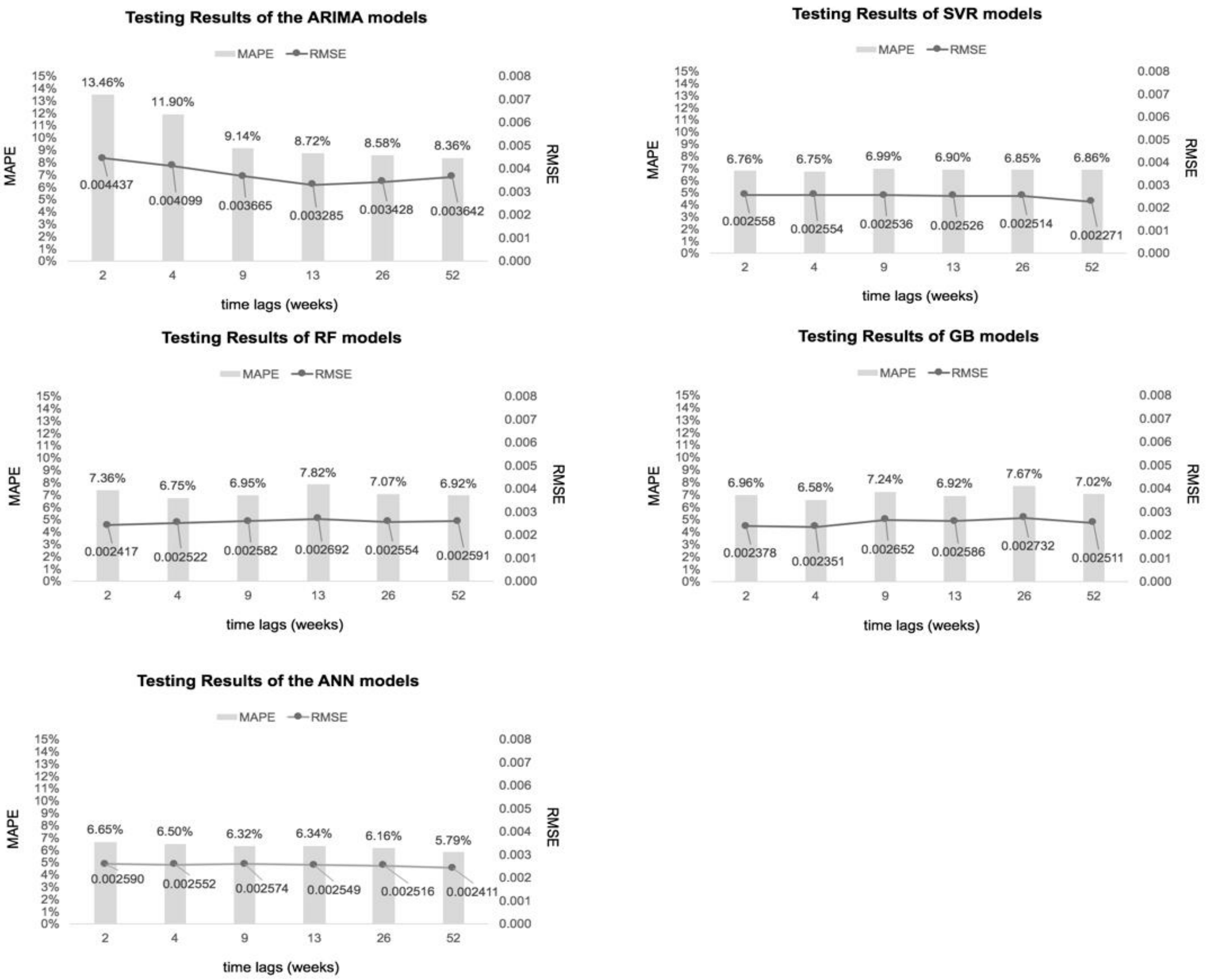


Figure 5.1: The MAPEs and RMSEs of the ARIMA, SVR, RF, GB, and ANN models with the different time lags as the feature spaces.

The X-axis represents the different time lags. The Y-axis represents the MAPEs and RMSEs of prediction. The percentage represents MAPE. The numeric number in lines represent RMSEs.

Table 5.1: The standard deviations of the MAPEs of the LSTM models of 3, 4, and 5 layers without regularization and of 4, 5, 6, and 10 layers with regularization.

	time lag = 2	time lag = 4	time lag = 9	time lag = 13	time lag = 26	time lag =52	row mean
3 layer (%)	6.80	7.00	7.00	6.87	6.93	6.71	6.89
4 layers (%)	6.69	6.42	6.28	6.17	6.06	5.44	6.18
5 layers (%)	6.85	6.61	7.20	6.64	6.53	6.28	6.69
standard deviation of the 3, 4, 5 layers without regularization (%)	0.08	0.30	0.49	0.36	0.44	0.64	0.37
4 layers with regularization (%)	6.74	6.32	6.22	6.09	6.07	5.45	6.15
5 layers with regularization (%)	6.56	6.38	6.11	6.01	5.91	5.53	6.08
6 layers with regularization (%)	6.61	6.52	6.20	6.12	5.91	5.46	6.14
10 layers with regularization (%)	6.46	6.42	5.98	5.90	5.75	5.72	6.04
standard deviation of the 4, 5, 6, 10 layers with regularization (%)	0.12	0.08	0.11	0.10	0.13	0.12	0.05

The columns represent different time lags. The rows represent different models with or without regularization or dropout. The last column represents the average MAPE of the same LSTM structure but different time lags. The last row represents the average MAPE of the different LSTM structure but same time lags.

models with regularization were less than those of the LSTM models without regularization when we used almost all the time lags except the time lag of 2 weeks (Figure 5.1). The probable explanation for this finding is that regularization made the models more robust, and the robust models made the prediction relatively stable. Although we achieved the lowest MAPE (5.44%) when we used the 4-layer LSTM model without regularization, the gap between the MAPEs of the 4-layer LSTM model without and with regularization is very limited (5.45% - 5.44% = 0.01%). Considering that unstable models may lead to poor accuracy if we changed the testing data, we recommend the use of the model with regularization for U.S. flu prediction.

5.1.4 Dropout

In addition to regularization, dropout can also usually help prevent overfit and make the model more robust. We found that the MAPE of the LSTM models with regularization is lower than those with dropout. “Dropout” randomly drops neurons, while “Regularization” selectively drops neurons. Although both suppress the number of neurons, in this study, the selective dropping performed much better than the random dropping (Figure 5.3).

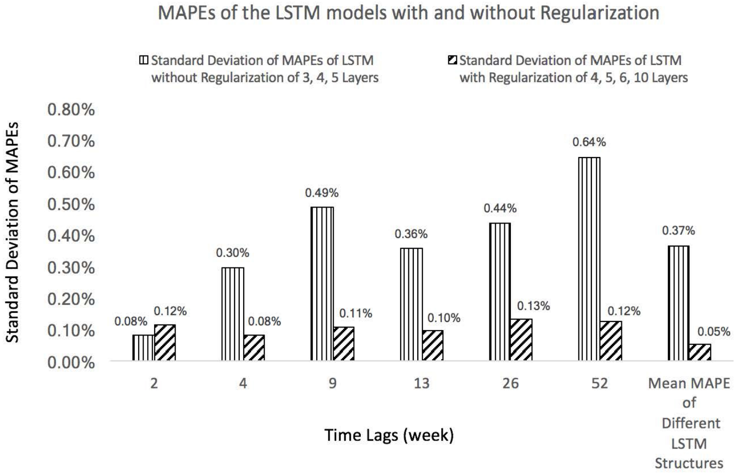


Figure 5.2: The standard deviation of the MAPEs of the LSTM models with and without regularization.

The X-axis represents the different time lags. The Y-axis represents the standard deviation of predictive MAPEs. The standard deviations of the MAPEs of the LSTM models with regularization were less than those of the LSTM models without regularization when we used almost all the time lags except the time lag of 2 weeks.

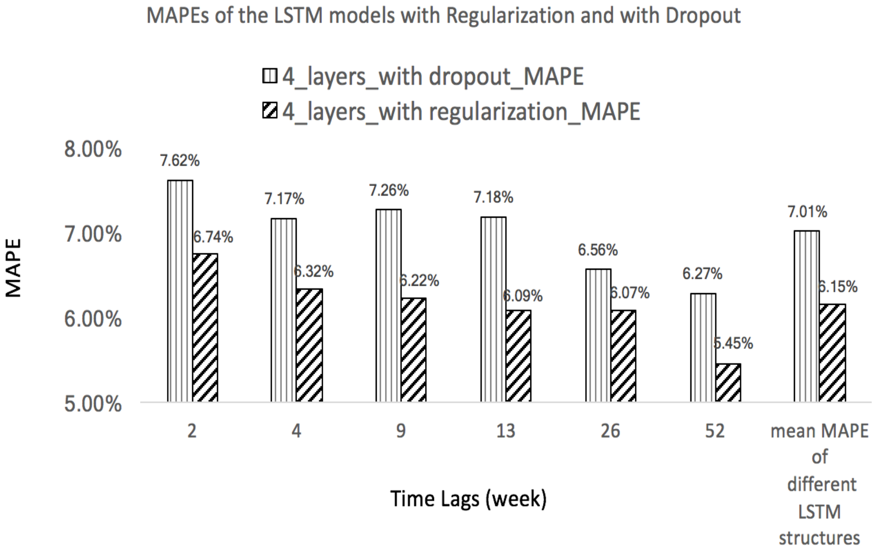


Figure 5.3: MAPEs of the LSTM models with Regularization and with Dropout.

The X-axis represents the different time lags. The Y-axis represents the MAPEs of prediction.

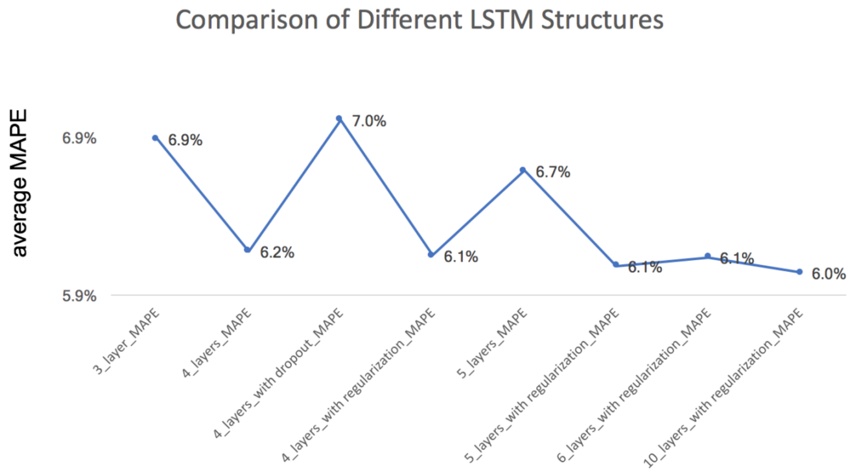


Figure 5.4: Comparison of the different layers for the LSTM with regularization.

The X-axis represents different models. The Y-axis represents average MAPEs.

5.1.5 Number of Layers

Why did a neural network of 4+ layers contribute little to the accuracy? Figure 5.4 compares the accuracy of the different number of layers. Generally, we found extra layers (more than 4 layers) contributed little to improve the predicting accuracy. AS we know, a neural network of more layers means a more complicated structure, which brings a more complicated non-linear regression. This is why, theoretically, a neural network of more layers could bring better accuracy. However, a neural network with more layers could also cause the problem of over-fitting. That is why we usually look for a balance between a more complicated non-linear regression and the problem of over-fitting. The ”regularization” can suppress the problem of the over-fitting. The regularization conquers the problem of the over-fitting by decreasing a model's complication. To show the over-fitting more clearly, we performed some supplement experiment for discussion. Table 5.2 shows the result of the supplement experiment. Table 5.2 shows the MAPEs of neural networks of 3, 4, 5, 6, 7, 8, 9, 10 layers with or without regularization in the first study. Figure 5.5 shows the average MAPEs of neural networks of 3, 4, 5, 6, 7, 8, 9, 10 layers with or without regularization. The X-axis represents the number of layers, and the Y-axis represents the predictive MAPEs.

We found the curve of the MAPEs of the neural networks without regularization (the blue curve in Figure 5.5) presents a “U” shape (the left side was too short, but the right side was obvious). It means that the MAPE decreased (i.e. the accuracy increased) as we added more layers at first, but then increased (i.e. the accuracy decreased) as we added more layers. That was because, in the beginning, a more complicated non-linear regression improved the regressive ability of the model. However, as we added more layers, the problem of over-fitting occurred and spoiled the accuracy of the models. Comparatively, we found, the curve of the MAPEs of the neural networks with regularization (the red curve in Figure 5.5) presents an “L” shape. As aforementioned, the regularization conquers the problem of the over-fitting by decreasing a model's complication. In the beginning, in a shallow neural network such as a 3-layer network, the regularization decreased a model's complication unduly. As a result, the MAPE of the 3-layer model with regularization was even higher than the MAPE of the 3-layer model without regularization. As we added more

Table 5.2: Accumulated MAPEs of the neural networks of 3, 4, 5, 6, 7, 8, 9, 10 layers with or without regularization

the number of layers	adjustment	time lag = 2	time lag = 4	time lag = 9	time lag = 13	time lag = 26	time lag = 52	MAPE
3	without regularization	6.80%	7.00%	7.00%	6.87%	6.93%	6.71%	6.89%
4		6.69%	6.42%	6.28%	6.17%	6.06%	5.44%	6.18%
5		6.85%	6.61%	7.20%	6.64%	6.53%	6.28%	6.69%
6		7.24%	6.72%	6.88%	6.52%	6.82%	6.07%	6.71%
7		44.21%	44.20%	44.25%	44.26%	6.80%	6.04%	31.63%
8		44.22%	44.25%	44.21%	44.22%	44.22%	44.23%	44.22%
9		44.22%	44.22%	44.20%	44.23%	44.21%	44.21%	44.21%
10		44.18%	44.19%	44.22%	44.23%	44.20%	44.22%	44.21%
3	with regularization	10.72%	11.70%	13.55%	13.12%	13.46%	16.24%	13.13%
4		6.74%	6.32%	6.22%	6.09%	6.07%	5.45%	6.15%
5		6.56%	6.38%	6.11%	6.01%	5.91%	5.53%	6.08%
6		6.61%	6.52%	6.20%	6.12%	5.91%	5.46%	6.14%
7		6.86%	6.53%	6.14%	6.01%	5.83%	5.49%	6.14%
8		6.74%	6.32%	6.22%	6.10%	5.92%	5.65%	6.16%
9		6.59%	6.32%	6.18%	5.91%	5.79%	5.48%	6.04%
10		6.46%	6.42%	5.98%	5.90%	5.75%	5.72%	6.04%

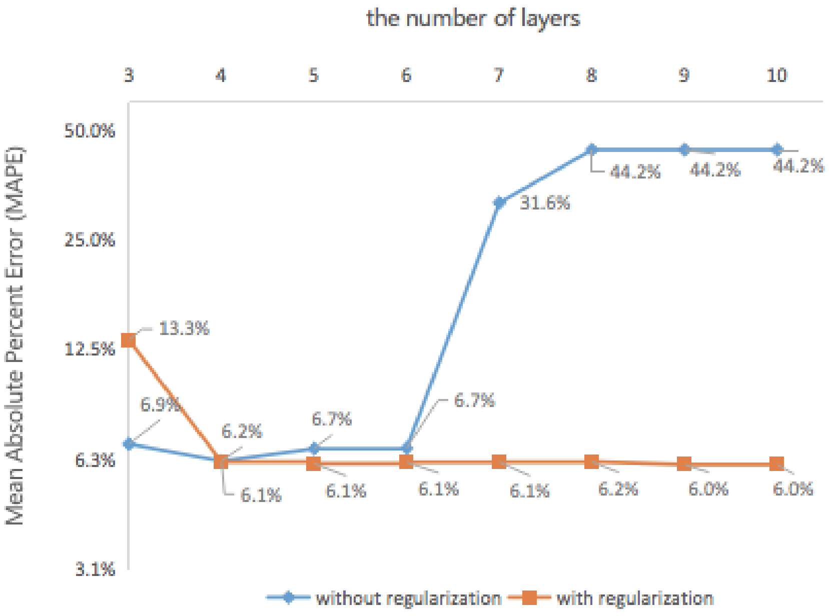


Figure 5.5: The average MAPEs of neural networks of 3, 4, 5, 6, 7, 8, 9, and 10 layers with or without regularization.

The X-axis represents the number of layers. The Y-axis represents MAPEs of prediction. The curve of the MAPEs of the neural networks without regularization (the blue curve) presents a “U” shape (the left side was too short, but the right side was obvious). The curve of the MAPEs of the neural networks with regularization presents an “L” shape.

layers into the predictive model, the MAPE decreased to a relatively constant level, i.e. around 6.1% (between 6.2% and 6.0%). We can conclude the ”regularization” neutralized the effect of the problem of over-fitting and a model’s complication, and therefore successfully suppressed the problem of the over-fitting. Besides, the 4-layer neural network means to perform a non-linear function for 3 times. That could tell us, in mathematics, the time-series predictions of flu outbreaks possible need three-order calculation, which coincidently accords with the SIR model, which also includes three compartments. Anyhow, we need further studies to explore the explainable deep learning in the future. Does a structure of 4 layers still achieve the same result if more data can be achieved? We can divide this question into two questions to answer. (1) Does a structure of 4 layers still achieve the same accuracy when more data can be achieved with that of this study? (2) Does a structure of 4 layers still achieve the best accuracy when more data can be achieved in all the numbers of the layers?

(1) Does a structure of 4 layers still achieve the same accuracy when more data can be achieved with that of this study?

It depends on the distribution of more data. If the distribution of more data is similar to that of the current data, in our opinion, the answer should be ”No. That is because various data from the same distribution can help model more easily learn predictive algorithms. However, as we found, the 4-layer structure has already learned the predictive algorithms best. More data might help the 4-layer structure learn better, which causes better accuracy. Precisely speaking, when the distribution of the more data is similar to that of the current data, a structure of 4 layers will achieve a no-worse accuracy. Besides, if the distribution of the more data is different from that of the current data, we regard this answer ”No” since the distribution of the whole data becomes more complicated and thus the same neural network cannot learn all the predictive algorithms. Usually, the predictive accuracy will become worse because the new distribution will confuse the model’s understanding and thus a more complicated structure, such as more layers, more neurons in layers, etc., will be needed.

(2) Does a structure of 4 layers still achieve the best accuracy when more data can be achieved in all the numbers of the layers?

As (1), it depends on the distribution of more data. If the distribution of more data is similar to that of the current data, in our opinion, a structure of a fewer number of layers could achieve the best predictive accuracy. That is because various data from the same distribution can help model learn the predictive algorithms more effectively and more efficiently, which decreases the difficulties for the model to learn and the necessity of a more layer structure, and more layers without regularization would just increase the non-linearity of models and simply overfit the dataset, to some extent. Although the overfit can be removed by the regularization, the regularization can just remove the unnecessary non-linearity predictive algorithm but cannot improve the predictive accuracy. As a result, the 4-layer or the-fewer-number-of-layer would lead to the best predictive accuracy. Moreover, if the distribution of the more data is different from that of the current data, 4-layer or the-more-number-of-layer could achieve the best predictive results when the distribution is more complicated to learn. A more complicated data distribution needs a more complicated neural network to learn. Therefore, 4-layer or the-more-number-of-layer could achieve the best predictive results. Whether 4-layer or the-more-number-of-layer could achieve the best predictive results will depend on the data distribution as well as the number of neurons in each layer in the neural network, respectively. Briefly, case by case, although it could be hard to simply conclude a structure of 4 layers still achieve the best accuracy if more data can be achieved, it usually needs more layers to acquire the predictive algorithms.

5.2 Discussion of Comparative Study on Algorithms of Multistep Prediction

This section describes the discussion of the second research. It is (2) to find the best algorithms of multistep prediction for flu outbreaks.

5.2.1 Past studies on Multistep prediction

We did not find past studies that performed auto-regression in the multistep prediction for flu outbreaks. Regarding multistep prediction for studies in other fields, MSP is one of the most popular methodologies probably because many types of models can be used for this purpose, such as linear regression, Support Vector Regression [172], Random Forest, Gradient Boosting, Multilayer Perception [173], and so on. However, any such model inevitably introduces errors and tends to suffer from error accumulation problem when the perdition period is long. This is because the bias and variance from previous predictions impact future predictions [168], and these compounding errors change the input distribution for future prediction steps, breaking the train-test independent and identically distributed (i.i.d) assumption common in supervised learning [168].

5.2.2 Comparison of the Accuracy of MSP, AMSP, MSOP, and MOP

When comparing four different multistep predicting algorithms, we found the MAPEs of AMSP were less than those of MSP, which demonstrated the adjusting algorithms of AMSP worked effectively. Besides, the MAPEs of MSOP are less than those of MOP. As we mentioned in the session of “Methods”, to predict the ILI rates of the coming 2nd -13th weeks, MOP trained only one model while MSOP trained 13 models. As a result, MSOP can predict with no necessity of sharing weights and neurons in LSTM structure; while MOP has to share weights and neurons in LSTM structure. Consequently, the accuracy of the MSOP performed better. Moreover, the average MAPEs of MSOP are slightly less than those of AMSP. The explanation is that MSOP does not accumulate errors at all while AMSP just adjusted its accumulated errors by training new models. Therefore, MSOP performed best.

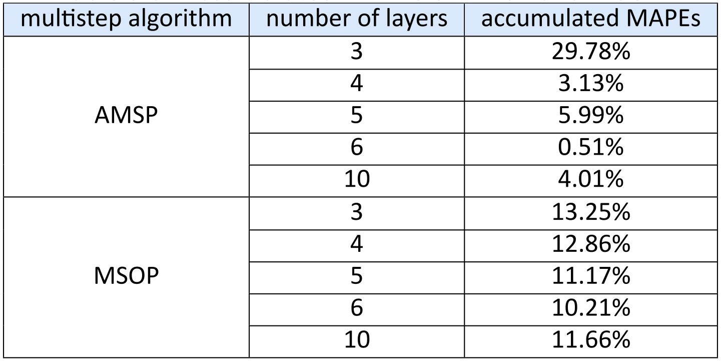
5.2.3 Accumulated MAPE

Table 5.3 shows the accumulated MAPEs of AMSP and MSOP. The accumulated MAPE is calculated in the whole testing set. One of the objectives of multistep prediction is to arrange a vaccine manufacturing plan. The calculation of the accumulated MAPEs helps us to find the best algorithm for vaccine manufacturing, and we found that AMSP with 6 layers achieved the best accuracy, i.e. the accumulated MAPE of 0.51% in the whole testing set.

5.2.4 Other Probable Features

In fact, including other features in multistep predicting models impacts models’ accuracy positively and negatively. For one thing, when predicting future values, other features may help predict more

Table 5.3: Accumulated MAPEs of AMSP and MSOP



multistep algorithm	number of layers	accumulated MAPEs
AMSP	3	29.78%
	4	3.13%
	5	5.99%
	6	0.51%
	10	4.01%
MSOP	3	13.25%
	4	12.86%
	5	11.17%
	6	10.21%
	10	11.66%

accurately, especially at turning points, such as an abrupt decrease in temperature. For another thing, before forecasting future ILI rates, we need to forecast other features (e.g. we need weather forecast for temperature and humidity). The error in former prediction could enlarge the error in later prediction since we predict recursively. The mechanism is similar to MSP, which accumulates the error step by step. Whether the accuracy improves or deteriorates might depend on different data in different seasons from different countries. In this study, we only performed auto-regression based on two pieces of consideration. Firstly, we regard historic values as a response to all related features, such as temperature, humidity, and so on. Therefore, to some extent, taking historic values as feature space includes all related features in models. Besides, how to include temperature or humidity of the whole country in models is a challenging job. Simply averaging temperature or humidity of all the places (cities and towns) of the United States might bring other problems, such as overlooking in population size, population density, lifestyles, and so on. in different places. For this topic, we need another research. We performed an additional experiment and compared the results in Section 5.3.6.

5.3 Discussion of Multistep Geolocational-Temporal Prediction of Worldwide Flu Outbreaks

This section describes the discussion of the third research. It is (3) to build up an effective and efficient approach to perform geolocational-temporal multistep of flu.

5.3.1 Source of Source Data

The data of Flunet are provided remotely by National Influenza Centres (NICs) of the Global Influenza Surveillance and Response System (GISRS) and other national flu reference laboratories collaborating actively with GISRS, or are uploaded from World Health Organization regional databases. [171] Take Japan as an example. In Japan, Infectious Disease Surveillance Center (IDSC) of National Institute of Infectious Diseases (NIID) is notified the results of isolation/detection of infectious agents from prefectural and municipal public health institutes (PHIs). The notified results are based on the laboratory identification done by PHIs for the specimens collected at sentinel clinics/ hospitals under the National Epidemiological Surveillance of Infectious Diseases (NESID), occasionally at non-sentinel sites and at health centers. [174] However, some flu patients might not go to clinics or hospitals but stay at home. Therefore, the flu-data collection may have some deviation. As a result, a part of the whole error of the multistep perdition may not be a predictive error from the models but a source-data deviation from the data collection.

5.3.2 Verification of Source Data

To verify the source data, we took the flu data in the United States as an example. We achieved the flu data from the World Health Organization and Centers for Disease Control and Prevention and compared their correlation (Figures 5.3.2 and 5.3.2). The ”R-score” was 0.8760, which shows the number of ILI from Centers for Disease Control and Prevention highly correlates to the number of flu patients from the World Health Organization. Moreover, the source data may have statistical variance due to the collective methodologies. Here we took Japan as an example. Luckily, we found Japan flu data from the four sources, and then we scraped the flu data of Japan from the four source URLs [175–178]. It seemed that the statistical data was being modified since the technique examination for flu needs to be updated. One can easily find the numbers from different sources cannot accord with each other. That means that there is a statistical variance due to the collective methodologies. This phenomenon can also make prediction inaccurate, to some extent. The basic reason for the change of the source data was that flu virus determination needs a duration of time. Usually, whether a patient is infected with the flu virus cannot be detected in a real-time manner. Besides, there is unavoidable false positive and false negative in flu virus detection. As a result, the aggregated number changes even after several months. Unfortunately, we were using the most recent year for testing dataset. Thereby, the predictive error may not only come from the models' construction but also result from the ground truth's error. How do we conquer this problem? Using the flu data several years ago when the data may not be revised any more could be an effective solution. However, we just want to mention here, the flu data several years ago still have some aggregation error although the probabilities and the number of errors could be better since they had been revised for several times. Another drawback could be the model learning flu data several years ago might not be able to learn the recent characteristics of the flu outbreaks and thus harder for recent use. As we discuss in the first chapter, the objective of the series of these researches was to find a pragmatic way for human being s to prevent flu outbreaks. Using relatively obsolete data seems a little bit of useless for the objective. Moreover, due to some problems, such as there were not so many flu data, the current model needs more data to train. The cut of the flu data of recent years will make insufficient training worse. In conclusion, using the flu data several years ago might bring merit and demerit. What is more, we might use advanced predictive algorithms, such as graph embedding, to decrease the necessity of enough data for training. However, these methodologies are still being developed, and the real effect of these advanced methodologies still needs to be researched in real-world datasets.

5.3.3 Explanation of Results

In brief, the results of the third research were concluded as follows:

(1) in the Southern Hemisphere, the MAPEs of prediction without flu data of other countries were lower than those with other countries

(2) in the Northern Hemisphere, the MAPEs of prediction without flu data of other countries were lower than those with other countries, when we performed 1-week-ahead prediction.

(3) in the Northern Hemisphere, the MAPEs of prediction without flu data of other countries were higher than those with other countries, when we performed 2-week-ahead, 3-week-ahead, 4-

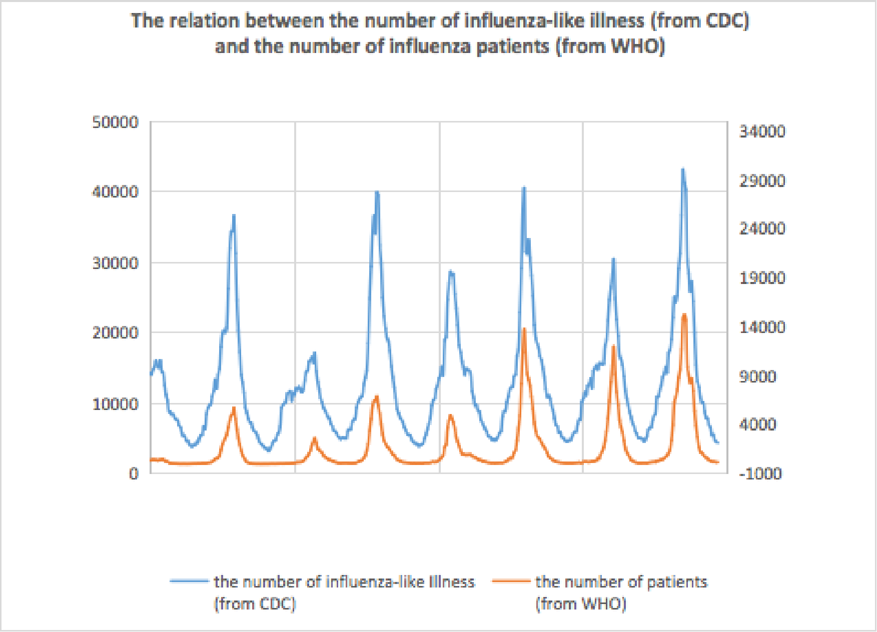


Figure 5.6: The comparison between the number of ILI from Centers for Disease Control and Prevention and the number of flu patients from the World Health Organization.

This figure compares the trend between the number of flu patients from World Health Organization and the number of ILI from CDC.

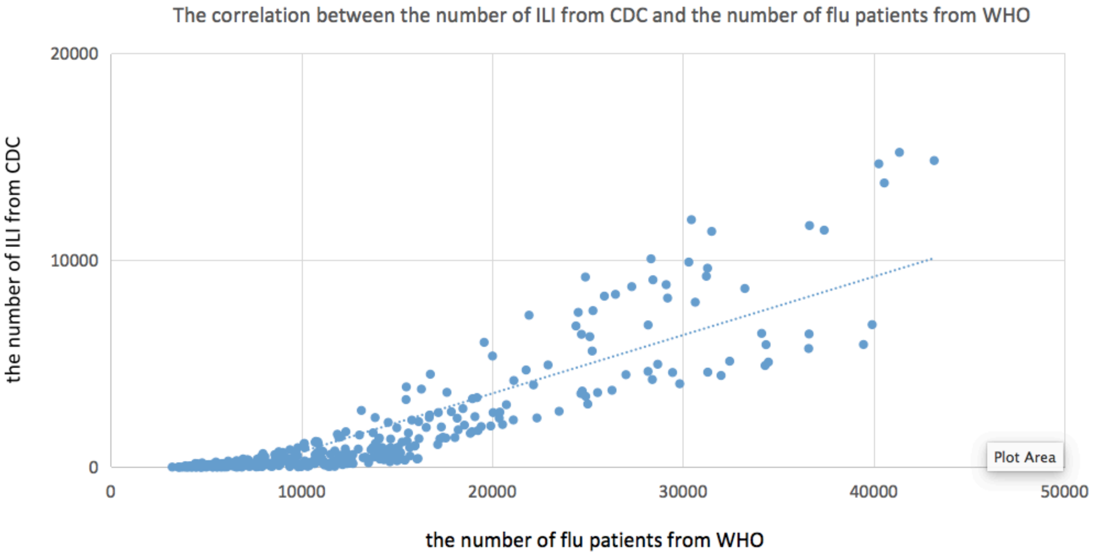
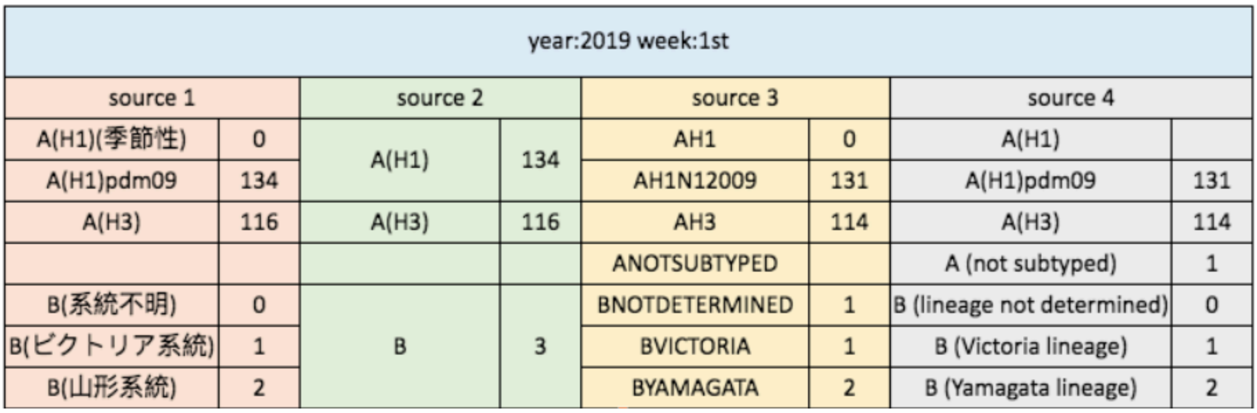


Figure 5.7: The correlation between the number of ILI from Centers for Disease Control and Prevention and the number of flu patients from the World Health Organization.

The ”R-score” was 0.8760. Therefore, the number of flu patients from World Health Organization correlates to the number of ILI from CDC.



year:2019 week:1st							
source 1		source 2		source 3		source 4	
A(H1)(季節性)	0	A(H1)	134	AH1	0	A(H1)	
A(H1)pdm09	134			AH1N12009	131	A(H1)pdm09	131
A(H3)	116	A(H3)	116	AH3	114	A(H3)	114
				ANOTSUBTYPED		A (not subtyped)	1
B(系統不明)	0	B	3	BNOTDETERMINED	1	B (lineage not determined)	0
B(ビクトリア系統)	1			BVICTORIA	1	B (Victoria lineage)	1
B(山形系統)	2			BYAMAGATA	2	B (Yamagata lineage)	2

Figure 5.8: The collective variance among the flu data of Japan from the different source URLs.

The flu data of Japan were scraped from the four source URLs. We can find the numbers from different sources cannot accord with each other. Therefore, we conclude there is a statistical variance due to the collective methodologies, which also made the predictive model inaccurate.

week-ahead prediction.

Practically, the probable reason that the MAPEs of predicting flu data with flu data of other countries are higher than the MAPEs without flu data of other countries in the Southern Hemisphere (i.e. Australia and Brazil) is countries in the Southern Hemisphere have different flu seasons, and the countries, the historical flu data of which were selected as geolocational-temporal factors in this study, are mostly in the Northern Hemisphere and their flu data are barely correlated to the flu data of the countries in the Southern Hemisphere. As for the Northern Hemisphere, the MAPEs of predicting flu data of 2nd, 3rd, and 4th week ahead with flu data of other countries were lower than those without other countries. That is countries in the Northern Hemisphere share similar flu seasons. However, the MAPEs of predicting flu data of the 1st week ahead with flu data of other countries were lower than those without other countries. That is probably because flu infection among countries is supposed to have a time lag resulting from a geographical distance. Theoretically, to find a reasonable explanation of the performance of the LSTM neural nets in the third research, we tried performing another supplement experiment. We divided the extra-experiment into two subparts: (1) qualitative analytics and (2) quantitative analytics.

(a) Qualitative Analytics

As the first part of the extra-experiment, we explored the correlation between the flu data of 1 week ahead, 2 weeks ahead, 3 weeks ahead, and 4 weeks ahead of Australia, China, and United States and the current flu data of all the 22 countries. Table 5.4 presents the correlation coefficients between the flu data of 1 week ahead and 2 weeks ahead of Australia and the current flu data of all the 22 countries. Table 5.5 presents the correlation coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of Australia and the current flu data of all the 22 countries. In each table, all the 22 countries were ranked by the absolute values of correlation coefficients. In the 1-week-ahead (Table 5.4), 2-week-ahead (Table 5.4), 3-week-ahead (Table 5.5), 4-week-ahead correlation (Table 5.5) of Australia, the second largest correlation coefficients (Indonesia, Indonesia, Indonesia, Republic_of_Korea) were all less than 0.3. In statistics, correlation coefficients of (0.00,

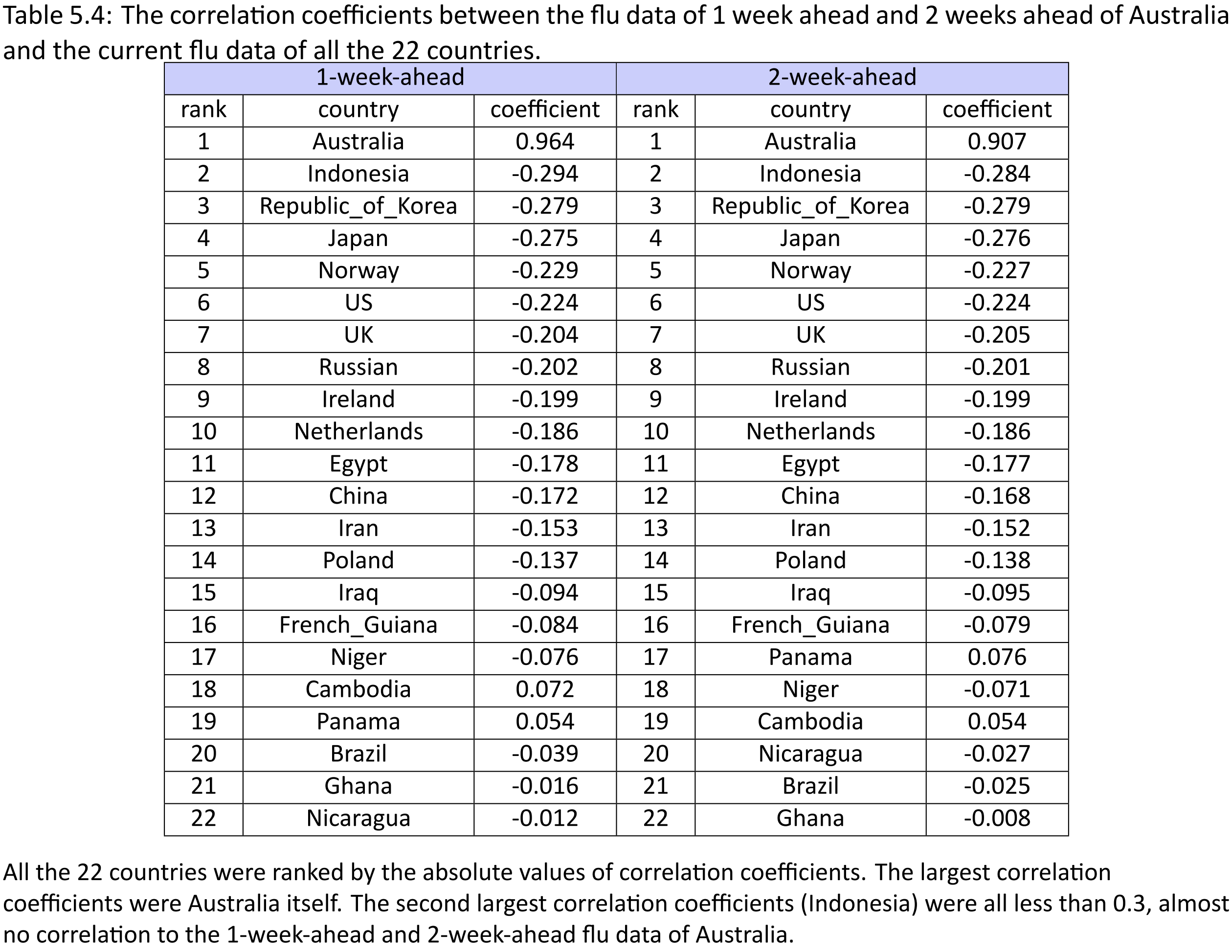


Table 5.4: The correlation coefficients between the flu data of 1 week ahead and 2 weeks ahead of Australia and the current flu data of all the 22 countries.

1-week-ahead			2-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	Australia	0.964	1	Australia	0.907
2	Indonesia	-0.294	2	Indonesia	-0.284
3	Republic_of_Korea	-0.279	3	Republic_of_Korea	-0.279
4	Japan	-0.275	4	Japan	-0.276
5	Norway	-0.229	5	Norway	-0.227
6	US	-0.224	6	US	-0.224
7	UK	-0.204	7	UK	-0.205
8	Russian	-0.202	8	Russian	-0.201
9	Ireland	-0.199	9	Ireland	-0.199
10	Netherlands	-0.186	10	Netherlands	-0.186
11	Egypt	-0.178	11	Egypt	-0.177
12	China	-0.172	12	China	-0.168
13	Iran	-0.153	13	Iran	-0.152
14	Poland	-0.137	14	Poland	-0.138
15	Iraq	-0.094	15	Iraq	-0.095
16	French_Guiana	-0.084	16	French_Guiana	-0.079
17	Niger	-0.076	17	Panama	0.076
18	Cambodia	0.072	18	Niger	-0.071
19	Panama	0.054	19	Cambodia	0.054
20	Brazil	-0.039	20	Nicaragua	-0.027
21	Ghana	-0.016	21	Brazil	-0.025
22	Nicaragua	-0.012	22	Ghana	-0.008

All the 22 countries were ranked by the absolute values of correlation coefficients. The largest correlation coefficients were Australia itself. The second largest correlation coefficients (Indonesia) were all less than 0.3, almost no correlation to the 1-week-ahead and 2-week-ahead flu data of Australia.

0.30] presents almost no linear relationship; correlation coefficients of (0.30, 0.50] presents a weak linear relationship; correlation coefficients of (0.50, 0.70] presents a moderate linear relationship; correlation coefficients of (0.70, 1.00) presents a strong linear relationship; and correlation coefficients of 1.00 presents an exact linear relationship. This phenomenon tells us that only the past flu data of Australia itself helps perform 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead prediction. This phenomenon explains one of the performances of the LSTM neural nets, why in the Southern Hemisphere MAPEs of prediction without flu data of other countries were lower than those with other countries.

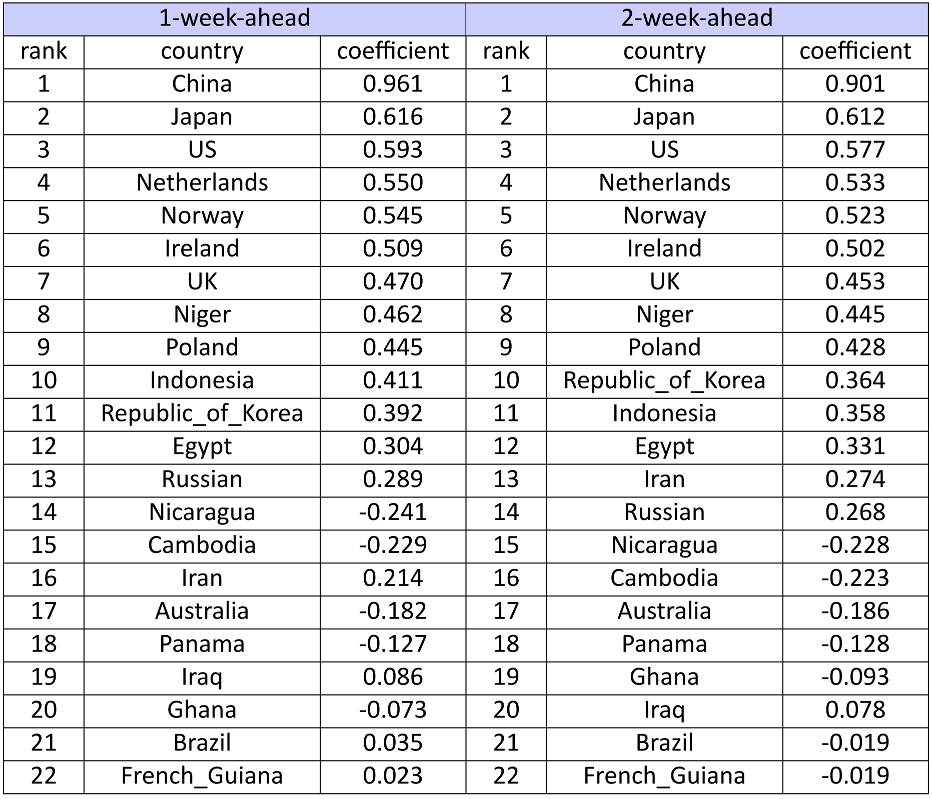
Table 5.6 presents the correlation coefficients between the flu data of 1 week ahead and 2 weeks ahead of China and the current flu data of all the 22 countries. Table 5.7 presents the correlation coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of China and the current flu data of all the 22 countries. In each table, all the 22 countries were ranked by the absolute values of correlation coefficients. In the 1-week-ahead (Table 5.6), 2-week-ahead (Table 5.6), 3-week-ahead (Table 5.7), 4-week-ahead correlation (Table 5.7) of China, many countries showed strong or week correlations. Table 5.14 shows there were 12, 13, 13, and 11 countries, current flu data of which had strong or week correlations (correlation coefficient > 0.3) to the 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead flu data of China. Table 5.8 presents the correlation coefficients between the flu data of 1 week ahead and 2 weeks ahead of the US and the current flu data of all the 22 countries. Table 5.9 presents the correlation coefficients between the flu data of 3 weeks ahead and

Table 5.5: The correlation coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of Australia and the current flu data of all the 22 countries.

3-week-ahead			4-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	Australia	0.819	1	Australia	0.710
2	Indonesia	-0.277	2	Republic_of_Korea	-0.276
3	Republic_of_Korea	-0.277	3	Japan	-0.274
4	Japan	-0.275	4	Indonesia	-0.268
5	Norway	-0.224	5	US	-0.222
6	US	-0.223	6	Norway	-0.220
7	UK	-0.205	7	Russian	-0.203
8	Russian	-0.202	8	UK	-0.203
9	Ireland	-0.198	9	Ireland	-0.195
10	Netherlands	-0.186	10	Netherlands	-0.186
11	Egypt	-0.177	11	Egypt	-0.174
12	China	-0.165	12	China	-0.165
13	Iran	-0.147	13	Iran	-0.143
14	Poland	-0.138	14	Panama	0.142
15	Panama	0.105	15	Poland	-0.140
16	Iraq	-0.094	16	Iraq	-0.094
17	Niger	-0.074	17	Niger	-0.079
18	French_Guiana	-0.061	18	Nicaragua	-0.047
19	Cambodia	0.041	19	French_Guiana	-0.039
20	Nicaragua	-0.037	20	Cambodia	0.024
21	Brazil	-0.011	21	Ghana	0.016
22	Ghana	0.003	22	Brazil	0.009

All the 22 countries were ranked by the absolute values of correlation coefficients. The largest correlation coefficients were Australia itself. The second largest correlation coefficients (Indonesia and Republic_of_Korea) were all less than 0.3, almost no correlation to the 1-week-ahead and 2-week-ahead flu data of Australia.

Table 5.6: The correlation coefficients between the flu data of 1 week ahead and 2 weeks ahead of China and the current flu data of all the 22 countries.



1-week-ahead			2-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	China	0.961	1	China	0.901
2	Japan	0.616	2	Japan	0.612
3	US	0.593	3	US	0.577
4	Netherlands	0.550	4	Netherlands	0.533
5	Norway	0.545	5	Norway	0.523
6	Ireland	0.509	6	Ireland	0.502
7	UK	0.470	7	UK	0.453
8	Niger	0.462	8	Niger	0.445
9	Poland	0.445	9	Poland	0.428
10	Indonesia	0.411	10	Republic_of_Korea	0.364
11	Republic_of_Korea	0.392	11	Indonesia	0.358
12	Egypt	0.304	12	Egypt	0.331
13	Russian	0.289	13	Iran	0.274
14	Nicaragua	-0.241	14	Russian	0.268
15	Cambodia	-0.229	15	Nicaragua	-0.228
16	Iran	0.214	16	Cambodia	-0.223
17	Australia	-0.182	17	Australia	-0.186
18	Panama	-0.127	18	Panama	-0.128
19	Iraq	0.086	19	Ghana	-0.093
20	Ghana	-0.073	20	Iraq	0.078
21	Brazil	0.035	21	Brazil	-0.019
22	French_Guiana	0.023	22	French_Guiana	-0.019

All the 22 countries were ranked by the absolute values of correlation coefficients. The largest correlation coefficients were China itself. The second to largest correlation coefficients (from Japan to Egypt) were all larger than 0.3. Flu data of any countries have correlations to the 1-week-ahead and 2-week-ahead flu data of China.

4 weeks ahead of the US and the current flu data of all the 22 countries. In each table, all the 22 countries were ranked by the absolute values of correlation coefficients. In the 1-week-ahead (Table 5.8), 2-week-ahead (Table 5.8), 3-week-ahead (Table 5.9), 4-week-ahead correlation (Table 5.9) of US, many countries showed strong or week correlations. Table 5.14 shows there were 10, 10, 10, and 11 countries, current flu data of which had strong or week correlations (correlation coefficient > 0.3) to the 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead flu data of US.

(b) Quantitative Analytics

In Table 5.6, Table 5.6, Table 5.8, and Table 5.8, we found the current flu data of other countries have correlation to the 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead flu data of China and US, both of which locate in Northern Hemisphere. These qualitative results introduce further research on how much the current flu data of other countries impact the 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead flu data of China and the US. That is quantitative research, as follows. Table 5.10 presents the linear coefficients between the flu data of 1 week ahead and 2 weeks ahead of the US and the current flu data of all the 22 countries. Table 5.11 presents the linear coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of the US and the current flu data of all the 22 countries. To compare the impact from different countries, we scaled all the flu data so that the comparison would avoid the impact from the order of magnitude of the flu data. Table 5.15 presents the times between the largest linear coefficients (China itself) and the second largest linear

Table 5.7: The correlation coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of China and the current flu data of all the 22 countries.

3-week-ahead			4-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	China	0.820	1	China	0.728
2	Japan	0.601	2	Japan	0.584
3	US	0.540	3	US	0.484
4	Netherlands	0.499	4	Netherlands	0.451
5	Norway	0.483	5	Ireland	0.427
6	Ireland	0.478	6	Niger	0.426
7	Niger	0.439	7	Norway	0.425
8	UK	0.421	8	UK	0.375
9	Poland	0.391	9	Iran	0.343
10	Egypt	0.345	10	Egypt	0.340
11	Iran	0.327	11	Poland	0.339
12	Republic_of_Korea	0.327	12	Republic_of_Korea	0.283
13	Indonesia	0.307	13	Indonesia	0.268
14	Russian	0.245	14	Russian	0.217
15	Cambodia	-0.212	15	Australia	-0.199
16	Nicaragua	-0.205	16	Cambodia	-0.196
17	Australia	-0.193	17	Nicaragua	-0.178
18	Panama	-0.127	18	Panama	-0.123
19	Ghana	-0.100	19	Ghana	-0.108
20	Iraq	0.087	20	Brazil	-0.107
21	Brazil	-0.067	21	Iraq	0.097
22	French_Guiana	-0.053	22	French_Guiana	-0.090

All the 22 countries were ranked by the absolute values of correlation coefficients. The largest correlation coefficients were China itself. The second to largest correlation coefficients (from Japan to Indonesia; from Japan to Poland) were all larger than 0.3. Flu data of any countries have correlations to the 3-week-ahead and 4-week-ahead flu data of China.

Table 5.8: The correlation coefficients between the flu data of 1 week ahead and 2 weeks ahead of United States and the current flu data of all the 22 countries.

1-week-ahead			2-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	US	0.714	1	US	0.812
2	Netherlands	0.643	2	Netherlands	0.695
3	UK	0.582	3	UK	0.605
4	Norway	0.571	4	Norway	0.599
5	Japan	0.552	5	Japan	0.571
6	Ireland	0.502	6	Poland	0.552
7	Poland	0.484	7	China	0.499
8	China	0.447	8	Ireland	0.489
9	Niger	0.392	9	Niger	0.442
10	Russian	0.338	10	Russian	0.362
11	Egypt	0.224	11	Indonesia	0.247
12	Australia	-0.222	12	Australia	-0.225
13	Indonesia	0.201	13	Egypt	0.202
14	Iran	0.160	14	Nicaragua	-0.160
15	Republic_of_Korea	0.151	15	Republic_of_Korea	0.159
16	Cambodia	-0.137	16	Cambodia	-0.154
17	Nicaragua	-0.137	17	Iran	0.142
18	Ghana	-0.129	18	Ghana	-0.137
19	Iraq	0.092	19	Iraq	0.097
20	Panama	-0.076	20	Panama	-0.082
21	Brazil	-0.069	21	French_Guiana	-0.047
22	French_Guiana	-0.065	22	Brazil	-0.032

All the 22 countries were ranked by the absolute values of correlation coefficients. The largest correlation coefficients were the United States itself. The second to largest correlation coefficients (from the Netherlands to Russian) were all larger than 0.3. Flu data of any countries have correlations to the 1-week-ahead and 2-week-ahead flu data of the United States.

Table 5.9: The correlation coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of United States and the current flu data of all the 22 countries.

3-week-ahead			4-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	US	0.902	1	US	0.970
2	Netherlands	0.730	2	Netherlands	0.742
3	Norway	0.616	3	Norway	0.620
4	UK	0.615	4	UK	0.613
5	Poland	0.593	5	Poland	0.595
6	Japan	0.574	6	China	0.573
7	China	0.542	7	Japan	0.564
8	Ireland	0.466	8	Ireland	0.437
9	Niger	0.447	9	Niger	0.412
10	Russian	0.372	10	Russian	0.378
11	Indonesia	0.297	11	Indonesia	0.342
12	Australia	-0.225	12	Australia	-0.224
13	Egypt	0.182	13	Nicaragua	-0.188
14	Nicaragua	-0.176	14	Cambodia	-0.185
15	Cambodia	-0.170	15	Republic_of_Korea	0.169
16	Republic_of_Korea	0.164	16	Egypt	0.165
17	Ghana	-0.139	17	Ghana	-0.139
18	Iran	0.126	18	Iraq	0.129
19	Iraq	0.110	19	Iran	0.110
20	Panama	-0.088	20	Panama	-0.094
21	French_Guiana	-0.033	21	Brazil	0.049
22	Brazil	0.007	22	French_Guiana	-0.024

All the 22 countries were ranked by the absolute values of correlation coefficients. The largest correlation coefficients were the United States itself. The second to largest correlation coefficients (from the Netherlands to Russian; from the Netherlands to Indonesia) were all larger than 0.3. Flu data of any countries have correlations to the 3-week-ahead and 4-week-ahead flu data of the United States.

coefficients. As we increase the number of steps of prediction, the time decreased dramatically, from 10 to 2. Table 5.12 presents the linear coefficients between the flu data of 1 week ahead and 2 weeks ahead of the US and the current flu data of all the 22 countries. Table 5.13 presents the linear coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of the US and the current flu data of all the 22 countries. Table 5.15 presents the times between the largest linear coefficients (the US itself) and the second largest linear coefficients. As we increase the number of steps of prediction, the time decreased dramatically, from 13 to 2. In Table 5.15, we found a magnitude gap between the times between the largest coefficient and the second largest coefficient of the 1 week ahead and those of the 2, 3 and 4 weeks, ahead. Both of the times between the largest coefficient and the second largest coefficient of the 1 week ahead in China and US outnumber 9, an order of magnitude. In other words, when we perform 1-week-ahead flu data prediction, current flu data of other countries help improve predicting accuracy limitedly or even deteriorates the predicting accuracy due to the noise from the flu data in all the other countries, and when we perform 2-, 3-, and 4-week-ahead flu data prediction, current flu data of other countries help improve predicting accuracy more largely than current flu data of other countries do when we perform 1-week-ahead flu data prediction. This phenomenon explains the other two performances of the LSTM neural nets, why in the Northern Hemisphere, the MAPEs of prediction without flu data of other countries were lower than those with other countries, when we performed 1-week-ahead prediction; and the MAPEs of prediction without flu data of other countries were higher than those with other countries, when we performed 2-week-ahead, 3-week-ahead, 4-week-ahead prediction. In conclusion, by our extra qualitative and quantitative experiments, we answered the questions why the LSTM neural nets could perform the geolocational-temporal multistep prediction of flu data and why the LSTM neural nets showed the three performances in the third research.

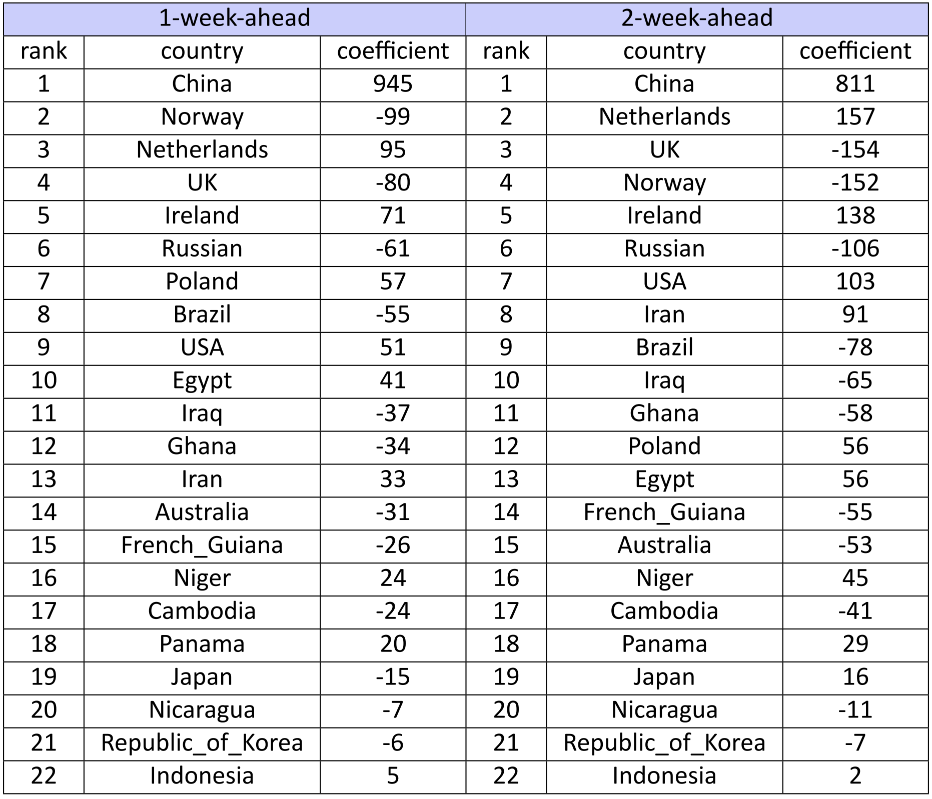
5.3.4 One-Step-Ahead Prediction of Flu Data in UK

As for the UK, even when we performed 1-step-ahead prediction, the predictive MAPEs without flu data of other countries were higher than those with other countries. That was probably because of the rapidly increasing number of travelers in 2017 and 2018. UK tourism is set to see a record-breaking 2018. Actually, in 2017, UK tourism just saw record highs, with overseas visitors reaching 40.3 million [179]. This was an increase of 4.6 percent from 38.5 million in 2016 [179]. However, the population of the whole UK was only around 65.6 million in 2016. In other words, the number of foreign tourists was around two-thirds of the whole population of the UK. This phenomenon can also explain why the MAPEs of UK were relatively high (some even over 110%) while the MAPEs of all other countries were relatively low (no more than 55%). Table 5.16 shows the number of tourists and the population in the six countries in the third study. The data were from [180–184]. We found that the tourism ratio in the UK was extremely high when we compare the data of the other countries in Table 5.16.

5.3.5 High MAPEs

In this study, The best MAPEs of LSTM models achieved were still very high. The probable reason is that we used the flu data in 2017-2018 as a testing set. The 2017-2018 flu season, a pandemic-like season, was quite different from and seriously heavier than the past few seasons, and other machine learning meteorologies, such as SVR and RF, also resulted in high MAPEs in this study.

Table 5.10: The linear coefficients between the flu data of 1 week ahead and 2 weeks ahead of China and the current flu data of all the 22 countries.



1-week-ahead			2-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	China	945	1	China	811
2	Norway	-99	2	Netherlands	157
3	Netherlands	95	3	UK	-154
4	UK	-80	4	Norway	-152
5	Ireland	71	5	Ireland	138
6	Russian	-61	6	Russian	-106
7	Poland	57	7	USA	103
8	Brazil	-55	8	Iran	91
9	USA	51	9	Brazil	-78
10	Egypt	41	10	Iraq	-65
11	Iraq	-37	11	Ghana	-58
12	Ghana	-34	12	Poland	56
13	Iran	33	13	Egypt	56
14	Australia	-31	14	French_Guiana	-55
15	French_Guiana	-26	15	Australia	-53
16	Niger	24	16	Niger	45
17	Cambodia	-24	17	Cambodia	-41
18	Panama	20	18	Panama	29
19	Japan	-15	19	Japan	16
20	Nicaragua	-7	20	Nicaragua	-11
21	Republic_of_Korea	-6	21	Republic_of_Korea	-7
22	Indonesia	5	22	Indonesia	2

All the 22 countries were ranked by the absolute values of correlation coefficients.

5.3.6 Weather Features

Adding weather features into multistep predicting models may impact models' accuracy positively and negatively. For one thing, when predicting future values of flu, weather features may help predict more accurately, especially at turning points, such as an abrupt decrease in temperature. For another thing, before forecasting future flu data, we need to forecast weather, such as temperature and humidity. The error in forecasting (former prediction) could enlarge the error in later prediction. In brief, whether the accuracy improves or deteriorates might depend on different data in different seasons from different countries. To rapidly explore the impact of adding other features, we conducted an additional experiment to see if the temperature and humidity would improve the accuracy of the models. We took the country of Japan as an analytical experiment. Here were the steps of this additional experiment.

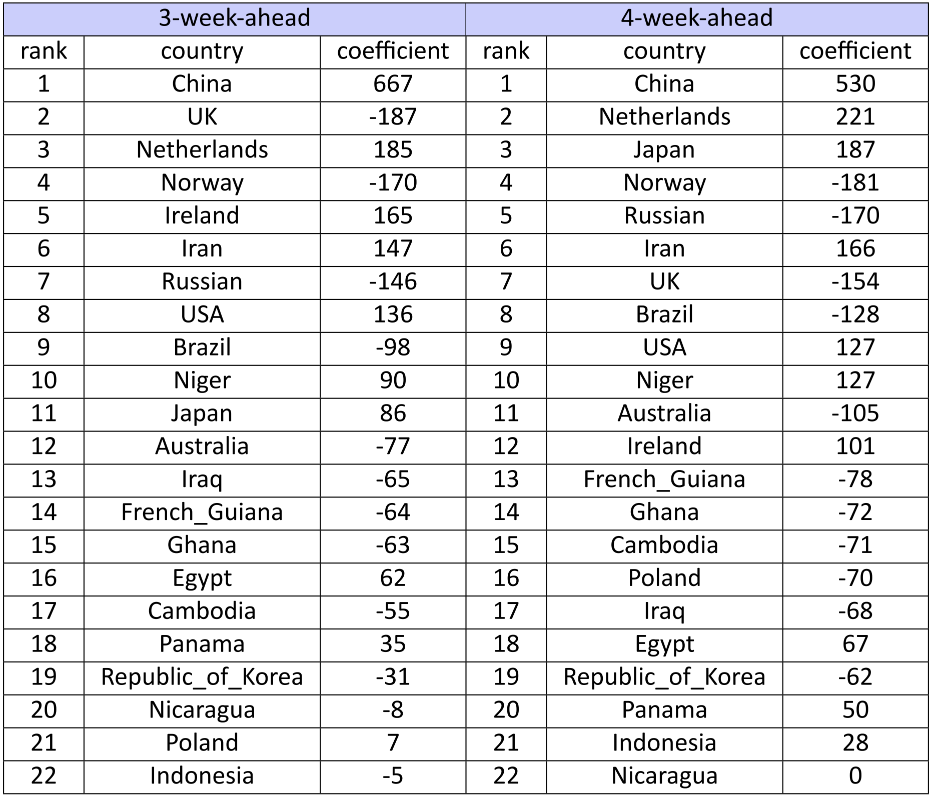
(1) We selected the TOP 10 cities, which have the largest populations of Japan, to represent the whole of Japan. They were Tokyo, Osaka, Yokohama, Nagoya, Sapporo, Kobe, Kyoto, and Fukuoka.

(2) We scraped the daily temperature and humidity of the 10 cities from Japan Meteorological Agency [185] from Dec 1, 2009, to Jan 31, 2019, by Python.

(3) We took the mean of the daily temperature and humidity of the 10 cities, to represent the whole country's temperature and humidity.

(4) We aggregated the daily mean temperature and humidity of the whole country into the weekly mean. Table 5.17 shows the image of the processed data.

Table 5.11: The linear coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of China and the current flu data of all the 22 countries.



3-week-ahead			4-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	China	667	1	China	530
2	UK	-187	2	Netherlands	221
3	Netherlands	185	3	Japan	187
4	Norway	-170	4	Norway	-181
5	Ireland	165	5	Russian	-170
6	Iran	147	6	Iran	166
7	Russian	-146	7	UK	-154
8	USA	136	8	Brazil	-128
9	Brazil	-98	9	USA	127
10	Niger	90	10	Niger	127
11	Japan	86	11	Australia	-105
12	Australia	-77	12	Ireland	101
13	Iraq	-65	13	French_Guiana	-78
14	French_Guiana	-64	14	Ghana	-72
15	Ghana	-63	15	Cambodia	-71
16	Egypt	62	16	Poland	-70
17	Cambodia	-55	17	Iraq	-68
18	Panama	35	18	Egypt	67
19	Republic_of_Korea	-31	19	Republic_of_Korea	-62
20	Nicaragua	-8	20	Panama	50
21	Poland	7	21	Indonesia	28
22	Indonesia	-5	22	Nicaragua	0

All the 22 countries were ranked by the absolute values of correlation coefficients.

(5) Since in a realistic prediction, we need to use the weather forecast to predict the number of flu patients in the future. In our experiment, to simulate a real analytical way, we used ARIMA to predict temperature and humidity for the testing period of the predictive model for the number of flu patients.

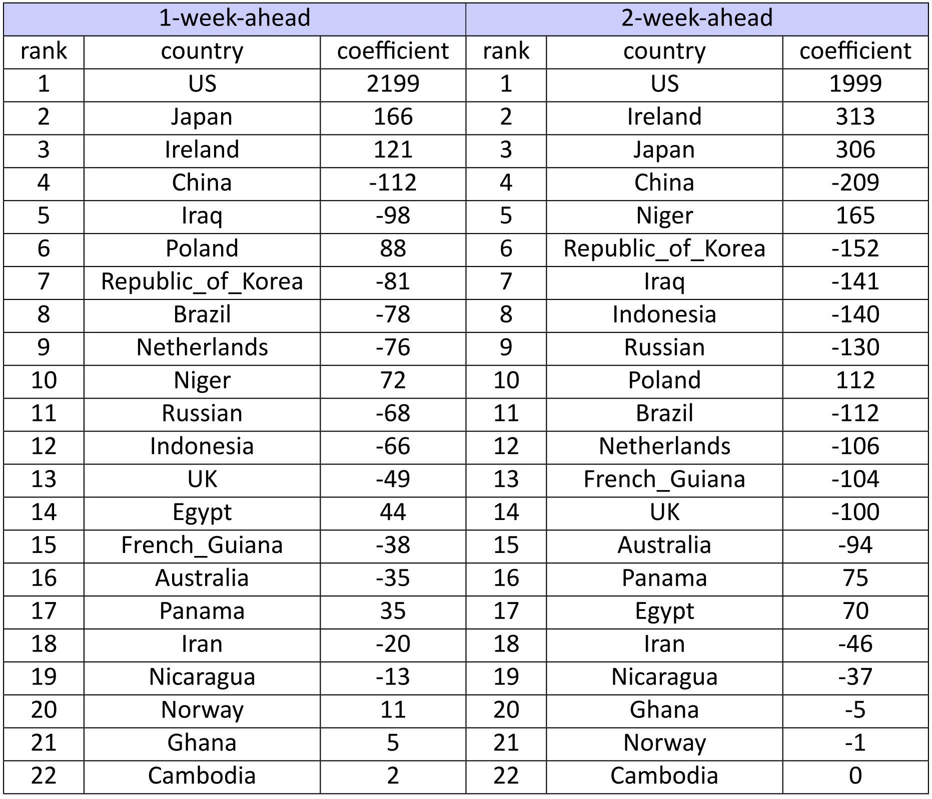
(6) We used the true weather data [from Step (4)] as well as the flu data of Japan to train a predictive model to forecast the number of flu patients. The only difference from the experiments in the 3rd research was that we added the true weather data of Japan.

(7) We used the predicted weather data [from Step (5)] as well as the flu data of Japan to predict the number of flu patients. The only difference from the experiments in the 3rd research was that in this additional experiment we added the predicted weather data of Japan into the models.

(8) We compared the predictive results. Table 5.18 shows the predictive accuracy with and without weather information and with and without the flu data of other countries. The highlighted cells were the best accuracy we achieved for the different multistep prediction. In Table 5.18, the best results were all from the models without adding weather data into the models. In other words, there was no improvement after we inputted weather (temperature and humidity) into the models as features.

The probable reason is that the historical data have already included the influential factors from weather condition since weather condition such as the temperature and humidity would not change dramatically year by year. For example, the temperatures and humidities in the 49th week of different years are usually similar. After we added the historical flu data into models, the models

Table 5.12: The linear coefficients between the flu data of 1 week ahead and 2 weeks ahead of United States and the current flu data of all the 22 countries.



1-week-ahead			2-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	US	2199	1	US	1999
2	Japan	166	2	Ireland	313
3	Ireland	121	3	Japan	306
4	China	-112	4	China	-209
5	Iraq	-98	5	Niger	165
6	Poland	88	6	Republic_of_Korea	-152
7	Republic_of_Korea	-81	7	Iraq	-141
8	Brazil	-78	8	Indonesia	-140
9	Netherlands	-76	9	Russian	-130
10	Niger	72	10	Poland	112
11	Russian	-68	11	Brazil	-112
12	Indonesia	-66	12	Netherlands	-106
13	UK	-49	13	French_Guiana	-104
14	Egypt	44	14	UK	-100
15	French_Guiana	-38	15	Australia	-94
16	Australia	-35	16	Panama	75
17	Panama	35	17	Egypt	70
18	Iran	-20	18	Iran	-46
19	Nicaragua	-13	19	Nicaragua	-37
20	Norway	11	20	Ghana	-5
21	Ghana	5	21	Norway	-1
22	Cambodia	2	22	Cambodia	0

All the 22 countries were ranked by the absolute values of correlation coefficients.

are, more or less, able to learn the trend that is regarded as a reflection of all relevant factors including temperature, humidity, and so on.

(a) Do the colder and drier weather help the spread and infection of the flu?

The flu season in the Northern Hemisphere usually begins as early as October, spreads in December, peaks in February, and ends in March. [186] It seems that flu will spread whenever winter lasts. Some previous studies found that flu viruses survived longer at low humidity and low temperatures. [187, 188] The lower the temperature is, the longer the flu viruses will survive. However, another previous study found a contradictory phenomenon. For one thing, it reported that colder and drier weather caused higher numbers of flu infections. For another, in geolocation with warmer climates, flu infection rates are positively correlated closely to high humidity and lots of rain. Therefore, the study concluded that, rather than the colder and drier weather, it was still unclear why the flu behaves so differently in disparate environments. [189] In fact, some other previous study argued that it was not cold temperatures that make the flu popular in winters. Instead, the research attested that the lack of sunlight or the different lifestyles in winter resulted in flu's population. The following are the most popular theories why the flu strikes in winter:

(a) During winters, people spend more time indoors with the windows sealed, so they are more likely to breathe the same air as someone who has the flu and thus contract the virus. [190]

(b) Days are shorter during the winter, and lack of sunlight leads to low levels of Vitamin D and melatonin, both of which require sunlight for their generation. This compromises our immune

Table 5.13: The linear coefficients between the flu data of 3 weeks ahead and 4 weeks ahead of United States and the current flu data of all the 22 countries.

3-week-ahead			4-week-ahead		
rank	country	coefficient	rank	country	coefficient
1	US	1694	1	US	1331
2	Japan	501	2	Japan	722
3	Ireland	406	3	Republic_of_Korea	-442
4	Republic_of_Korea	-289	4	Ireland	428
5	China	-244	5	China	-263
6	Indonesia	-196	6	French_Guiana	-250
7	French_Guiana	-181	7	Australia	-221
8	Niger	163	8	Brazil	-208
9	Australia	-161	9	Iraq	-190
10	Iraq	-161	10	Poland	-175
11	Brazil	-153	11	Indonesia	-165
12	Netherlands	-140	12	Egypt	134
13	Russian	-122	13	Panama	127
14	Panama	107	14	Russian	-102
15	Egypt	93	15	Norway	83
16	Iran	-75	16	Ghana	80
17	Norway	74	17	Iran	-80
18	Nicaragua	-51	18	UK	79
19	Ghana	25	19	Niger	63
20	UK	-18	20	Netherlands	-63
21	Poland	-17	21	Nicaragua	-49
22	Cambodia	-8	22	Cambodia	-32

All the 22 countries were ranked by the absolute values of correlation coefficients.

Table 5.14: The number of correlation coefficients that were larger than 0.3 in 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead flu data in China and US, which represents the countries in Northern Hemisphere.

hemisphere	countries	multistep ahead	the number of correlation >0.3
Southern	Australia	1	1
		2	1
		3	1
		4	1
Northern	China	1	12
		2	12
		3	13
		4	11
	US	1	10
		2	10
		3	10
		4	11

Table 5.15: The times between the largest linear coefficient and the second largest linear coefficient in 1-week-ahead, 2-week-ahead, 3-week-ahead, 4-week-ahead flu data in China and US, which represents the countries in Northern Hemisphere.

countries	multistep ahead	the times between the largest coefficient and the second largest coefficient
China	1	10
	2	5
	3	4
	4	2
US	1	13
	2	6
	3	3
	4	2

Table 5.16: The number of tourists and population in the six countries in the third study.

countries	tourists	population	ratio
Australia	7,428,600	25,240,300	29%
Brazil	6,626,000	204,519,000	3%
China	56,910,000	1,394,560,000	4%
Japan	19,704,000	126,891,000	16%
UK	40,300,000	66,040,229	61%
US	77,800,000	327,929,000	24%

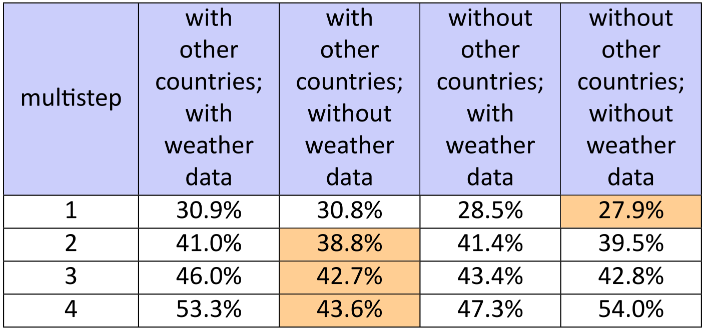
The table shows the number of tourists and population in the six countries: Australia, Brazil, China, Japan, the UK, and the US. The tourism ratio in the UK is extremely high when we compare the data of the other countries

Table 5.17: The weekly mean temperature and humidity of the whole Japan.

Year_Week_Number	Temperature	Humidity
200949	9.8	66.5
200950	9.6	63.5
200951	4.9	53.2
200952	6.2	60.7
200953	5.6	57.7
201001	5.0	55.2
201002	2.9	56.7
...	...	...
201850	5.7	65.2
201851	8.7	72.3
201852	5.1	57.6
201901	5.1	60.6
201902	4.9	60.1
201903	5.5	61.9
201904	5.4	55.4

The first column represents the year and the week sequence. For example, ”200949” means the 49th week of the year of 2009. The second column and third column represent the temperature and the humidity of the whole of Japan in the form of the weekly mean.

Table 5.18: The accuracy with/without the weather data (temperature and humidity) and with/without the flu data of the other countries.



multistep	with other countries; with weather data	with other countries; without weather data	without other countries; with weather data	without other countries; without weather data
1	30.9%	30.8%	28.5%	27.9%
2	41.0%	38.8%	41.4%	39.5%
3	46.0%	42.7%	43.4%	42.8%
4	53.3%	43.6%	47.3%	54.0%

This table shows the predictive accuracy with and without weather information and with and without the flu data of other countries. The highlighted cells were the best accuracy we achieved for the different multistep prediction.

systems, which in turn decreases the ability to fight the virus. [190]

(c) The flu virus may survive better in colder, drier climates, and therefore be able to infect more people. [190]

(b) Why does the flu season usually end at the end of the February or the beginning of the March?

There are two reasons, as follow.

(I) The number of people who may be infected with an epidemic will always decrease. Practically, the number of susceptible individuals falls rapidly as some of them are infected and thus enter the infectious compartments and some of them may get cured and thus enter the recovered compartments. Therefore, As a result, the number of suspectable people who may be infected with any epidemic will always decrease. Theoretically, by formula derivation of the SIR model, we can also get the conclusion that "The number of people who may be infected with any epidemic will always decrease."

(II) The infectious probabilities decreases. As aforementioned in Section 5.3.6, the following are the most popular theories why the flu strikes in winter:

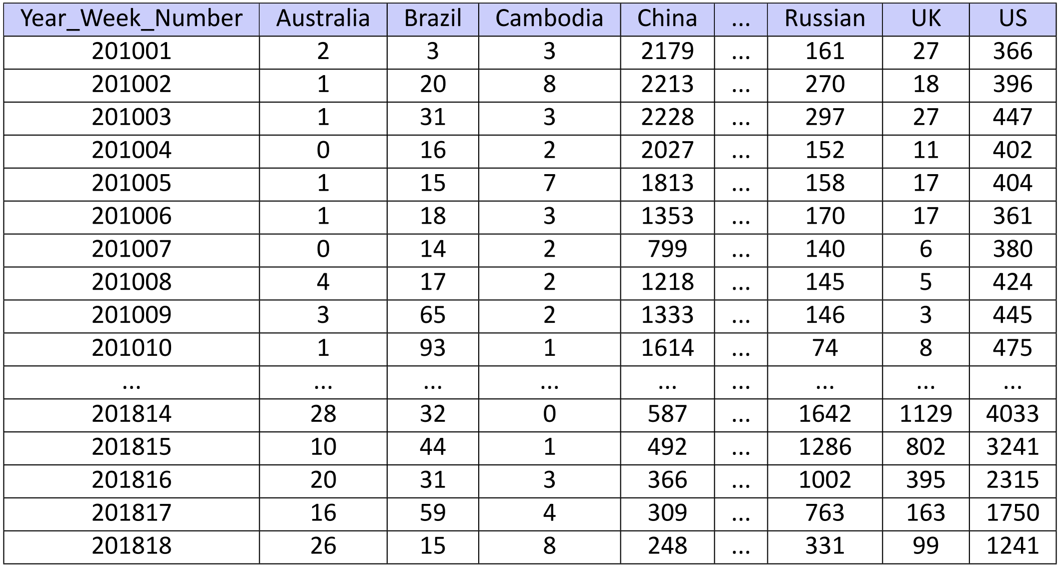
(a) During winters, people spend more time indoors with the windows sealed, so they are more likely to breathe the same air as someone who has the flu and thus contract the virus. [190]

(b) Days are shorter during the winter, and lack of sunlight leads to low levels of Vitamin D and melatonin, both of which require sunlight for their generation. This compromises our immune systems, which in turn decreases the ability to fight the virus. [190]

(c) The flu virus may survive better in colder, drier climates, and therefore be able to infect more people. [190]

When Marches come, days become longer and people spend more time outdoors and expose to more sunlight. Besides, the flu viruses also survive more hardly in warmer and more moister climates. In addition, people spend more time outdoors with the windows open. As a result, infectious probabilities decreased. In other words, not only do temperature and humidity change in winter, people's social interactions also change. Determining which factors are responsible for flu's increased prevalence in the winter months is very difficult and needs further studies.

Table 5.19: An image of the raw data used for training a predictive mode in the third study



Year_Week_Number	Australia	Brazil	Cambodia	China	...	Russian	UK	US
201001	2	3	3	2179	...	161	27	366
201002	1	20	8	2213	...	270	18	396
201003	1	31	3	2228	...	297	27	447
201004	0	16	2	2027	...	152	11	402
201005	1	15	7	1813	...	158	17	404
201006	1	18	3	1353	...	170	17	361
201007	0	14	2	799	...	140	6	380
201008	4	17	2	1218	...	145	5	424
201009	3	65	2	1333	...	146	3	445
201010	1	93	1	1614	...	74	8	475
...	...	...	...	...	...	...	...	...
201814	28	32	0	587	...	1642	1129	4033
201815	10	44	1	492	...	1286	802	3241
201816	20	31	3	366	...	1002	395	2315
201817	16	59	4	309	...	763	163	1750
201818	26	15	8	248	...	331	99	1241

5.3.7 Latitudes

The latitude could be another important factor influencing the flu popularity, as flu season behaves totally differently in temperate zone and tropic zone. Table 5.19 shows an image of the raw data used for training a predictive mode in the third study. For one thing, if we add latitudes of all countries as several columns, all the rows share the same data of latitudes, which is meaningless to model training since since there is no difference among the rows. For another, the interactive calculation in neural networks helps the model learn the correlation among data indirectly. In other words, the network structure have included the correlation of the latitudes although we did not add latitude directly. Although we cannot directly add latitudes as features into model, we may have an indirect way to show the effect of the latitude. Taking Egypt as an example since Egypt has the largest number of flu patients in all the low-latitude countries. Here were the steps of the additional experiment:

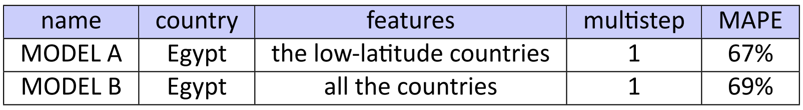
(a) We inputted the flu data of only the low-latitude countries (Egypt, French Guiana, Ghana, Indonesia, Nicaragua, Niger, and Panama) into a predictive model. Let us call this MODEL A.

(b) We inputted the flu data of all the countries (Australia, Brazil, Cambodia, China, Egypt, French Guiana, Ghana, Indonesia, Iran (Islamic Republic of), Iraq, Ireland, Japan, Netherlands, Nicaragua, Niger, Norway, Panama, Poland, Republic of Korea, Russian Federation, United Kingdom of Great Britain and Northern Ireland, and United States of America) into the predictive model. Let us call this MODEL B.

(c) The only difference between MODEL A and MODEL B is whether we input the flu data of the middle- and high-latitude countries into the predictive model. Therefore, if the accuracy of MODEL B is worse than that of MODEL A, we can say the fact of latitude works because inputting the flu data of middle- and high-latitude countries make the prediction worse. Otherwise, If the accuracy of MODEL B is no worse than that of MODEL A, we can say the method of inputting the flu data of all the countries into a deep learning model has already considered the importance of the latitudes since the neural network has learned the correlation among data.

(d) We compared the predictive accuracy.

Table 5.20: Comparison of the accuracy by inputting the number of patients of all countries and low-latitude countries



name	country	features	multistep	MAPE
MODEL A	Egypt	the low-latitude countries	1	67%
MODEL B	Egypt	all the countries	1	69%

(e) Table 5.20 shows the comparison. There is almost no significant difference between MODEL A and MODEL B.

As we discussed in Step (c) at least we can say the method of inputting the flu data of all the countries into a deep learning model has already considered the importance of the latitudes by neural network's learning of the correlation among data.

Chapter 6

Conclusion

6.1 Conclusion

In this Ph.D. research, we performed three studies on predicting flu outbreaks. We found a proper methodology to perform a geolocational-temporal prediction of flu outbreaks.

In the first research, we performed ARIMA, SVM, RF, GB, ANN, and LSTM models with different time lags (2, 4, 9, 13, 26, 52 weeks) to forecast the weekly ILI rate of U.S. flu data. We found the ARIMA, ANN and LSTM models with a lag time of 52 weeks (i.e., the periodicity of the flu season) resulted in the best MAPEs, while SVR, RF, and GB performed with almost no changes when we used the time lags. We also found the MAPEs of the machine learning models (SVR, RF, and GB) with the first differences were lower than those of ARIMA, and the MAPEs of the deep learning models (ANN and LSTM) with multiple layers were lower than those of the ML models (SVR, RF, and GB). To the best of our knowledge, this is the first time LSTM has been used to predict flu outbreaks. In all the models (with different model types, different hyperparameters, and different time lags), the LSTM model of 4 layers reached the lowest MAPE of 5.4%, and the LSTM model of 5 layers with regularization reached the lowest RMSE of 0.00210. Additionally, the LSTM models with 4 - 6 layers with regularization resulted in very low MAPEs of approximately 5.4% - 5.5% and more than 6 layers contributed little to improving the predictive accuracy.

In the second research, we adjusted the LSTM model by four multi-step prediction algorithms: MSP, AMSP, MSOP, MOP. Both MSP and AMSP are “recursive” prediction. Generally, the method of “recursive” predicts step-by-step: using predicted value to predict further values. For example, when performing the two-step-ahead prediction, ”recursive” method firstly predict 1-week-ahead value (i.e. a single step prediction) and then uses the predicted 1-week-ahead value to predict the 2-week-ahead value, and when predicting the 3-week-ahead value, the model will also use the predicted 1-week-ahead value and 2-week-ahead value. By recursive predicting, the models predict some-step-ahead values. The MSOP and MOP are “jumping” prediction. The method of “jumping” prediction only uses current and past values to predict some-step-ahead value directly instead of predicting step-by-step. The result showed that implementing MSOP in a 6-layer LSTM structure achieved the best accuracy. The MAPEs from 2-step-ahead to the 13-step-ahead prediction for the U.S. ILI rates were all less than 15%, averagely 12.930%.

In the third research, we developed a geolocational-temporal method to perform multi-step prediction of flu outbreaks. When we performed multistep prediction in the Northern Hemisphere, feeding those geolocational-temporal factors into models helped to improve the predictive accuracy; comparatively, when we performed prediction in the Southern Hemisphere, feeding the irrelevant geolocational-temporal factors from the Northern Hemisphere into the models exacerbate the predictive accuracy. Further experiments looked for an explanation. The fact is that the 22 countries were mostly in the Northern Hemisphere. Due to the high correlation among the flu data in the Northern Hemisphere, inputting the historical flu data of the 22 countries into the predictive models helped forecast flu data in the Northern Hemisphere, but exacerbate the predictive accuracy in the Southern Hemisphere since flu seasons in the Southern Hemisphere usually peaked in June, July, and August, totally different / unrelated to those in the Northern Hemisphere. Furthermore, the spread of flu among countries need some time more or less and thereby there should be a time lag, which explained the reason that the 1-week-ahead predictive MAPEs without feeding flu data of other countries were better than those with feeding flu data of other countries in the Northern Hemisphere. Also, as for the United Kingdom, the extremely high ratio of tourists (around two-thirds of the population of the United Kingdom) disrupted the time lag of flu spread. Thereby, even in the case of the 1-week-ahead prediction, inputting the flu data of other countries helped improve predictive accuracy. Therefore, we concluded feeding relevant geolocational-temporal factors in the same hemisphere helped to improve the predictive accuracy of the worldwide flu outbreaks.

Hopefully, these modeling approaches will positively help hospitals, pharmaceutical companies, individuals, and governments prepare better for the flu seasons and therefore help prevent and control flu outbreaks worldwide.

6.2 Originality

The originality of this study can be divided into two categories: (1) originality of methodology, and (2) originality of application:

(1) originality of methodology

In this part, the methodologies we leveraged in this study were firstly developed by ourselves. Almost all the originalities were developed for practical uses, and we hope these methods can help other researches more or less. The following list is the originality of methodology in details:

1. firstly explored different models as well as different hyperparameters for flu prediction
2. firstly explored different analytical methods, such as time lags, and so on.
3. firstly focused on the correlation of flu outbreaks among countries.

(2) the originality of application:

In this part, the application we leveraged in this study were firstly developed for flu prediction. The following list is the originality of application in details:

1. firstly performed a multistep prediction of flu outbreaks.
2. firstly used the two new algorithms (AMSP and MSOP) for multistep prediction of flu.
3. firstly performed a geolocational-temporal prediction for worldwide flu outbreaks.

In conclusion, as we discussed in Chapter 1, the main originality is to leverage frontier methodologies and develop new methodologies to perform wide-applicable accurate research for worldwide flu outbreaks since the model can effectively and efficiently help trillions of people. Our research is not designed to improve the current algorithm or technologies but explore a pragmatic approach to a real project all over the world. We believe these types of research are quite necessary since they aim at real-world problems.

6.3 Application

Hopefully, in our perspective of views, our research could help all countries better prepare for the annual flu outbreak. The multistep prediction is quite practical for hospitals and pharmaceutical manufacturers. For hospitals, we can use the predicted number to dynamically assign hospital beds of geolocational hospitals to flu patients. For pharmaceutical manufacturers, we can use the predicted number to formulate a dynamic manufacturing plan since the time for manufacturing the flu vaccine is very limited every year due to the high mutation rates of flu.

As we describe in Chapter 1, manufacturing flu vaccine is a dynamic time-consuming work because flu virus undergoes high mutation rates and frequent genetic re-assortment (combination and rearrangement of genetic material) [66–70]. In Februaries, World Health Organization assesses the strains of flu virus that are most likely to be circulating over the following winter. Every year, the first batch of vaccine is usually unavailable for the patients until every September [71]. Vaccine manufacturers can only produce flu vaccines in a very limited time [71]. Thus, how to arrange manufacturing plan is quite important for the timely delivery of flu vaccines as well as flu medicines. In other words, if we can, to some extent, precisely predict the number of flu patients, we can dynamically conceive manufacturing plan so that every ILI patients in a different location can gain the vaccine and medicine on time, which will help suppress the spread of flu outbreaks.

Besides, the prediction of the current research will help to construct a dynamic model to arrange patients to different hospitals to dynamically make the most of the source of the hospitalization. During flu peak periods, clinics and hospitals are overwhelmed. Beds assignment to flu patients in hospitals is a challenging task due to the limited capacity of hospital beds, time-dependencies of bed request arrivals, and unique treatment requirements of flu patients [72]. Besides, flu seasons vary in timing, severity, and duration from one season to another [71]. Therefore, flu hospitalization also varies by sites and time in each season [73], which makes beds assignment to flu patient more difficult for hospitals. This research is supposed to be the basis of a reinforcement learning to assign the source of the hospitalization in a city or an area since the number of flu patients, the number of beds, and the duration of flu hospitalization can be known by prediction of this study, the statistics of hospital sources, and the previous medical researches, respectively. This reinforcement learning model will be a future research topic and help humans to fight against flu outbreaks more directly. The integration of this reinforcement learning into the hospitalization system will be a quite necessary improvement for human beings.

6.4 Drawbacks of This Study

Although, we believe our study is useful and helpful. However, every research has merits and demerits. In our opinion, the drawbacks of this study are as follows.

Flu spread is not only a natural process but also a social phenomenon. Therefore, flu spread has too many impactors. However, in this study, we just used historical data as features. Historical data can help models understand periodicity and increasing or decreasing trend. However, historical data can hardly predict a turning point since there must be something happened at the turning point or just before the turning point. in the case of a flu outbreak, it could be hard to predict the first week when the flu outbreak occurs and the peak weak when the number of flu patients reaches most in some year. At the first week, there could be not enough medicine for flu patients. At the peak week, hospitals and drug stores may continue to increase stocking of flu medicine so that some wastes may happen to some extent. Including more features into the predictive model could be an effective way.

This study focused on the predictive accuracy and therefore we applied and refined deep learning models. However, the notable drawback of deep learning model is that a deep learning model lacks explainabilities. For example, in this study, we do not know if the short or the long weeks impact the flu spread in the coming weeks. In future studies, a more explainable model will be welcomed because a more explainable model can bring more hint and inspiration for human beings to design and improve solutions against flu outbreaks. Generally, developing a deep learning model with explainabilities is a popular trend. More and more models are leveraging technologies and skills, such as attention, graph, and so on. because these technologies and skills can help to explain the reason, and human society is a reasoning society. That is why the models' explainabilities is quite important.

We scraped the source data from all the countries. However, every country has a different range. For example, the range of Singapore is just around a city and the range of Russia looks like a continent. Simply connecting Singapore and Russia unbalances the granularity of data collection. Ideally, the granularity of data collection can depend on an area range or a population's density. If we can have the data, the granularity of which is on an area range or a population's density, a better predictive model can be achieved. Moreover, we can combine the data on area range and a population's density and learn the relation between flu outbreaks and area and population.

In the past years for the Ph.D. studies, we leveraged the models that were popular at that time. However, more and more advanced analytical algorithms have been developed, such as attention, embedding, reinforcement learning, and so on. In future studies, we might leverage and refine those algorithms to improve the accuracy and explainabilities of the predictive models. As almost every previous study did, a Ph.D. study needs a duration, and when a candidate concludes his or her research contents, most of the contents could be 2-3 years ago. It is quite hard for a candidate to closely follow every aspect of a specific research area.

6.5 Future Studies

This section describes the future work of flu prediction. After the Ph.D. studies, we will put more efforts to leverage and refine the cutting-edge technologies and skills to further improve the predictive accuracy and widen the research area to the relevant range, and thus better help human beings to fight against flu outbreaks. Generally, there are two promising directions: (1) feature improvement; and (2) medical application.

The following part describes how to involve more helpful and feasible features to improve the model. In our mind, there could be two types of features we need to pay attention to. Those are (a) physical features and (b) human features.

First, we talk about (a) physical features. Flu outbreaks closely correlate with some physical factors, such as temperature and humidity. Predictive models including temperature and humidity are supposed to have a lower MAPE. Nonetheless, we did not leverage those two features due to some practical reasons. First, we can hardly get temperatures of cities of a country due to the research fund. We tried to look for open temperature data from the cyberspace. Unfortunately, although some countries open their weather information, such as Japan, more countries, especially developing countries, have no open access to historical weather information. The possible reason could be lack of the public fund, and so on. Second, even if we get temperature readings of some cities of a country, a calculation of representing the temperature situation of a country is a difficult problem. Third, we have to use weather forecast information to predict future flu data. Nevertheless, weather forecast, especially humidity, has predicting errors, which could be accumulated further in predicting flu data. The disadvantage of error accumulation could invalidate the advantage of including more information. Although we have many potential problems before experimenting, we could continue trying by some tricks to avoid some problems. Regarding the second problem, one promising idea is how to represent the temperature situation of one country. The probable solution is to input temperatures of cities (as many as possible) of a country to constitutional or fully connected layers. The backpropagation algorithm will automatically adjust the focus among all temperatures by refining the weights. Another possibly feasible idea is to convolute or fully connect the weighted temperature (by population). Regarding the third problem, we might delay inputting temperatures to models by a time lag, such as one week. We might have two theoretical explanations. For one thing, the temperature of this week usually highly correlates with the one of last week. Second, is that the impact from temperature to flu virus may also have a time lag. Due to both the two reasons above, inputting a previous temperature to the model could replace the reliance on the accuracy of the weather forecast.

The second idea is related to (b) human features, such as traveling and flu vaccine. The number of travelers positively correlate with flu data since traveling increases the spread of the flu virus. However, the achievement of the number of travelers could be hard. There are many traveling tools, such as cars, bus, train, airlines, and so on. Different tools may have varying impact. Besides, there are many distances of traveling, such as intercity, domestic moving, and international traveling. Different distances may also have a different impact. Moreover, there are different moving objectives, such as school bus, business trip, event (such as the World Cup), and so on. Different objectives may gather different cohort, which is or is not sensitive to the flu virus. For example, the school bus could have a larger impact, since children are weaker to the flu virus. Another example is worldwide events, such as the Olympics. These worldwide events, to some extent, would break the relative isolation in flu infection due to season reverse between the Northern Hemisphere and the Southern Hemisphere. We can leverage a similar solution to include different traveling tools, distances, objectives is to aggregate data by data's type and convoluted and/or fully connect them to leverage the advantage of backpropagation. The effect of the vaccine and the population of flu vaccine could negatively correlate with flu data. The number of vaccine spread and the quantified effect of different flu vaccine will decrease the number of flu patient. This could be a probable reason why in the UK, a Northern Hemisphere country, by including other country's data, the number of flu patients of the coming week was predicted more accurately, since family doctors (the characteristics of UK medical system) helps residents better prevent flu virus in the UK than in other countries. The local influential factors are largely deteriorated and thus the global influential factors assume a pivotal role in spreading flu virus in the UK.

In this part we discuss (2) medical application. The first model improvement is "Attention". A neural net of "Attention" is a series of matrix multiplications and element-wise non-linearities, where elements of the input or feature vectors interact with each other only by addition. Comparatively, attention mechanisms compute a mask which is used to multiply features, by which the space of functions that can be well approximated by a neural net is vastly expanded. There are two types of attention, soft attention, and hard attention. Soft attention multiplies features with a (soft) mask of values between zero and one. Hard attention multiplies features with a (hard) mask of values, which are exactly zero or one, namely a $\in \{0,1\}^k$. In the latter case, we can use the hard attention mask to directly index the feature vector: $\tilde{g}$ =z[a] (in Matlab notation), which changes its dimensionality and now $\tilde{g} \in \mathrm{R}^m$ with m$\leq$k. We might leverage attention algorithms in two levels: (1) temporal attention, and (2) global attention. Regarding temporal attention, let $\chi_1 \in \mathrm{R}^p$ be a_1 time-series input vector (i.e. a time lag with length p), $\mathrm{z}_1 \in \mathrm{R}^j$ be a feature vector, $\mathrm{f}_{lstm,\theta}(\chi_1)$ be an LSTM with parameters θ, $\mathrm{a}_1 \in [0,1]^j$ be an attention vector, $\mathrm{g}_1 \in \mathrm{R}^j$ be an attention glimpse, and $\mathrm{f}_{temporalAttn,\phi}(\chi_1)$ be an attention network with parameters ϕ. Such an idea is implemented as

$a_1 = f_{lstm,\theta}(\chi_1)$, $z_1 = f_{temporalAttn,\phi}(\chi)$, and $g_1 = a_1 \odot z_1$. Regarding global attention, let $\chi_2 \in$ R^q be an input vector including features extracted from q countries, z$_2$ $\in$ R^k be a feature vector, f$_{nn,\eta}(\chi_2)$ be a neural network with parameters η, a $\in$ [0,1]^k be an attention vector, g $\in$ R^k be an attention glimpse, and f$_{globalAttn,\psi}(\chi_2)$ be an attention network with parameters ψ. Such an idea is implemented as$a_2 = f_{nn,\eta}(\chi)$, $z_2 = f_{globalAttn,\psi}(\chi)$, and $g_2 = a_2 \odot z_2$. The second model improvement is reinforcement learning. If we can collect the data on an area range and a population's density, we might construct a reinforcement learning to dynamically arrange hospitalization beds to ILI patients. This research is supposed to be the basis of a reinforcement learning to assign the source of the hospitalization in a city or an area. The reinforcement learning to dynamically arrange hospitalization beds to ILI patients needs three factors: (1) number of flu patients, (2) the number of beds, and (3) the duration of flu hospitalization. Among them, (1) can be known by prediction of this study; (2) can be achieved by the statistics of hospital sources, and (3) can be surveyed bt the previous medical researches. In other words, we have already got enough basis for this reinforcement learning model. Therefore, this reinforcement learning model will be a future research topic and help humans to fight against flu outbreaks more directly. The integration of this reinforcement learning into the hospitalization system will be a quite necessary improvement for human beings.

Moreover, in future studies, we may not only perform prediction of the number of the flu patients totally but also the number of patients of subtypes of flu, such as IVA, IVB, IVC, and IVD. The more diverse the flu prediction supplies, the better we can respond to flu public health emergencies on time. As we know, the vaccines of different subtypes of flu are different. How to precisely manufacturing vaccines of different subtypes is more important than predicting the number of flu patients. However, in the current stage, if we divide flu patients into subtypes, there will not be enough data to perform prediction. In the future, as the predictive technologies and skills develop, there could be a more efficient model type that use a small amount of data to train and to understand what is happening. At that time, we might say we could confidently fight against flu outbreaks

Furthermore, the set of geolocational-temporal time-series prediction we have explored in this Ph.D. research can be applied for other infection, such as hepatitis, tuberculosis, and so on. Almost every infection has a similar spreading path, periodical duration, etc, and almost all of them adapt to the predictive models, such as LSTM or ABMs. In other words, those models were not developed only for one infectious disease. Our research approach can also aim at general infectious disease, and we strongly believe our research approach can help perform an accurate prediction for other infectious diseases if implementation adjusts hyperparameter more or less. We might start with another famous infectious disease, measles, a highly contagious infectious disease caused by the measles virus.

In addition, comparing flu outbreaks among different countries or regions could be a good method to explain the influential factors. For example, if we compare the model of Tokyo and Singapore, we may find whether climate could be a pivotal factor for flu outbreaks or not, and if we compare the model of Korea and Germany, we may find population density could be an indispensable factor for flu outbreaks or not. In the future study, if we need to find impacting factors, comparing flu outbreaks among different countries or regions could be a valid way. However, there are some problems when we compare the flu outbreaks among different countries. There is at least one problem: too many impactors could influence the prediction. Take the comparison between Tokyo and Singapore as an example. Although both of them are big cities, they have so many different factors, such as climate, the human race, different relation to other countries, and so on. Firstly, Singapore is typically a tropical climate, where flu outbreaks last whole years. Comparatively, Tokyo is in a temperate zone in Northern Hemisphere. Flu outbreaks in Tokyo only occur in winters and peaks in February. The different climate means the humidity and the temperature, which could impact flu outbreaks indirectly, are different. Besides, population density is different. Singapore has a fewer population density (731.5 people/km^2) while Tokyo has a more population density (6000 people/km^2). The population density plays a key role in flu spread. Moreover, population construction is also different. Singapore has a median age of 37 years while Tokyo's is 44.7 years. Thereby, there are more elderly people in Tokyo. Singapore annual population growth of 2.1% and Tokyo's is only 0.77%, which means there are more babies in Singapore. Both the elderly and babies are the susceptible population of flu infection. In conclusion, there could be too many different impactors between two regions, and determining which factor is important for flu infection could be still very difficult to pinpoint. How to solve this problem? One solution could be to choose two regions as similar as possible. For instance, choose Singapore and Kuala Lumpur, both of which have similar climates. However, generally speaking, choosing two regions with only one or two difference is almost impossible since human society is too diverse. Another solution could be comparing the current status of one region with its historical status. For example, compare Singapore now with Singapore 10 years ago. We might find how virus mutation and climate changes impact the flu spread. The third solution could be using reinforcement learning or ABM. Both of the methods simulate the real environment. Therefore, we can construct any environment as we need. This could be an inexpensive and efficient way to perform comparative studies. However, the simulated environment cannot include every social factor so that researchers should be very careful when performing comparative studies. In brief, just as stock market analytics, there could be too many social factors that influence flu spread positively and /or negatively. Comparative studies could be an effective and efficient approach although it is still very difficult to conclude some valuable ideas from then comparative studies, from the perspective of our views. And alike, there could also be a regime switch in flu outbreaks just as those in the stock market. A recent regime switch of flu could be in 2009 and 2018 when the type of flu outbreaks seemed quite different but researchers cannot easily find a feature to represent these changes. However, in future studies, as the technologies and skills develop, there could be better methodologies for the human being to understand simulation and prediction of annual worldwide flu outbreaks. The relevant development includes the following items:

(1) a more computing power to simulate a more complicated environment;

(2) a more refined neural network or other algorithms to calculate a complicated environment more precisely; and

(3) more understanding of the mechanism of flu spread.

List of Figures

List of Tables

Bibliography

[1] World Health Organization. Influenza (seasonal): Fact sheet, 2014.

[2] Preeti N Malani. Harrison's principles of internal medicine. *JAMA*, 308(17):1813–1814, 2012.

[3] Ron Eccles. Understanding the symptoms of the common cold and influenza. *The Lancet infectious diseases*, 5(11):718–725, 2005.

[4] MA Urban. Influenza: Viral infections: Merck manual home edition. Technical report, Technical report, 2009.

[5] Eitaro Suzuki, Kiyoshi Ichihara, and Andrew M Johnson. Natural course of fever during influenza virus infection in children. *Clinical pediatrics*, 46(1):76–79, 2007.

[6] Seema Jain, Laurie Kamimoto, Anna M Bramley, Ann M Schmitz, Stephen R Benoit, Janice Louie, David E Sugerman, Jean K Druckenmiller, Kathleen A Ritger, Rashmi Chugh, et al. Hospitalized patients with 2009 h1n1 influenza in the United States, april–june 2009. *New England journal of medicine*, 361(20):1935–1944, 2009.

[7] Centers for Disease Control, Prevention, et al. Key facts about influenza (flu) & flu vaccine. *Atlanta, GA: Centers for Disease Control and Prevention*, 2014.

[8] Centers for Disease Control, Prevention, et al. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the United States. *Updated April*, 19, 2017.

[9] R Vainionpää and T Hyypiä. Biology of parainfluenza viruses. *Clinical microbiology reviews*, 7(2):265–275, 1994.

[10] Alan J Hay, Victoria Gregory, Alan R Douglas, and Yi Pu Lin. The evolution of human influenza viruses. *Philosophical Transactions of the Royal Society of London. Series B*, 356(1416):1861, 2001.

[11] Eri Nobusawa and Katsuhiko Sato. Comparison of the mutation rates of human influenza a and b viruses. *Journal of virology*, 80(7):3675–3678, 2006.

[12] Maria C Zambon. Epidemiology and pathogenesis of influenza. *Journal of Antimicrobial Chemotherapy*, 44(90002):3–9, 1999.

[13] Yoko Matsuzaki, Noriko Katsushima, Yukio Nagai, Makoto Shoji, Tsutomu Itagaki, Michiyo Sakamoto, Setsuko Kitaoka, Katsumi Mizuta, and Hidekazu Nishimura. Clinical features of influenza c virus infection in children. *The Journal of infectious diseases*, 193(9):1229–1235, 2006.

[14] Susumu Katagiri, Akiko Ohizumi, and Morio Homma. An outbreak of type c influenza in a children's home. *Journal of Infectious Diseases*, 148(1):51–56, 1983.

[15] Y Matsuzaki, K Sugawara, K Mizuta, E Tsuchiya, Y Muraki, S Hongo, H Suzuki, and K Nakamura. Antigenic and genetic characterization of influenza c viruses which caused two outbreaks in Yamagata city, Japan, in 1996 and 1998. *Journal of clinical microbiology*, 40(2):422–429, 2002.

[16] Jeffery K Taubenberger and David M Morens. The pathology of influenza virus infections. *Annu. Rev. pathmechdis. Mech. Dis*, 3:499–522, 2008.

[17] Shuo Su, Xinliang Fu, Gairu Li, Fiona Kerlin, and Michael Veit. Novel influenza d virus: Epidemiology, pathology, evolution and biological characteristics. *Virulence*, 8(8):1580–1591, 2017.

[18] Ben M Hause, Emily A Collin, Runxia Liu, Bing Huang, Zizhang Sheng, Wuxun Lu, Dan Wang, Eric A Nelson, and Feng Li. Characterization of a novel influenza virus in cattle and swine: proposal for a new genus in the orthomyxoviridae family. *MBio*, 5(2):e00031–14, 2014.

[19] Emily A Collin, Zizhang Sheng, Yuekun Lang, Wenjun Ma, Ben M Hause, and Feng Li. Cocirculation of two distinct genetic and antigenic lineages of proposed influenza d virus in cattle. *Journal of virology*, 89(2):1036–1042, 2015.

[20] Mariette F Ducatez, Claire Pelletier, and Gilles Meyer. Influenza d virus in cattle, france, 2011–2014. *Emerging infectious diseases*, 21(2):368, 2015.

[21] Hao Song, Jianxun Qi, Zahra Khedri, Sandra Diaz, Hai Yu, Xi Chen, Ajit Varki, Yi Shi, and George F Gao. An open receptor-binding cavity of hemagglutinin-esterase-fusion glycoprotein from newly-identified influenza d virus: basis for its broad cell tropism. *PLoS pathogens*, 12(1):e1005411, 2016.

[22] Zizhang Sheng, Zhiguang Ran, Dan Wang, Adam D Hoppe, Randy Simonson, Suvobrata Chakravarty, Ben M Hause, and Feng Li. Genomic and evolutionary characterization of a novel influenza-c-like virus from swine. *Archives of virology*, 159(2):249–255, 2014.

[23] Megan Quast, Chithra Sreenivasan, Gabriel Sexton, Hunter Nedland, Aaron Singrey, Linda Fawcett, Grant Miller, Dale Lauer, Shauna Voss, Stacy Pollock, et al. Serological evidence for the presence of influenza d virus in small ruminants. *Veterinary microbiology*, 180(3-4):281–285, 2015.

[24] Donald B Smith, Eleanor R Gaunt, Paul Digard, Kate Templeton, and Peter Simmonds. Detection of influenza c virus but not influenza d virus in scottish respiratory samples. *Journal of Clinical Virology*, 74:50–53, 2016.

[25] Fabrice Carrat, Elisabeta Vergu, Neil M Ferguson, Magali Lemaitre, Simon Cauchemez, Steve Leach, and Alain-Jacques Valleron. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *American journal of epidemiology*, 167(7):775–785, 2008.

[26] Susan Murin and K Smith Bilello. Respiratory tract infections: another reason not to smoke. *Cleveland clinic journal of medicine*, 72(10):916–920, 2005.

[27] Jeremy D Kark, Moshe Lebiush, and Lotte Rannon. Cigarette smoking as a risk factor for epidemic a (h1n1) influenza in young men. *New England Journal of Medicine*, 307(17):1042–1046, 1982.

[28] Stephen R Palmer, EJL Soulsby, Paul Torgerson, and David WG Brown. *Oxford Textbook of Zoonoses: Biology, clinical practice, and public health control.* Oxford University Press, 2011.

[29] Gabrielle Brankston, Leah Gitterman, Zahir Hirji, Camille Lemieux, and Michael Gardam. Transmission of influenza a in human beings. *The Lancet infectious diseases*, 7(4):257–265, 2007.

[30] Thomas P Weber and Nikolaos I Stilianakis. Inactivation of influenza a viruses in the environment and modes of transmission: a critical review. *Journal of infection*, 57(5):361–373, 2008.

[31] Caroline Breese Hall. The spread of influenza and other respiratory viruses: complexities and conjectures. *Clinical Infectious Diseases*, 45(3):353–359, 2007.

[32] John W Ward. Twelve diseases that changed our world. *Emerging Infectious Diseases*, 14(5):866, 2008.

[33] Eugene C Cole and Carl E Cook. Characterization of infectious aerosols in health care facilities: an aid to effective engineering controls and preventive strategies. *American journal of infection control*, 26(4):453–464, 1998.

[34] B Bean, BM Moore, B Sterner, LR Peterson, DN Gerding, and HH Balfour Jr. Survival of influenza viruses on environmental surfaces. *Journal of Infectious Diseases*, 146(1):47–51, 1982.

[35] Yves Thomas, Guido Vogel, Werner Wunderli, Patricia Suter, Mark Witschi, Daniel Koch, Caroline Tapparel, and Laurent Kaiser. Survival of influenza virus on banknotes. *Applied and environmental microbiology*, 74(10):3002–3007, 2008.

[36] World Health Organization et al. Ten things you need to know about pandemic influenza. 2005, 2007.

[37] World Health Organization et al. World now at the start of 2009 influenza pandemic. statement to the press by who director-general dr margaret chan. 11 june 2009, 2009.

[38] Gregory A Poland. Vaccines against avian influenza—a race against time, 2006.

[39] Samuel K Peasah, Eduardo Azziz-Baumgartner, Joseph Breese, Martin I Meltzer, and Marc-Alain Widdowson. Influenza cost and cost-effectiveness studies globally–a review. *Vaccine*, 31(46):5339–5348, 2013.

[40] Anthony T Newall and Paul A Scuffham. Influenza-related disease: the cost to the australian healthcare system. *Vaccine*, 26(52):6818–6823, 2008.

[41] Emile Levy. French economic evaluations of influenza and influenza vaccination. *Pharmacoeconomics*, 9(3):62–66, 1996.

[42] Yiting Xue, Ivar Sønbø Kristiansen, and Birgitte Freiesleben de Blasio. Modeling the cost of influenza: the impact of missing costs of unreported complications and sick leave. *BMC Public Health*, 10(1):724, 2010.

[43] X Llach Badia, M Gamisans Roset, JM Tudel Francés, C Sanz Alvarez, and C Terrés Rubio. Study of flu costs. *Atencion primaria*, 38(5):260–267, 2006.

[44] Thomas Szucs, Monika Behrens, and Timm Volmer. Public health costs of influenza in germany 1996-a cost-of-illness analysis. *Medizinische Klinik (Munich, Germany: 1983)*, 96(2):63–70, 2001.

[45] Y Hara, H Ikematu, A Nabeshima, A Hagihara, K Nobutomo, and S Kashiwagi. Cost of medication for influenza infected elderly inpatients. *Kansenshogaku zasshi. The Journal of the Japanese Association for Infectious Diseases*, 73(7):689–693, 1999.

[46] Karen A Fitzner, KF Shortridge, SM McGhee, and AJ Hedley. Cost-effectiveness study on influenza prevention in hong kong. *Health Policy*, 56(3):215–234, 2001.

[47] James Mark Simmerman, Jongkol Lertiendumrong, Scott F Dowell, Timothy Uyeki, Sonja J Olsen, Malinee Chittaganpitch, Supamit Chunsutthiwat, and Viroj Tangcharoensathien. The cost of influenza in thailand. *Vaccine*, 24(20):4417–4426, 2006.

[48] Noelle-Angelique M Molinari, Ismael R Ortega-Sanchez, Mark L Messonnier, William W Thompson, Pascale M Wortley, Eric Weintraub, and Carolyn B Bridges. The annual impact of seasonal influenza in the us: measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, 2007.

[49] Wayan CWS Putri, David J Muscatello, Melissa S Stockwell, and Anthony T Newall. Economic burden of seasonal influenza in the United States. *Vaccine*, 36(27):3960–3966, 2018.

[50] Centers for Disease Control, Prevention, et al. Questions & answers 2009 h1n1 flu (" swine flu") and you. *http://www. cdc. gov/h1n1flu/qa. htm*, 2009.

[51] M Lindsay Grayson, Sharmila Melvani, Julian Druce, Ian G Barr, Susan A Ballard, Paul DR Johnson, Tasoula Mastorakos, and Christopher Birch. Efficacy of soap and water and alcohol-based hand-rub preparations against live h1n1 influenza virus on the hands of human volunteers. *Clinical Infectious Diseases*, 48(3):285–291, 2009.

[52] Tom Jefferson, Chris B Del Mar, Liz Dooley, Eliana Ferroni, Lubna A Al-Ansary, Ghada A Bawazeer, Mieke L Van Driel, Sreekumaran Nair, Mark A Jones, Sarah Thorning, et al. Physical interventions to interrupt or reduce the spread of respiratory viruses. *Cochrane database of systematic reviews*, (7):CD006207–1, 2011.

[53] C Raina MacIntyre, Simon Cauchemez, Dominic E Dwyer, Holly Seale, Pamela Cheung, Gary Browne, Michael Fasher, James Wood, Zhanhai Gao, Robert Booy, et al. Face mask use and control of respiratory virus transmission in households. *Emerging infectious diseases*, 15(2):233, 2009.

[54] Robert A Weinstein, Carolyn Buxton Bridges, Matthew J Kuehnert, and Caroline B Hall. Transmission of influenza: implications for control in health care settings. *Clinical infectious diseases*, 37(8):1094–1101, 2003.

[55] Julia E Aledort, Nicole Lurie, Jeffrey Wasserman, and Samuel A Bozzette. Non-pharmaceutical public health interventions for pandemic influenza: an evaluation of the evidence base. *BMC public health*, 7(1):208, 2007.

[56] David JD Earn, Daihai He, Mark B Loeb, Kevin Fonseca, Bonita E Lee, and Jonathan Dushoff. Effects of school closure on incidence of pandemic influenza in alberta, canada. *Annals of internal medicine*, 156(3):173–181, 2012.

[57] Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boelle, Antoine Flahault, and Neil M Ferguson. Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750, 2008.

[58] AD Heymann, I Hoch, L Valinsky, E Kokia, and DM Steinberg. School closure may be effective in reducing transmission of respiratory viruses in the community. *Epidemiology & Infection*, 137(10):1369–1376, 2009.

[59] Mark Jit, Anthony T Newall, and Philippe Beutels. Key issues for estimating the impact and cost-effectiveness of seasonal influenza vaccination strategies. *Human vaccines & immunotherapeutics*, 9(4):834–840, 2013.

[60] Anthony T Newall, Mark Jit, and Philippe Beutels. Economic evaluations of childhood influenza vaccination. *Pharmacoeconomics*, 30(8):647–660, 2012.

[61] Maarten J Postma, Rob PM Baltussen, Abraham M Palache, and Jan C Wilschut. Further evidence for favorable cost-effectiveness of elderly influenza vaccination. *Expert review of pharmacoeconomics & outcomes research*, 6(2):215–227, 2006.

[62] Scott A Harper, Keiji Fukuda, Timothy M Uyeki, Nancy J Cox, and Carolyn B Bridges. Prevention and control of influenza: recommendations of the advisory committee on immunization practices (acip). *Morbidity and Mortality Weekly Report: Recommendations and Reports*, 54(8):1–41, 2005.

[63] Centers for Disease Control, Prevention, et al. *Key facts about influenza and the influenza vaccine*. 2006.

[64] Anthony T Newall, Juan Pablo Dehollain, Prudence Creighton, Philippe Beutels, and James G Wood. Understanding the cost-effectiveness of influenza vaccination in children: methodological choices and seasonal variability. *Pharmacoeconomics*, 31(8):693–702, 2013.

[65] Anthony T Newall, Heath Kelly, Stuart Harsley, and Paul A Scuffham. Cost effectiveness of influenza vaccination in older adults. *Pharmacoeconomics*, 27(6):439–450, 2009.

[66] World Health Organization et al. Recommended composition of influenza virus vaccines for use in the 2006-2007 influenza season. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 81(09):82–86, 2006.

[67] Michael D Lubeck, Jerome L Schulman, and Peter Palese. Antigenic variants of influenza viruses: marked differences in the frequencies of variants selected with different monoclonal antibodies. *Virology*, 102(2):458–462, 1980.

[68] J Stech, X Xiong, C Scholtissek, and RG Webster. Independence of evolutionary and mutational rates after transmission of avian influenza viruses to swine. *Journal of virology*, 73(3):1878–1884, 1999.

[69] P Suárez, J Valcárcel, and J Ortín. Heterogeneity of the mutation rates of influenza a viruses: isolation of mutator mutants. *Journal of virology*, 66(4):2491–2494, 1992.

[70] EC Holmes, E Ghedin, N Miller, J Taylor, Y Bao, et al. Whole-genome analysis of human influenza a virus reveals multiple persistent lineages and reassortment. 2005.

[71] Catherine Gerdil. The annual production cycle for influenza vaccine. *Vaccine*, 21(16):1776–1779, 2003.

[72] Nathan Proudlove, Ruth Boaden, and Julie Jorgensen. Developing bed managers: the why and the how. *Journal of nursing management*, 15(1):34–42, 2007.

[73] Joan Puig-Barberà, Anita Tormos, Anna Sominina, Elena Burtseva, Odile Launay, Meral A Ciblak, Angels Natividad-Sancho, Amparo Buigues-Vila, Sergio Martínez-Úbeda, and Cedric Mahé. First-year results of the global influenza hospital surveillance network: 2012–2013 northern hemisphere influenza season. *BMC Public Health*, 14(1):564, 2014.

[74] Andreas S Weigend. *Time series prediction: forecasting the future and understanding the past*. Routledge, 2018.

[75] Li Zhang, Wei-Da Zhou, Pei-Chann Chang, Ji-Wen Yang, and Fan-Zhang Li. Iterated time series prediction with multiple support vector regression models. *Neurocomputing*, 99:411–422, 2013.

[76] Shahrokh Akhlaghi and Ning Zhou. Adaptive multi-step prediction based ekf to power system dynamic state estimation. In *Power and Energy Conference at Illinois (PECI), 2017 IEEE*, pages 1–8. IEEE, 2017.

[77] Feng Wang, Haiyan Wang, Kuai Xu, Ross Raymond, Jaime Chon, Shaun Fuller, and Anton Debruyn. Regional level influenza study with geo-tagged Twitter data. *Journal of medical systems*, 40(8):189, 2016.

[78] 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.

[79] Mamunur R Malik, Zaeem Ul Haq, Quaid Saeed, Ruth Riley, and Wasiq M Khan. Distressed setting and profound challenges: Pandemic influenza preparedness plans in the eastern mediterranean region. *Journal of infection and public health*, 2017.

[80] Hongyan Wu, Yunpeng Cai, Yongsheng Wu, Ren Zhong, Qi Li, Jing Zheng, Denan Lin, and Ye Li. Time series analysis of weekly influenza-like illness rate using a one-year period of factors in Random Forest regression. *Bioscience trends*, 11(3):292–296, 2017.

[81] M Goeijenbier, P van Genderen, BJ Ward, A Wilder-Smith, R Steffen, and ADME Osterhaus. Travellers and influenza: risks and prevention. *Journal of travel medicine*, 24(1), 2017.

[82] Daihai He, Roger Lui, Lin Wang, Chi Kong Tse, Lin Yang, and Lewi Stone. Global spatio-temporal patterns of influenza in the post-pandemic era. *Scientific reports*, 5:11013, 2015.

[83] Barbara Michiels, Van Kinh Nguyen, Samuel Coenen, Philippe Ryckebosch, Nathalie Bossuyt, and Niel Hens. Influenza epidemic surveillance and prediction based on electronic health record data from an out-of-hours general practitioner cooperative: model development and validation on 2003–2015 data. *BMC infectious diseases*, 17(1):84, 2017.

[84] Yan Bu, Jinhong Bai, Zhuo Chen, Mingjing Guo, and Fan Yang. The study on China's flu prediction model based on web search data. *Journal of Data Analysis and Information Processing*, 6(03):79, 2018.

[85] Pi Guo, Li Wang, and Yuantao Hao. Building a prediction system of influenza epidemics with lasso regression model and baidu search query data. *Chinese Journal of Health Statistics*, 34(2):186–191, 2017.

[86] Pi Guo, Jianjun Zhang, Li Wang, Shaoyi Yang, Ganfeng Luo, Changyu Deng, Ye Wen, and Qingying Zhang. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. *Scientific reports*, 7:46469, 2017.

[87] Feng Liang, Peng Guan, Wei Wu, and Desheng Huang. Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in liaoning, from 2011 to 2015. *PeerJ*, 6:e5134, 2018.

[88] Chunli Wang, Yongdong Li, Wei Feng, Kui Liu, Shu Zhang, Fengjiao Hu, Suli Jiao, Xuying Lao, Hongxia Ni, and Guozhang Xu. Epidemiological features and forecast model analysis for the morbidity of influenza in Ningbo, China, 2006–2014. *International journal of environmental research and public health*, 14(6):559, 2017.

[89] Shweta Chaudhary. Using big data for computational epidemiology in india. *International Journal of Advanced Research in Computer Science*, 8(2), 2017.

[90] Nikita E Seleznev and Vasiliy N Leonenko. Boosting performance of influenza outbreak prediction framework. In *International Conference on Digital Transformation and Global Society*, pages 374–384. Springer, 2017.

[91] Chia-liang FU, Ray-jade CHEN, and Yu-sheng LO. Earlier prediction of influenza epidemic by hospital-based data in taiwan. *DEStech Transactions on Social Science, Education and Human Science*, (emse), 2017.

[92] Chen-yuan Tung, Tzu-Chuan Chou, and Jih-wen Lin. Using prediction markets of market scoring rule to forecast infectious diseases: a case study in taiwan. *BMC public health*, 15(1):766, 2015.

[93] Tsung-Hau Chen, Yung-Chiao Chen, Jiann-Liang Chen, and Fu-Chi Chang. Flu trend prediction based on massive data analysis. In *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 304–308. IEEE, 2018.

[94] Josephine LK Murray, Diogo FP Marques, Ross L Cameron, Alison Potts, Jennifer Bishop, Beatrix von Wissmann, Naoma William, Arlene J Reynolds, Chris Robertson, and Jim McMenamin. Moving epidemic method (mem) applied to virology data as a novel real time

tool to predict peak in seasonal influenza healthcare utilisation. the scottish experience of the 2017/18 season to date. *Eurosurveillance*, 23(11), 2018.

[95] Armin Spreco, Olle Eriksson, Örjan Dahlström, and Toomas Timpka. Influenza detection and prediction algorithms: comparative accuracy trial in östergötland county, sweden, 2008–2012. *Epidemiology & Infection*, 145(10):2166–2175, 2017.

[96] Armin Spreco, Olle Eriksson, Örjan Dahlström, Benjamin John Cowling, and Toomas Timpka. Integrated detection and prediction of influenza activity for real-time surveillance: Algorithm design. *Journal of medical Internet research*, 19(6), 2017.

[97] Balsam Alkouz and Zaher Al Aghbari. Analysis and prediction of influenza in the uae based on arabic tweets. In *Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on*, pages 61–66. IEEE, 2018.

[98] Ali Alessa and Miad Faezipour. A review of influenza detection and prediction through social networking sites. *Theoretical Biology and Medical Modelling*, 15(1):2, 2018.

[99] Kathy Lee, Ankit Agrawal, and Alok Choudhary. Forecasting influenza levels using real-time social media streams. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, pages 409–414. IEEE, 2017.

[100] Batuhan Bardak and Mehmet Tan. Disease outbreak prediction by data integration and multi-task learning. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on*, pages 1–7. IEEE, 2017.

[101] Shikha Verma, Younghee Park, and Mihui Kim. Predicting flu-rate using big data analytics based on social data and weather conditions. *Advanced Science Letters*, 23(12):12775–12779, 2017.

[102] Hongxin Xue, Yanping Bai, Hongping Hu, and Haijian Liang. Influenza activity surveillance based on multiple regression model and artificial neural network. *IEEE Access*, 6:563–575, 2018.

[103] Xiangjun Du, Aaron A King, Robert J Woods, and Mercedes Pascual. Evolution-informed forecasting of seasonal influenza a (h3n2). *Science translational medicine*, 9(413):eaan5325, 2017.

[104] Xiangjun Du and Mercedes Pascual. Incidence prediction for the 2017-2018 influenza season in the United States with an evolution-informed model. *PLoS currents*, 10, 2018.

[105] Jaber Belkhiria, Robert J Hijmans, Walter Boyce, Beate M Crossley, and Beatriz Martínez-López. Identification of high risk areas for avian influenza outbreaks in california using disease distribution models. *PloS one*, 13(1):e0190824, 2018.

[106] Fred Sun Lu, Suqin Hou, Kristin Baltrusaitis, Manan Shah, Jure Leskovec, Rok Sosic, Jared Hawkins, John Brownstein, Giuseppe Conidi, Julia Gunn, et al. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the boston metropolis. *JMIR public health and surveillance*, 4(1), 2018.

[107] Susannah Paul, Osaro Mgbere, Raouf Arafat, Biru Yang, and Eunice Santos. Modeling and forecasting influenza-like illness (ili) in houston, texas using three surveillance data capture mechanisms. *Online journal of public health informatics*, 9(2), 2017.

[108] Haruka Morita, Sarah Kramer, Alexandra Heaney, Harold Gil, and Jeffrey Shaman. Influenza forecast optimization when using different surveillance data types and geographic scale. *Influenza and other respiratory viruses*, 2018.

[109] HT Thrastarson, J Teixeira, EA Serman, A Parekh, and E Yeo. Analysis and modeling of influenza outbreaks as driven by weather. In *AGU Fall Meeting Abstracts*, 2017.

[110] Mihajlo Grbovic and Haibin Cheng. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 311–320. ACM, 2018.

[111] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11. ACM Press, 2003.

[112] Jeremy Ginsberg, Matthew Mohebbi, Rajan Patel, Lynnette Brammer, Mark Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009. doi:10.1038/nature07634.

[113] Robert L Axtell, Clinton J Andrews, and Mitchell J Small. Agent-based models of industrial ecosystems. *Rutgers University, October*, 6, 2003.

[114] Melissa Tracy, Magdalena Cerdá, and Katherine M Keyes. Agent-based modeling in public health: current applications and future directions. *Annual review of public health*, 39:77–94, 2018.

[115] Joshua M Epstein and Robert Axtell. *Growing artificial societies: social science from the bottom up.* Brookings Institution Press, 1996.

[116] Gerardo Chowell, Lisa Sattenspiel, Shweta Bansal, and Cécile Viboud. Mathematical models to characterize early epidemic growth: A review. *Physics of Life Reviews*, 18:66–97, 2016.

[117] Jon Parker and Joshua M Epstein. A distributed platform for global-scale agent-based models of disease transmission. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 22(1):2, 2011.

[118] 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.

[119] Philip Cooley, Bruce Y Lee, Shawn Brown, James Cajka, Bernadette Chasteen, Laxminarayana Ganapathi, James H Stark, William D Wheaton, Diane K Wagener, and Donald S Burke. Protecting health care workers: a pandemic simulation based on allegheny county. *Influenza and other respiratory viruses*, 4(2):61–72, 2010.

[120] Bruce Y Lee, Shawn T Brown, George W Korch, Philip C Cooley, Richard K Zimmerman, William D Wheaton, Shanta M Zimmer, John J Grefenstette, Rachel R Bailey, Tina-Marie Assi, et al. A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 h1n1 influenza pandemic. *Vaccine*, 28(31):4875–4879, 2010.

[121] Models of infectious disease agent study.

[122] Nathaniel Hupert, Wei Xiong, and Alvin Mushlin. The virtue of virtuality: The promise of agent-based epidemic modeling. *Translational Research*, 151(6):273–274, 2008.

[123] Marek Laskowski, Bryan CP Demianyk, Marcia R Friesen, Robert D McLeod, and Shamir N Mukhi. Improving agent based models and validation through data fusion. *Online journal of public health informatics*, 3(2), 2011.

[124] José Manuel Galán, Luis R Izquierdo, Segismundo S Izquierdo, José Ignacio Santos, Ricardo Del Olmo, Adolfo López-Paredes, and Bruce Edmonds. Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation*, 12(1):1, 2009.

[125] Niko Speybroeck, Carine Van Malderen, Sam Harper, Birgit Müller, and Brecht Devleesschauwer. Simulation models for socioeconomic inequalities in health: a systematic review. *International journal of environmental research and public health*, 10(11):5750–5780, 2013.

[126] The SIR Model for Spread of Disease - The Differential Equation Model, 2004.

[127] The mathematical modeling of epidemics, 2005.

[128] Svr with different epsilons, 2017.

[129] An example of decision tree, 2017.

[130] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *In , O. Bousquet, U.v. Luxburg, and G. Rsch (Editors*, pages 169–207. Springer, 2004.

[131] David J. Hand, Padhraic Smyth, and Heikki Mannila. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001.

[132] The algorithm of random forest, 2017.

[133] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

[134] Tin Kam Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, pages 278–, Washington, DC, USA, 1995. IEEE Computer Society.

[135] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[136] Artificial neural network with layer coloring, 2017.

[137] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[138] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[139] Understanding LSTM networks.

[140] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. 1999.

[141] wikipedia of LSTM.

[142] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.

[143] Daniel Hsu. Time series forecasting based on augmented long short-term memory. *arXiv preprint arXiv:1707.00666*, 2017.

[144] Franoise Beaufays. The neural networks behind google voice transcription. *Google Research blog*, 2015.

[145] H Sak, A Senior, K Rao, F Beaufays, and J Schalkwyk. Google voice search: faster and more accurate. *Google Research blog*, 2015.

[146] Pranav Khaitan. Chat smarter with allo, 2016.

[147] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[148] An infusion of ai makes google translate more powerful than ever | wired.

[149] Apple's machines can learn too.

[150] iphone, ai and big data: Here's how apple plans to protect your privacy.

[151] ios 10: Siri now works in third-party apps, comes with extra ai features.

[152] Bringing the magic of amazon ai and alexa to apps on aws.

[153] Microsoft's speech recognition system is now as good as a human.

[154] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[155] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[156] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[157] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[158] Wavenet: A generative model for raw audio, 2017.

[159] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016.

[160] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. *Frontiers of Computer Science*, 10(1):96–112, 2016.

[161] Mikołaj Bińkowski, Gautier Marti, and Philippe Donnat. Autoregressive convolutional neural networks for asynchronous time series. *arXiv preprint arXiv:1703.04122*, 2017.

[162] Prajit Ramachandran, Tom Le Paine, Pooya Khorrami, Mohammad Babaeizadeh, Shiyu Chang, Yang Zhang, Mark A Hasegawa-Johnson, Roy H Campbell, and Thomas S Huang. Fast generation for convolutional autoregressive models. *arXiv preprint arXiv:1704.06001*, 2017.

[163] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*, 2017.

[164] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[165] Why attention is more accurate?

[166] Andrew Ng. What data scientists should know about deep learning. *URL https://www. slideshare. net/ExtractConf*, 44, 2015.

[167] Jie Zhang and Kazumitsu Nawata. A comparative study on predicting influenza outbreaks. *Bioscience trends*, 11(5):533–541, 2017.

[168] Haibin Cheng, Pang-Ning Tan, Jing Gao, and Jerry Scripps. Multistep-ahead time series prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 765–774. Springer, 2006.

[169] Arun Venkatraman, Martial Hebert, and J Andrew Bagnell. Improving multi-step prediction of learned time series models. In *AAAI*, pages 3024–3030, 2015.

[170] Yukun Bao, Tao Xiong, and Zhongyi Hu. Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing*, 129:482–493, 2014.

[171] Flunet, 2018.

[172] K-R Müller, Alexander J Smola, Gunnar Rätsch, Bernhard Schölkopf, Jens Kohlmorgen, and Vladimir Vapnik. Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks*, pages 999–1004. Springer, 1997.

[173] Kumpati S Narendra and Kannan Parthasarathy. Identification and control of dynamical systems using neural networks. *IEEE Transactions on neural networks*, 1(1):4–27, 1990.

[174] National institute of infectious diseases, Japan, 2018.

[175] 4 2014/15 2018/19 .

[176] 2014/15 2018/19 .

[177] http://apps.who.int/flumart/default?reportno=1&countrycode=jp.

[178] http://apps.who.int/flumart/default?reportno=12.

[179] UK tourism to hit record number of visitors in 2018 - despite brexit warnings, 2018.

[180] World tourism rankings, 2018.

[181] Tourism in the Australia, 2018.

[182] Tourism in the Australia, 2018.

[183] Tourism in the United States, 2018.

[184] China inbound tourism in 2015, 2018.

[185] Japan Meteorological Agency, 2018.

[186] The Centers for Disease Control and Prevention. The flu season., 2014.

[187] G Kolata. Study shows why the flu likes winter. Technical report, New York Times, 2007.

[188] Anice C Lowen, Samira Mubareka, John Steel, and Peter Palese. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151, 2007.

[189] R Roos. Study: Flu likes weather cold and dry or humid and rainy. Technical report, University of Minnesota Center for Infectious Disease Research and Policy, 2013.

[190] E Elert. Fyi: Why is there a winter flu season? . Technical report, Popular Science., 2013.