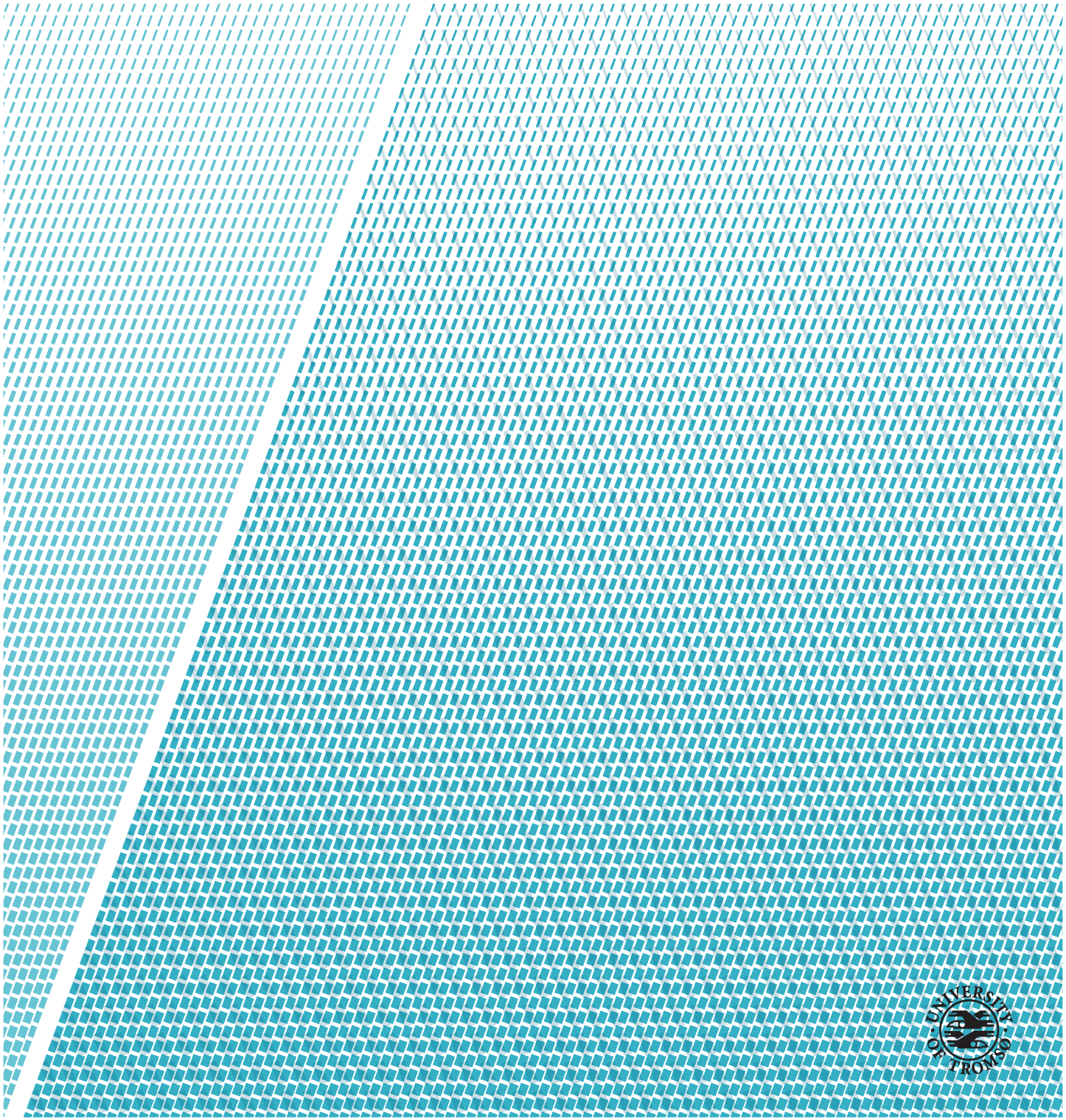
UiT
THE ARCTIC UNIVERSITY OF NORWAY

Faculty of Science and Technology
Department of Physics and Technology

On the Feasibility of Using Twitter Data to Assess the Global Circulation Patterns of Influenza Viruses

Inga Setså Holmstrand

EOM-3901 Master's thesis in Energy, Climate and Environment, 30 SP - June 2018



Abstract

Having the flu is something that everyone is familiar with, and the influenza season hits every year. The intensity and timing vary from year to year driven by climatic conditions and antigenic evolution, through mechanisms that are only partially understood. Most research agree that the virus originates in East and South-East (E-SE) Asia and spread throughout the world through human movement. In this thesis we explore the possibility of modelling this circulation pattern using a simple semi-stochastic mathematical model. Interestingly, this model exhibits chaotic behavior and is unable to confirm the above mentioned hypothesis. A separate approach is to analyze influenza incidence data. However, these data are subject to substantial underreporting (or complete lack of reporting) during the low-seasons. Some recent works have suggested using social media data to obtain proxies of influenza-like illness (ILI) data. In this thesis we discuss if it is possible to discern pattern or tendencies using data from Twitter. As the data used is collected only during a short time window, we can only say something about the feasibility of using this approach to analyze the global circulation of influenza viruses.

Acknowledgements

Writing my master's thesis has been an interesting and educational experience. As it now has come to an end, I would like to thank my supervisor professor Martin Rypdal, for his help and advice on this thesis.

Another person I would like to thank, is Sandra S. S. Nesse for proofreading, critique and the comments she gave. A thank you to the others who have contributed on my thesis as well.

After five years of studying hard, a thank you to my fellow classmates and friends is needed. For laughs, inspiration, and support throughout my years at the university.

Finally, I need to thank my family for their support and encouragements over the past years, as a student and while making this thesis.

Inga Setså Holmstrand
Tromsø, June 2018

Contents

List of Figures

List of Tables

Abbreviations

ILI Influenza Like-Illness

HA Haemagglutinin

NA Neuraminidase

E-SE East-Southeast Asia.

CDC Center for Disease Control and Prevention

API Application Programming Interface

REST Representational State Transfer

AI Artificial Intelligence

GFT Google Flu Trends

CDC ILISN CDC's US Outpatient Influenza-Like Illness Surveillance Network

CDC's ISPSN CDC's Influenza Sentinel Provider Surveillance Network

1 Introduction

1.1 Influenza

Influenza is a virus that everybody is familiar with, as it emerges with new types every year, which in turn gives yearly outbreaks of the disease. Influenza is, what we call, a contagious respiratory disease, and the virus itself is subdivided into three different influenza types for humans, A, B and C. Influenza type B is further broken into two groups, and influenza type A, is further classified into different classes, depending on the combination of the two proteins, haemagglutinin (HA) and neuraminidase (NA). Influenza type C is detected less often than the other types, and only causes mild infections in humans. Type C does not have any subclasses. There is an influenza type D as well, but this is not a type that is known to affect humans [CDC, 2014, WHO, 2014].

The well-known symptoms of the influenza virus, or the flu, as it is called, is fever, body aches, headaches, coughing and tiredness. And for most people, it is harmless. There are people that are at a higher risk of getting infected and seriously ill. That is elders, newborns, pregnant women, and people with diseases such as asthma, these are said to be in risk groups. For the people in these risk groups, the virus could be lethal. And therefore, it is an important

virus to monitor and prevent, so that the morbidity and mortality numbers goes down for these groups [CDC, 2014].

It is known that both influenza type A and B cause epidemics, but only influenza type A is known to cause pandemics, like the swine influenza in 2009 [WHO, 2014], which was of the type H1N1. Another subtype of A, H3N2, is the major cause of human influenza morbidity and mortality, world-wide, and on average, 5 to 15 % of the World's population are infected with this type of influenza at any time [Russell et al., 2008a].

Despite progress in many areas of influenza research, it is largely unknown when or to what extent the virus will change, and to what extent it will spread throughout the world. It is known that in temperate countries, the influenza seasonality is typically during the coldest part of the year, but it does not have to be. This makes the influenza season to some extent, predictable.

In tropical countries it is much more difficult to say anything about the seasonality, but it often coincides with the rainy season, but we can see influenza activity throughout the whole year in this region. World-wide, the yearly epidemic result in approximately 3 to 5 million cases of serious illness, and consequently being able to predict influenza epidemics would be of great benefit for healthcare, society and economic welfare [Azziz Baumgartner et al., 2012, Viboud et al., 2006a].

The influenza virus seasonality does vary with latitude, but why it does, is not exactly known. There has not been found any environmental links that have been convincing to describe this. But, as stated above it often coincide with the rainy season or in the coldest months. Estimating the burden of influenza is difficult to measure, and in tropical countries it is even more difficult to do. Since there are many unknown facts and many questions on how the influenza virus behave in the tropical region, more research need to be done. To get a good model of the influenza burden in tropical countries, good surveillance data is needed. And since there are big variation from year to year in the impact of influenza, the model depends on several years of data. Since good surveillance data just started in this region, the research studies are of short burden duration, and they will get better as time goes by [Viboud et al., 2006a].

The influenza season of 2017-2018 has been a really bad influenza season compared to previous seasons. Because of a low efficiency in the influenza vaccine. There has been an increase of deaths and hospitalization in this season, and scientist are therefore wondering why and how this could happen. They are also wondering how they could make the vaccine preparation more efficient in the following years. Trying to figure out this question, they analyzed the influenza virus, the circulation of that year influenza strain, and analyzed the predictions they made, when they made that seasons vaccine. One way to make the influenza vaccine better is to make mutation to the influenza strain in the vaccine. Which will lead to an increase in the immune response, making the vaccine better [Teitzel, 2018].

1.2 Prevention, Complication and Transmission of the Influenza Virus

Since influenza is so well-known, the virus has different prevention strategies. How the virus is transmitted and what complications they may arose, are also therefore well-known. Even because of this, there are still some uncertainties attached to this problem.

1.2.1 Prevention

Since there is a high morbidity and mortality each year because of the influenza virus, prevention of this virus is very important, so that these number potentially can go down.

Each person is susceptible of the new virus, but some are at a higher risk, which is mention earlier. Increasing age, pregnancy, chronicle illnesses, and residential care all increases the risk of being infected, and with a higher risk of complication and deaths. Today there are two ways to lessen and prevent the impact of the influenza virus, a vaccine that contains inactivated virus-organism and prophylaxis with antivirals, a drug that works on viruses. Different countries have policies considering prevention, but they all recommend people in the

risk group to take the influenza vaccine against influenza annually [Cooper et al., 2003].

The Influenza Vaccine

From the recent paper "Factors associated with influenza vaccination among healthcare workers in acute care hospitals in Canada", the author suggested that the influenza vaccine uptake would increase exponentially with every year the vaccine was taken. Other studies have shown the same or similar thing as well. This will thus suggest the individual perceptions, that are associated with vaccine recognition and rejection will be stable over several years.

In the paper, they determined that physicians with a higher knowledge about influenza and vaccination were less likely to expect a severe reaction to the vaccine, and more likely to consider influenza vaccine effective than what a person with a lower knowledge would [Hussain et al., 2018].

Vaccines that not contains the correct influenzas strain, because the strain has changed after the vaccine decisions was made, will of course not be as effective as a year where the vaccine matches the strain that is in the influenza season. One example is for the 2014-2015 season in the United States. This season more than 80 % of the influenza viruses that circulated where different from what the vaccine would protect from. The influenza vaccine effectiveness was only 13 % that year. Which led to an increased number of mortalities and morbidities. Even in years when the vaccine is matched to the circulating viruses, the effectiveness is not 100 % of it, but normally somewhere between 40 to 60 %. Which is actually lower than for most of the non-influenza virus vaccines [Paules et al., 2018].

It is also recommended that health care workers get vaccinated to stop spreading the virus at their work place [Hayes, 2008].

Tracking the virus, where it is and what kind of it that is circulating, helps prevent influenza. Since it helps with the vaccination. It is possible to figure out when the vaccination should be given, and what kind of vaccine that it should

be. Different types of the influenza virus, needs different types of vaccines. The best time for the vaccination to be given, is just before the season starts. After the vaccination is given to a person, it takes approximately 2 weeks before a person is immune to the annual virus. How well the vaccines help depends on the age of the person getting the vaccine [Hayes, 2008].

The vaccine that is developed each year targets the virus strains that is predicted to be the most prevalent by the Centers of Disease Control and Prevention (CDC). And is therefore not effective to every type of influenza strain [Hayes, 2008]. Vaccination on both health care personnel and patients, is the best way to prevent an influenza spread. Since vaccinated health care personnel has been associated with a decrease of influenza illnesses among the patients and mortality in long-term care facilities [Weinstein et al., 2003].

Other Prevention Strategies:

In the influenza season, preventing the virus itself, strict hand washing is a very effective strategy to prevent the spread of the virus [Hayes, 2008].

Another way to prevent influenza is by using antiviral medication, but this is not possible to buy in every country, as in some countries, only hospitals have access to it [Paul et al., 2014]. These medications can be a helpful medication with the vaccination. These are useful at health care facilities, since they can effectively reduce the spread of influenza, when used in combination with other control measures [Weinstein et al., 2003].

It is also possible to have isolation precautions to prevent influenza spread. This prevention procedure is thus very important in health care facilities. This precaution could be, placing patients alone in a room, or with other infected patients [Weinstein et al., 2003].

1.2.2 Complications

Complications that may occur for someone that are infected with the influenza virus, may be inflamed mucous membranes, that is sinuses, ears and bronchi,

and also pneumonia. More than 200,000 people are hospitalized each year, and about 36,000 in the United States of America, because of complications. And therefore, prevention work is important for all [Hayes, 2008].

Is has been shown that pregnant women who get infected by the virus has a three-to four-fold higher risk than the non-pregnant women. Pregnant women are therefore hospitalized more than non-pregnant women. It has been shown that fetal and newborn conditions that are related to maternal influenza, are congenital malformations, altered brain development, miscarriage and stillbirth. Some recent studies have found a correlation between utero exposure to influenza and increased risk of Parkinson's disease and schizophrenia [Hayes, 2008].

We have seen influenza pandemics before, the latest in 2009, which was known as the Swine Flu. All of the latest pandemics are studied and especially the latest three. The four pandemics that where in 1918, 1957, 1968 and in 2009, where all influenza type A [Kilbourne, 2006]. To get a pandemic, at a minimum, the virus needs itself to have a major change in the HA antigen. One could see that in 1957, there were changes in both HA and NA antigens. Which again caused a higher rate of illnesses and deaths. The Spanish flue in 1918, may have been special because of wartime conditions and also a less important bacterial infections [Kilbourne, 2006]. When there is sufficient change in the virus to get a pandemic, the change is called an antigenic shift to the virus, whereas small changes is called antigenic drift.

In the brief period of the modern virology, the 15 different HA antigens that are known to exist. Only the three different antigens, H1, H2 and H3, are known to cause a pandemic [Kilbourne, 2006].

One of the worst complications of influenza is pneumonia. And for elders this is much worse than for adults and young adults. Treatment is more difficult for elders, and hospitalization and death is frequent among the elder patients [Meehan et al., 1997]. But, pneumonia is not the only cause of hospitalization and deaths, but with influenza, there is an increase in other pulmonary and cardiovascular diseases. There is also some hospitalization because of neuromuscular complications [Rothberg et al., 2008].

It is possible to prepare for an epidemic, but even though people do a big amount of hand washing, public education and masks to prevent spreading in health care services, there will be epidemics each year, and even pandemics of influenza in the future. Even though, it is very important to prevent it from spreading, since this will be a part of reducing deaths and hospitalization of infected people [Kilbourne, 2006]. Because of these preventions actions that have been put into place, we have seen a decrease in deaths over the years as more knowledge on the virus has been known [Doshi, 2008].

1.2.3 Transmission

The influenza virus is transmitted by aerosols, large droplets, or direct contact with secretions. Therefore, it is possible to be infected, if you are susceptible for the particular version of the virus, at any public place where an infected person has visited [Hayes, 2008].

The drier the air, the longer the viral particles live, which leads to that the virus is more prevalent in the winter months, or at least, this is what we think. Since the air is colder and drier, but also that the nasal passageways will be direr as well. And in colder months the heated buildings will contain a drier air than what it will in the other seasons, which makes it easier to spread [Hayes, 2008].

If a person gets infected by the influenza virus, the person will be contagious for 1 to 2 days before, and up to 5 days after symptoms begins. It appears so that children have a longer incubate time than adults. We know that the viral particles can live on non-porous surfaces for up to 24 hours, and on paper surfaces for up to 15 minutes, so that this need to be thought of [Hayes, 2008].

Aerosols are small particles that are suspensions in air. They are small enough to remain airborne for some time, because of their low settling. Aerosols transmission is the mode of transmission that may have the greatest impact for infection control, since this requires specialized personal protective equipment. Since these particles moves very slowly in still air, they are easily carried over

a long distance by air columns and air currents. Which can in turn cause long-distance infections [Tellier, 2009]. Coughing and sneezing will generate a substantial quantity of particles to infect others [Tellier, 2006].

Early studies of influenza transmission in humans, showed that infection is activated more efficiently when the virus is collected in the lower respiratory tract rather than the upper respiratory tract [Weinstein et al., 2003]. The respiratory tract is a part of the human anatomy. It is divided in two, the upper and the lower tract. The upper part includes, among other things, the nose and nasal passages. The lungs could be a part of the lower part of the respiratory tract, if it is not looked upon as a separate part. Trachea is a part of the lower tract [Weinstein et al., 2003].

From one research paper that where published in 2007, they looked at how the aerosol spread of the influenza virus where dependent on relative humidity and temperature on guinea pigs. In this paper they discovered that the virus transmission of influenza is in fact dependent on temperature and the relative humidity. They did 20 experiments where they had a range between 20 to 80 % in humidity and in three different temperatures, 5 °C, 20 °C and 30 °C. And the result was that it indicated that both the cold and the low humidity where favored for the virus to transmit. They suggest that these two environmental factors could be a part of the seasonal pattern of influenza. Not that it is not possible to get infected during the summer, but that it is much easier when temperatures are cold, and the humidity is low [Lowen et al., 2007].

In this research there where a lack of transmission at 30 °C, which question if their research represents human infections, as we have that the virus also transmit in tropical areas [Lowen et al., 2007].

Influenza does not always spread from human to human, and often the virus emerges in animals, like birds. Avian influenza is influenza where all birds are susceptible, and therefore we often see outbreaks in birds, especially turkeys and chickens. Humans are rarely infected by this type of the virus. Humans are believed to be infected through pigs, that act as a host. Where the virus need to go through mutation to the virus in airborne transmission. When a mammal first gets infected by the virus, the virus is transmitted from mammal

to mammal by the airborne route [Webster, 1997].

The influenza virus is proposed to transmit with aerosols, but the importance of this transmission tool is unclear. One study even suggests that it is enough to breath to spread the influenza virus [York, 2018].

1.3 Thesis Structure

Chapter 2: In this chapter we are looking at Twitter and how this can be used to tell us something about Influenza.

Chapter 3: In this chapter we are looking at the global circulation of influenza, and how the virus is changing.

Chapter 4: In this chapter The SIR model is represented.

Chapter 5: In this chapter we are looking at community structures, and how this is used in this thesis.

Chapter 6: In this chapter the results are presented.

Chapter 7: In this chapter the discussion is made.

Chapter 8: In this chapter the conclusion is made, with a summery and further work within this problem.

Finally, the appendix and the bibliography come, containing Mathematica codes that has been used in this thesis, and the time series of the Twitter influenza data.

/2

Collecting Data About Influenza Like Illnesses Using Twitter

Another way to see if we can see the pattern in real life and other interesting things, is to use influenza data. One way of collecting these data from the whole world, is to download data from the health care services. Another way, that has been showed to work well, is to use internet profiles and social media to collect these data.

In the past 20 years, there have been powerful advances in computer science, and with this, algorithms and advanced hardware on the known problems of understanding spoken and written text. Today this science is wildly used by everyone. Machine translation, speech synthesis is examples of things that is used every day. Social media may be one of the most used computer science today, and here I am looking at the most leading social networking and micro-blogging service, Twitter [Agogo and Hess, 2018].

Another way of downloading data that has been used, to get influenza like illnesses-data (ILI), is Google Flu Trends (GFT). Which is when Google is capturing the queries from people that search about influenza. There are even some that have used a service to analyze blogs, where people have written about themselves being sick [Corley et al., 2010].

2.1 What is Twitter?

Twitter is what we call a micro-blog, developed in 2006, where the users may post short messages, called *tweets*, which was original a maximum of 140 characters, but has since November 2017 doubled their character limitation[1]. Twitter is the most famous micro-blog service all over the world.

Each Twitter-profile have what they have called followers, which will get these messages in their own feed. These followers could be anybody, but most often they are friends and people you know. For well-known persons, they typically have many followers, and can therefore share their thoughts and opinions to many people. These messages, these *tweets*, can be *retweeted*, which is when another user take your *tweet* and post it on their profile. Each of this retweeted messages, will have RT in the beginning of the text, and with the original's profile name. This means that you will always be able to see the original *tweet*.

Because of this, one *tweet* could possibly spread to many different users. Twitter has multiple times shown to be a good source of information on what is going on in a country, and as well in the whole world, as many of the users post their opinion about the community and about what they see, on their public Twitter profile [Java et al., 2007, Kwak et al., 2010].

1. https://twitter.com/

2.2 Why Using Twitter as a Source of Influenza Data?

Since influenza is under-reported, as not everyone goes to see a doctor when they get sick, other sources needs to be found. And one may wish to look for other sources to find data. Twitter has shown to be a good source with a great correlation for Influenza-like illnesses (ILI). One reason for using this method of collecting influenza data, is that it is a quite fast way to collect data. Since it is, as stated above, that the users of Twitter often publish their thoughts on their public profile as it happens. It is possible to get the data simultaneously, whereas collecting data from the health reports could take several weeks [Signorini et al., 2011a]. Using twitter as a source opens up for easier access for collecting influenza data.

Twitter has more than 190 million users worldwide and produces over 55 million *tweets* each day from all over the world. Most of the *tweets* is mostly conversation between a few users, spam or general shatter. But even though there is a lot of noise in the site, it is possible to find useful information from this. Twitter profiles has previously and, it will most likely in the future, been or be used to measure political opinion, impact on earthquake effects, and national sentiment from the public [Signorini et al., 2011a].

As it says above, Twitter will give us a real time information of people with an ILI, while data from people who has confirmed the influenza virus, will be delayed by 1-2 week after the diagnosis has been made. Since the data system of influenza diagnosed patients is mainly manual. For the best intervention and prevention for an epidemic, the public health authorizes need to be informed as soon as possible as it is a growth of the influenza virus proportion in the public. So faster ways to get the influenza data for the healthcare services, the more efficient would the preventive intervention be for every year [Achrekar et al., 2011].

The reason for choosing Twitter over for example, Facebook or other different micro-blogs, is that the Twitter-profiles are often open for the public and has many users. You do not need your own profile to see others, like most of the

Facebook profiles are. The threshold for publishing something on Twitter is much lower than it is for publishing a text on Facebook amongst the people, particularly for young people. Since Twitter is made for publishing random thoughts and opinions [Dawar et al., 2018]. Most of the *tweets* are also posted with geographical coordinates, because of the heavy use of smart phones.

With the geographical coordinate on almost every *tweet*, it is possible to say something about the spread of the influenza virus [Lampos and Cristianini, 2010]. Here in this thesis, only the text-messages that have a geographic location is used, since the spreading pattern of influenza is what we are looking for.

Although there are most young people that have a Twitter profile, we still see a diversity in demographic groups. Twitter may not only be used to collecting data but can also be used to enlist people to studies [Sinnenberg et al., 2017].

Something that has been shown is that if the media talks about the influenza virus, there will be more *tweets* that mentions influenza, than if the media did not talk about it. So, in these periods when the are some talk about the influenza virus, in the media, there are more *tweets* that mentions influenza, but actually not *tweets* where people are sick. This has been seen in other web-based flu surveillance systems as well [Broniatowski et al., 2013]. But, since the media often increase their stories about influenza, in the influenza season, there will most likely be an increase also in the number of infected people.

In a paper from 2014, they showed that Twitter surveillance would highly improve influenza forecasting. The paper also states that it is possible to forecast the influenza prevalence rates some weeks into the future using only Twitter. They state that Twitter is more accessible, and that it will provide better forecasting of epidemics [Paul et al., 2014].

2.3 Hashtags and Queries

In these so called *tweets*, often people use hashtags in front of a word, which is the hash character #. These "hashtagged" words will be marked with a different color and on Twitter, these words are blue. Clicking on these words that are "hashtagged", will lead to many *tweets* with the same hashtag. And thus, it would be much easier to find exactly those *tweets* that contains the information or content that you are looking for.

To collect data we could be looking for text, or *tweets*, that contains words that could be symptoms, like "headache", "sore throat" and so on, but also search for "flu", and "influenza". It is also possible to look for words like "#flu", where we have used the hashtag [Lampos and Cristianini, 2010]. And as stated, the data that will be collected will not be forecasting the influenza season, but rather give us the real time reports of influenza. Since we are looking for people that are sick with influenza right now. But an increasing number of people that write that they are sick could indicate that an epidemic is in the starting.

Searching only for "headache" and "sore throat" could be symptoms for other things than the influenza virus. So, using more symptoms could be smarter. But that again would lead to less *tweets*. Only searching for "flu" or "influenza" lead to a lot of *tweets* in the biggest cities, but in small cities we get a nice and small number.

This means that in the biggest cities, many people use Twitter daily, but also that we have more noise in the downloaded data. The noise could be information from the health care services about vaccines or influenza precautions, or it could be statistics about this or previous influenza seasons. That not actually people that have the influenza virus or people that has influenza-like illnesses.

2.4 Twitter's REST API

It is possible to download data form Twitter, because of its free application programming interface (API). Which can be used to interact with users and feeds of the social media platform. But, to download data from Twitter, you

need to have your own user on the website. Twitter API allows users to analysis data (*tweets*), and trending topics in time. In this thesis, downloading the data from Twitter, the REST API way to downloading information has been used. Using this way to get the interesting *tweets*, it is only possible to download 180 queries per 15 minutes [Dawar et al., 2018].

REST stands for representational State Transfer, and is the word Roy Fielding, a computer scientist, gave his own description of his Web's architectural style [Masse, 2011]. These API's uses the pull strategy for collecting the wanted data. There is also another way to download data from twitter, called Streaming API, but this is not used in this thesis. Using the REST API will give us data from the last 10 days, unless we take a maximum of *tweets* we want to download.

Downloading less data, takes a shorter time, and since the *tweets* were downloaded within ten days, a maximum of downloads per Twitter search is used. This way of downloading data, we search for words that we are interested in, as queries. It is possible to search for several words in one *tweet* using comma between the words, or the queries, while downloading the data [Kumar et al., 2014].

The Twitter API is allowed access to 1 % its data, and in real time. This is one of the strengths of the Twitter database, since it allows free access to a large set of data immediately after the data was created and published.

Twitter as a source for collecting data within health services, is a new way of collecting data, and is a rapid growing field. Which can be seen by the number of publications. The most commonly researched topics within health and sickness on Twitter, are cases with high morbidity and mortality. Such as influenza, cancer and Ebola. But there is also research about other health behaviors such as smoking [Sinnenberg et al., 2017].

2.5 Geolocation of the *Tweets*

While we are downloading the data, we are only searching for *tweets* that have a location. Since we only want the *tweets* where we know what the location is,

since we are trying to analyze the spreading pattern. Because of this, we will not be able to get all of the *tweets* that contains influenza data. Not all of the user has a geographical location on their *tweet* or they may want the location of where they are to be private.

Getting a *tweet* with a geographical location is available from two different sources, which is geotagging information and from the profile descriptions from the users [Kumar et al., 2014].

Geotagging information is when the users have chosen to provide their location of their *tweets*, and with the smart phone's GPS, the location will be highly accurate. The profile of the users can have the location of the user in their biography. The biography is on every profile page, where someone can say something about themselves, which one could be where they live.

2.6 Problems in Using Different Cities From All Over the World

The proportion of *tweets* from different cities in the world is of course different, which we would suspect since the population in those cities are quite different. The more people in a population, the more Twitter users there might be. The culture of a population, might also have a significant saying in how many that have a Twitter profile. We also know that there are a bigger proportion of people that have a Twitter profile in the United States than in any other country [Statista, 2018]. Which will be influencing the Twitter-data.

From the difference in the number of Twitter-users, we can see that even in big countries with a big population, the proportion does not have to be the same. Some countries have a much bigger proportion of Twitter-users. What we also see is a difference in the age groups between countries.

One problem that also arose in downloading the influenza data is that not every city has English as first language, and that the code did not translate the queries to different languages, only a few do in what that has been downloaded,

or even non or a few English speakers. Because if this, we also need to search for *tweets* in the city's native language. To do this, Google Translate[2] was used to find the words in different languages. Some cities that has several first language which is used, the most used language of those was used, as it often where not many *tweets* on those other languages.

Since English is such a highly used language, and since the "culture" on Twitter is to write it in in English, every Twitter-search was also done in English in the non-English countries. One reason for users to write it in English, is that in this way it is possible to communicate to the rest of the world.

In this thesis data from over 30 countries was search for from all over the world. It is clearly a difference in number of influenza incidents which can be seen on the plots in the appendix Fig. A.1.

2.7 Scientific Papers

2.7.1 Using Yahoo

As stated, not only Twitter and GFT can be used to surveillance influenza. But, also the search engine Yahoo[3]. Which one study did and collected data influenza data from March 2004 to May 2008 [Polgreen et al., 2008]. They have used the idea that people search for influenza information online, when they do need it, and the fact that the large number of health-related information makes it more difficult to find precisely what you are looking for. As there are 8 million people that search for health-related issues every day makes it possible to find patterns in search history.

In this study they used 2 different types to measure the influenza occurrence. The first type to get data, were based on weekly influenza cultures. Which comes from clinical laboratories that report the total number of respiratory specimens tested, and the number of positive influenza tests in the influenza

2. https://translate.google.no/
3. https://www.yahoo.com/

season. The second type of data they used were of weekly mortality attributable to pneumonia and influenza. From this data, the study obtained figures of the influenza mortality in the USA. As the influenza query search data needed to match these numbers, they collected data from March 2004 to May 2008 [Polgreen et al., 2008].

They collected the search queries that were from the States only, as it were only in this region they had collected data, and the fact that the season of influenza vary geographically. They calculated the daily influenza search, by dividing the daily number of influenza search by the total number of all searches that had been done. As the influenza data they had collected were on a weekly basis, they calculated the weekly average of influenza search [Polgreen et al., 2008].

To see what relationship there was between culture-positive cases of influenza and influenza-related searches, they examine the relationship between these two at a national level. They discovered that the fractions between these two, have a similar pattern over time, but there is s sharp increase in the search for influenza that precedes the cultures that are tested positive for influenza. To be able to test the search queries data, they fitted it into a linear model, so that they could test the predictability of the search frequency on positive influenza culture results, which also include a time variable, and it is as follows:

$$c_t = \beta_0 + \beta_1 s_{t-x} + \beta_2 t + \epsilon_0 \tag{2.1}$$

In this equation, Eq. 2.1, t is a time trends that is measures in weeks, c_t is the rate of positive influenza cultures received during week t, and s_{t-x} is the search frequency in the week of t-x. To determine the appropriate lag, they examined 11 different possible values for x and compared it with R^2 value for each of these models. And the best fit for this model, was given for 1-week lag. The coefficient β_2 is not significant different form zero in any of the tried models. As for this model, the best fitting model predicted an increase in the number of cultures positive for influenza three weeks in advanced [Polgreen et al., 2008].

As for the search and the influenza mortality results, they also made a fitted linear model, so that they could test the predictability of search frequency with regard to the mortality rate, and it is as follows:

$$m_t = \beta_0 + \beta_1 s_{t-x} + \beta_2 t + \epsilon_0 \tag{2.2}$$

Where in this case m_t is the total number of deaths. All the other variables are defined as in Eq. 2.1. For the best fitted model in this case, the search data peaked 4-6 weeks prior an increase in mortality attributable to influenza and pneumonia [Polgreen et al., 2008].

They discovered that there is a distinct temporal association that exist between influenza-related search-terms frequency and disease activity. In the States, the search activity seems to increase some weeks prior to the positive influenza cultures and in influenza related deaths [Polgreen et al., 2008].

2.7.2 The Use of Twitter to Track Diseases

Another study that also use Twitter to see if it is able to detect disease activity is, "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S during the Influneza A H1N1 Pandemic" [Signorini et al., 2011b].

In this study they also looked at the public concert of the influenza pandemics, as stated in the title. They started to collect a number of *tweets*, starting in April 29, 2009, with the pre-specified search terms, *flu, swine, influenza, vaccine, tamiflu, oseltamivir, pneumonia, h1n1, symptom, syndrome* and *illness*. Each of their collected *tweets* where geolocated using the profiles home location. In October 1, 2009 they began downloading expanded sample of *tweets*, using Twitter's API [Signorini et al., 2011b].

As in the previous study, they were only interested in *tweets* from the United States, and *tweets* that were not in English. And because of the volume of post on Twitter varies over time, and varying across geographical regions, they used statistics that were expressed in terms of the fraction of the total *tweets* emitted

within the corresponding time interval and geographical region [Signorini et al., 2011b].

To determine the contribution of each of the influenza-related Twitter term, they used Support Vector Regression, which is a more general class of Support Vector Machine, a supervised learning method generally applied to solve classification problems. This model will produce a nonlinear model that will minimize a preselected linear-error-cost function where features serve as regression variables [Signorini et al., 2011b].

Their Results

They manage to get a data set that contained 951,697 *tweets* in their first data set. And these were collected from April 29 and June 1, 2009. Their second data set contained approximately 4.2 million *tweets*, that were selected from about 8 million influenza-related *tweets* that were observed between October 1, 2009 and until the end of the year. When they had collected these *tweets*, they made estimates on ILI based on this data set. To verify their method, they used a standard leaving-one-out cross-validation methodology. And they got an average error of 0.28 %, and a standard deviation of 0.23 % [Signorini et al., 2011b].

Their results showed that Twitter data not only can be used to track the users interest and their concern related to influenza, especially the H1N1-influenza pandemics in 2009, but it is also possible to estimate disease activity in this moment. They do mention, since there is no comparable data they are available, it is not possible to validate their results. But, the results and trends that are observed are reasonable and quite consistent with what we would expect. One example of this, where when there were a drop in the number of *tweets* that contained antiviral drugs, at the same time as official disease reports indicated that most of the cases were mild [Signorini et al., 2011b].

The *tweets* which reflects the user's own level of disease and discomfort, they researchers devised an estimation method that were based on well-understood machine learning methods. Which showed that the accuracy of the resulting ILI

estimates identified and used in their model, which contains closely information that were associated with disease activity. Their result was also able to create a distinct relationship between Twitter data and the epidemic curve of the H1N1 pandemic in 2009. Both at the national level and at a geographical level [Signorini et al., 2011b].

In this study, they did not try to forecast an influenza epidemic, as many others, but rather to be able to make real-time estimates using Twitter. Which will be much faster than traditional estimates, which will be 1-2 weeks delayed [Signorini et al., 2011b].

2.7.3 An Analysis of the 2012-2013 Influenza Epidemics using Twitter

In the paper "National and Local Influenza Surveillance through Twitter: An Analysis of the "2012-2013 Influenza Epidemic", the authors demonstrate that influenza surveillance using social media with a system build and deployed before the influenza season have started. They found out that the number of *tweets* declined as the media attention declined [Broniatowski et al., 2013].

In this research the authors were able to create a new classification model that overcomes the barrier of *tweets* that contains the word influenza but is not actually about an infected person. By separating *tweets* indicating influenza infection, and those who indicate concern or influenza awareness. Which makes the model able to estimate influenza prevalence from normalized *tweet* volume [Broniatowski et al., 2013].

Their downloading of Twitter data, started at September 30, 2012. Which was the start of the 2012-2012 epidemic defined by the Centers of Disease Control and Prevention (CDC). Which ended in May 31, 2013. Their collection contained 1.3 billion *tweets*.

To filter data, they authors used a binary classification models to identify relevant data for influenza surveillance at each stage. And these models indicated if the *tweet* were relevant to health, to influenza, or indicative of an actual infec-

tion. The first filter, indicated if the *tweet* were relevant of irrelevant of health, which the classifier was estimated to have 90 % precision. Each of the *tweets*, were labelled with three different labels, (1) if the *tweet* discussed influenza or not, (2) if the *tweet* indicated infection or the user's awareness of influenza and (3) whether the *tweet* referenced the user themselves, or someone else. The third classifier was not used in the final classifiers. The labelled data was then used to trained parameters of separate logistic regression models for the two classification tasks. Using this, they manage to get 570,000 *tweets* that indicated an infected user. After this has been identified, they normalized the weekly number of these infected *tweets* by the total number of *tweets* in that week so that they were able to produce a Twitter-based influenza prevalence measure. To evaluate this, they compared their result with the CDC's US Outpatient Influenza-Like Illness Surveillance Network (CDC ILISN) [Broniatowski et al., 2013].

To manage to get the geographical location of each *tweet*, they used their recently geolocation system, called Carmen. With the GPS information which were associated with the small percentage of the collected *tweet*, Carmen will collect information from the user's biographies profiles [Broniatowski et al., 2013].

Their Results

On the national level of the United States, their system managed to identify 104,200 influenza infections. These *tweets* correlated strongly with the CDC ILISN data, from October 2013 to May 2013 ($r=0.93$, $p<0.001$). On the contrary, the weekly number of *tweets* containing influenza keywords provided by the US Department of Health and Human Services is much less strongly correlated ($r=0.75$, $p<0.001$). And the difference between these are significant at a $p<0.001$ level. The absolute error of their estimates is 0.0102 after normalizing the weekly rates to sum 1. The mean absolute error of their infection estimates is 0.0046, a 45 % reduction error over the keyword filter [Broniatowski et al., 2013].

On the municipal level, they looked at New York, where they also used the

same technique. In this case they had 4,800 *tweets* which were identified form New York City. The New York City Department of Health and Mental Hygiene, did a blind evaluation of their algorithm, and it showed a strong correlation between the city's weekly emergent department visits for ILI, and the city's number of *tweets* in the same week ($r=0.88$, $p<0.001$) [Broniatowski et al., 2013].

They did weekly correlation with the Twitter-data and the national ILI-data, which the Pearson correlation coefficient varied between 0.91 to 0.97. The mean is 0.93, with a standard deviation of 0.02. Their system also matched the direction of the change in cases by 85 % accuracy, which for baseline keyword-based systems, is 46 % [Broniatowski et al., 2013].

Any correlation analysis of time series could be potential bias if the underlying data is not stationary. One example, if each week influenza infection count is a function of the previous week's count, then it would be expected these two weeks would be correlated. This additional time series analysis, shows that it is possible to capture the detail beyond the overall trend [Broniatowski et al., 2013].

Their algorithm of collecting data establish significant improvements and is less sensitive to noise on Twitter. As when there were talk about the H7N1 virus in China, which had massive media attention. They observed a large increase of *tweets* with influenza keywords, which is expected, but *tweets* with infection only, had just a slight increase, or not at all. Their Twitter data correlated strongly with the governments data over influenza throughout all of the weeks of the influenza season [Broniatowski et al., 2013].

2.7.4 Detecting Influenza Epidemics Using Search Engine

Their Model

In this research paper, "Detecting Influenza Epidemics Using Search Engine Query Data", they are looking at query data from search engine as the title states. They mention that to get a faster detection of influenza than the original ways

of doing it, which often has a lag from 1 to 2 weeks, many different surveillance system has been created so that it would be possible to monitor influenza with no lag. As that 90 million American are believed to search online to get information of dieseases or medical problems each year, which makes the web search queries uniquely source of information about all kinds of health problems [Ginsberg et al., 2009].

It has been showed that a set of Yahoo search queries that contains influenza keywords, have correlated with virology and mortality surveillance data over multiple years. In this research, they are looking at Google as a search engine. The authors of this paper have looked at hundreds of billions of search-logs from 5 years of Google searches. Their system generated a more comprehensive model which can be used in influenza surveillance, which has both national and regional estimates of ILI in the United States [Ginsberg et al., 2009].

They collected historical logs from 2003 to 2008, which they computed a time series of weekly counts for the 50 million of the most common search queries in the United States. Each of these time series were normalized by dividing the count for each query in a particular week by the total number of online searches that same week and in the same location. They wanted to make a simple model that would estimate the probability that a random physician would visit a particular region is related to an ILI, which is equivalent to the percentage if ILI-related physician visits. Only a single explanatory variable was used, the probability that a random search query submitted from the same region is related to an ILI. They fitted a linear model by using the log-odds of an ILI-physician visit the log-odds of an ILI search query. Their linear model, is as follows:

$$\text{logit}(I(\alpha)) = \alpha \text{logit}(\text{Q(t)}) + \epsilon \tag{2.3}$$

In this equation, Eq. 2.3, I(t) is the percentage of ILI physician visits, Q(t) is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ϵ is the error term [Ginsberg et al., 2009].

To help build this model, the research paper's authors used influenza data from the CDC's influenza Sentinel Provider Surveillance Network (CDC's ISPSN), which is free of cost. For each of the nine regions in the United States that

CDC have surveillance for, the CDC reported the average percentage for all outpatients visits to sentinel provides that were ILI-related on a weekly basis. No data were provided outside of the influenza season, and those ILI-data that were collected outside of this season are left unvalidated [Ginsberg et al., 2009].

They designed an automated method for selecting ILI-related search queries, which required no previous knowledge about influenza. They have also measured how efficient their model would fit the CDC ILI-data in each of the nine regions if they only would use one query, as the variable $Q(t)$. Each of the 50 million candidates queries which was in their database were separately tested in this manner, so that the queries that could most accurately could model the CDC ILI visit percentage in each of the nine regions [Ginsberg et al., 2009].

Their Results

In the 2007-2008 influenza season they used preliminary versions of their model to generate ILI estimates, and shared their result each week with the Epidemiology and Prevention Branch of Influenza Division at the CDC to evaluate the timeliness and accuracy. And across the nine regions in the United States, their model was able to estimate consistently the current ILI percentage 1-2 weeks prior of the publications of reports by the CDC's Influenza Sentinel Provider Surveillance Network [Ginsberg et al., 2009].

Since local surveillance is especially useful for health planning in the area, they wanted to validate their model even further against weekly ILI percentage for individuals state, instead of those nine regions. The CDC does not make state-level ILI-data public, but the authors were able to validate their data with the state of Utah ILI-reports, which they obtained a 0.90 correlation across 42 validation points. From the validation of the model, they concluded with that Google queries can be used to estimate the ILI percentage, and accurately, in the nine regions prior to the CDC's ILI surveillance reports manage [Ginsberg et al., 2009].

As we can see from all of these four papers, using social media and search engine, it is possible to use them to get ILI-data, which in fact make a good correlation of official health department's ILI-reports. Which, in use, could make the health care more prepared for an epidemic or a pandemic as it could see an increase in infected before reports can see it.

/3

Global Circulation and Antigenic Drift

How the global circulation of the influenza virus works, is something that has been wanted for many years. And if it is perfectly understood, it would help understand the influenza season, and help predicting when it will hit much better. Precisely the global circulation is what they try to figure out in the paper, "Global Circulation Patterns of Seasonal Influenza Viruses Vary with Antigenic Drift" from 2015 [Bedford et al., 2015]. Despite the better understanding in the complete genome sequence data of influenza, there are many aspects of how the virus evolves, and the epidemiological of it that are not known, that is, measurments of viral diversity across time, across space and among the influenza subtypes [Rambaut et al., 2008]. In this study, that study the antigenic drift, which means that they are analyzing the virus itself, and how it changes.

Most of the study of the influenza virus, has only focused on a single segment, without trying to see at how the subtypes of the virus interact with each other. Most of the studies have not determined how the viruses relates to antigenic

evolution. And even though the two influenza viruses of type A, H1N1 and H3N2 have a seasonality, the forces who decide the periodicity, and how they vary are unknown [Rambaut et al., 2008].

In "The Global Circulation of Seasonal Influenza A (H3N2) Virus" [Russell et al., 2008a] they are just looking at the evolution of the H3N2 virus, as it is states in the title. They looked at how the virus evolves and changes, and where this happens. Their result tells us that there are evidence of seeding from a region, against local persistence in temperate regions. Where the seeded region would be East-Southeast Asia (E-SE Asia). What they also discovered is that it seems that the virus is travelling from this region to Europa, Oceania, and North America, and after this it travels to South America. Which could be explained by these regions travel and trade connections. They did not either find evidence of influenza seeding back to this region.

There are some evidence that even though the most important contributions are from China and South-East Asia, it has been found out that small temperate regions outside of Asia could contribute to the global circulation of influenza. It has been found evidence of migration virus from temperate to tropical countries, and that their linage may exist outside of Asia for several seasons. They manage to persist because of dynamical migration between regions and different seasonality [Bedford et al., 2010]. Some studies have shown that China, South-East Asia and the United States contribute to the trunk of the influenza genealogy, and hence mutation of the virus have affected the global influenza population, where the virus who were contributed from the United States, often is the one found in South America. Which could be consistent with aviation [Bedford et al., 2010].

3.1 Studies of Different Influenza Viruses

The authors mention that studies have shown that, each year, the H3N2 epidemics, a type A influenza, results from the introduction of new genetic variants in E-SE Asia, where it is believed that the virus circulates all the time, because of a network of temporally epidemics, rather than local percitence [Bedford

et al., 2015, Russell et al., 2008a]. In addition to this particular influenza virus, H1N1 viruses, and two antigenically diverged lineages of type B, called B/Victoria/2/1987-lik (Vic) and B/Yamagata/16/1988-like (Yam), are viruses that circulate among humans, and they have considerable disease burden. The global circulation of these influenza viruses is overlooked, even though it is an important part of understanding influneza [Bedford et al., 2015].

Considering that both influenza type A and B gives comparable symptoms and that they evolve in a similar matter, the authors of [Bedford et al., 2015], suggest that these viruses will follow the same pattern for global circulation. Where the new variant of the influenza types originates in E-SE Asia, which will replace the already exciting variant. To test this in this paper, the researchers compared the global circulation of HA genes of H3N2, the former H1N1, Vic, and Yam viruses. They managed to cover viruses from 2000-2012, and they reduced the impact of surveillance biases by subsampled these data to more equitable spatiotemporal distributions.

What they were able to see, was that faster rates of nucleotide mutation and amino acid in H3N2 and in H1N1, than in the type B viruses, which was previous shown as well. But they also discovered genealogical diversity in the B virus than what it is in the A virus. It is possible to discover a consistent pattern for the H3N2 virus. In addition to China and Southeast Asia, India frequently contributed to new viruses. Which means that India is a part of the contributing countries in the E-SE Asia pattern. It has also been briefly periods where other regions outside of this leading pattern have contributed with new viruses, once in 2007-2008 Northern Hemisphere winter. But this is very rare, and those viruses descend directly from E-SE Asia [Bedford et al., 2015].

Studies have shown that the global circulation of H1N1 surprisingly do not follow the same global circulation pattern as H3N2 [Bedford et al., 2015]. What has been discovered is that the H1N1 virus' lineages do unite with the viruses from E-SE Asia and India, but at a much slower than for the H3N2 viruses.

Analyses of the influenza type B viruses, Vic and Yam have revealed further differences from the H3N2 virus. Where one can see the lineages circulating outside of E-SE Asia for many years, without any evidence of seeding from this

region. A good example of this, is the seeding of the North American 2006/2007 Vic season, it was directly from the 2005-2006 North American viruses, which also with the seeding of the North American 2001-2002 Yam viruses, directly being seeded by Northern American viruses. Which also the same pattern can be seen in E-SE Asia. That the viruses circulate exclusively in the same region for more than 1 year [Bedford et al., 2015].

What have been showed is that the persistence of the different types of influenza viruses, H3N2 for approximately 6 months [Bedford et al., 2015, Russell et al., 2008a], H1N1 for about 9 months, Vic about 13 months, and Yam for approximately 12 months [Bedford et al., 2015]. H3N2 has the shortest persistence time across the world, but it is longer in China and India. Patterns that have been seen inside of China, has shown a characterization by North and South contributing the same to persistence, as combining the North and South phylogeny nodes resulted in substantially greater persistence estimates then from North and South alone. For the type B viruses, in India and in China they have a persistence time which were over two years [Bedford et al., 2015].

To see differences in the global migration pattern of these four different types of influenza, two types of A and two types of B. A study estimated the amounts of virus movements between different regions [Bedford et al., 2015]. The rates between pairs of regions were highly correlated, which suggest a similar global connectivity for all the viruses. Nonetheless, even though the overall arrangement of the pattern were similar, it is possible to see that the H3N2 migrate between regions more often than the other type A virus H1N1, and the two B type viruses. [Bedford et al., 2015] hypothesize that this is because of a relationship between the global movement and the rates of antigenic drift. What they also hypothesize that there are lower rates of immune escape for B viruses and for H1N1, compared to the H3N2 virus.

In [Russell et al., 2008a], they do mention that Japan, Thailand and Malaysia are expectation of this E-SE Asian migration pattern.

3.2 The Genomic and Epidemiological Dynamics of Influenza

A study that used a data set of the two influenza type A viruses, H1N1 and H3N2, from New York from a 12-year period at the genomic and epidemiological scale from viral isolates fron New York state and New Zealand [Rambaut et al., 2008].

The viral isolated from New York state's and New Zealand's changing pattern in genetic diversity definitely show the seasonal dynamic of influenza. The peak of the epidemic in the two regions are clearly in their respectively winters. In New Zealand are offset of New York state with appropriately 6 months. A similar pattern is discovered when Australia is a part of the analysis. The genetic diversity of the H3N2 virus in New Zealand was in general lower than what it was in New York state. This could be because of the lower susceptible population in New Zealand than in New York state. The difference in this population could also explain why the virus type A, H3N2 in New Zealand are sometimes less diverse than the type A, H1N1 in New York state, even though the H3N2 is more epidemiologically dominant than the H1N1 virus. The genetic diversity that is seen, is modest compared the other evolving viruses that evolves rapid which also infect fewer people. Which suggest that there is strong natural selection, in addition to periodic bottlenecks, will reduce the level of diversity that is co-circulating at any time [Rambaut et al., 2008].

In both of the population of New York state and in New Zealand, the H1N1 virus's season highly described peaks in diversity are coinciding with the weakly described peaks in the H3N2 virus diversity, that is, the measure of the peaks of these two viruses are negative correlated. Where the Wilcoxon signed-rank test gave: $W = 348$, $n = 32$, $p < 0.002$. From this one can say that there is an interaction with these two viruses, that is, the H1N1 will be suppressed by herd immunity when the H3N2 virus is dominant. We have that the H1N1 virus will only dominate and cause an epidemic when there has been a mild H3N2 epidemics the previous year [Rambaut et al., 2008].

The persistence of the viral diversity in epidemic peaks of these to type A viruses

in these regions, H1N1 and H3N2, have two explanations, (1) the chains of infection will survive within each of the population and across inter-epidemics interval, (2) or that the genetic diversity is imported into the temperate regions each year. Taking what is stated above, and from other studies [Russell et al., 2008a, Bedford et al., 2010, Russell et al., 2008b], the second explanation is strongly weighted.

It is believed by many that the influenza virus circulates continually in the tropics, and if this is true, it would explain the E-SE Asia leading pattern better, that is that the virus is being able to persist in Asia [Russell et al., 2008a].

In "The Genomic and Epidemiological Dynamics of Human Influenza a Cirus" [Rambaut et al., 2008], to look for evolutionary interactions, they used multivariate statistics to summarize the difference in the history of the H3N2 virus from New York state. In this paper they are concluding with that the dynamic of how the type A viruses evolves, is a complex interplay with rapid mutation, frequently reassortment, widespread gene flow, natural selection, functional interactions among segments, and global epidemiological dynamics [Rambaut et al., 2008].

It is possible to see consistent patterns in the two population, New York state and New Zealand, two temperate regions in the Northern and Southern Hemisphere, with the persistence of viral lineages across multiple epidemics [Rambaut et al., 2008].

To fully understand the antigenic evolution of influenza, it will be essential to consider the complex spatial epidemiological dynamics, with the genome-wide interaction of the virus [Rambaut et al., 2008].

3.3 Flight Traffic

One of the most used transportation option that we have today to travel between different countries and cities are air planes. It is therefore a great way to spread diseases between humans and continents. An infected person travels with a plane to a new country or city, where there are many susceptible people of this

distinct influenza virus. For a spreading of the virus from one city to another, it could be interesting to look at the flight traffic between those same two cities. To see if there could be an option that flying could be a big part of the explanation of the influenza spread between regions.

It is obvious that air traffic is a part of the mechanism for spreading influenza, but to what extent is unclear. A better understanding of this issue is something that is wanted. Should air traffic be stopped if there is a pandemic is one question that arises with the better understanding of the relationship between air traffic and on influenza spreading. But recent modelling of a pandemic has shown that air traffic has little to say compared to other prevention mechanisms. But this model that say this, has not been challenged with an observation study [Viboud et al., 2006b].

Mathematical modelling is important within influenza research, as it provides information about changes and impacts to the human population and how the global spread of infectious agents. There are only a few studies that directly explore the importance of air travel, even though this is an important area of research. One of the biggest reason for this, is the difficulty of getting air travel data. In the model of air travel spreading of influenza of Rvachev and Longini [Rvachev and Longini Jr, 1985], they manage to reconstruct the migration of the 1968-1969 influenza pandemic where they used 52 different cities and compared this with the air traffic. In their model they assumed that this pandemic only spread through normal air transportation [Flahault et al., 2006].

In a study from 2006 [Flahault et al., 2006], they also looked at the impact of air travel with influenza, where they had 4 different scenarios to explore this. One with none control measures, one with immediate isolation and air travel restrictions. Another where the treatment of all symptomatic infectious individuals. In scenario 4, they included vaccination. They also did modelling where they varied the reduction of air traffic, varying the proportion of isolated infections individuals, varying in duration of antivirals, varying in vaccination coverage, and they looked at the impact of seasonality and transmissibility.

Their results from this study showed that an influenza pandemic cannot be

contained in that specific area, it would be much more difficult to control. They model showed that air traffic had little impact until it was almost stopped [Flahault et al., 2006].

With the increase of flights that has been over the past year, it is believed that in the next pandemics, flight will be a one of the causes of the transmission of influenza [Leitmeyer and Adlhoch, 2016]. As air planes are closed settings, transmission of influenza, so that person to person contact and contact with contaminated surfaces is major causes of transmission.

There is evidence that transmission of influenza does exist on air planes if there is an infected person on-board. But the data that has been published do not permit any conclusive estimates of the likelihood and extent. And many studies were often biased because of other potential exposures before or after the flight [Leitmeyer and Adlhoch, 2016].

Something that does not come as a surprise, is that, a longer flight time will lead to more people getting infected by the virus. The infection rate of influenza on planes do also depend on which class one is travelling on. And within flight transmission, economic class could be more significant, and especially on long flights. The transmission rate is as well as dependent on how many that is on the plane. Not only that there are less people on the flight, and less people that could get infected. But there will be less people getting infected due to people not sitting to close to each other [Wagner et al., 2009].

It has been found that aircrew have a high rate of ILI, where there was one study that showed a 33 % attack rate over a 7 month period in unvaccinated. In 1979, there was an airplane the was delayed because of engine failure during take-off, which resulted in a 3-hour delay. In this time, all of the passengers stayed in the plain during the delayed. While this was happening, the ventilation was turned off. The infected individual symptoms while on board, and within 72 hours, 72 % of the passengers, and 40% of the crew experienced symptoms in ILI. This made that ventilation must be on if there is a delay on over 30 minutes [Leder and Newman, 2005].

3.3.1 In This Thesis

In this thesis we are looking at direct flights at a given data between every city that is used when downloading Twitter data. To being able to analysis the flight data, we made a matrix with the cities at both the rows and columns. We made an assumption that where there was a direct flight between two cities, there would be a direct flight the other direction as well. Thus, making the 33x33-matrix much easier to do, as it in this case would be symmetric. On the diagonal, there will be zero, because that in this specific place in the matrix represents direct flight from the same city, which is not possible.

To look for direct flights between two cities, Expedia[1] was used. Which showed us how many direct flights there was. In all, there were made 512 searches between all cities. We did not consider different flights, and how many passengers there where room for. To get the data, a random date was used, which in this case were June 1st. This because one need to get an idea of how many flights that travels between cities. We did not look at travels where one needed to change flights to get to the destination.

If a city had more than one airport, all of the possible airports were of course chosen.

To compare the flight data with the influenza data, a fitted linear model was made between them, which make it easier to see how they fitted together. The influenza data was smoothed by using moving average and then correlated. So that the correlation number was the one we used in the fitted model with the number of direct flights.

1. https://www.expedia.no/

4

The SIR Model

The SIR model is a model that simulate influenza data, or other viruses. There are many different ways to use this model, and in this case, it is as follows:

$$\dot{S} = \frac{\beta \mathrm{SI}}{n} + \mu \tag{4.1}$$

$$\dot{I} = \frac{\beta \mathrm{SI}}{n} + \gamma \mathrm{I} \tag{4.2}$$

$$\dot{R} = \gamma \mathrm{I} \tag{4.3}$$

Here S stands for susceptible, that is the people in the population that are not immune to the virus, and can therefore possibly be infected with it. The I represents the size of the population that are infected with the influenza virus, and R is the part of the population that have recovered of the influenza virus, and which cannot be infected again with the same virus. Here μ is the birth rate, whereby at all times there are new individuals introduced into that are susceptible to get the influenza virus. Also, γ is the recovery rate, n is the population in the model, which in this simulation is set to 200, and β is the infection rate. In this run of the model, the time is given by weeks. The

infection rate is given by:

$$\beta = 0.1 \cdot 0.17 \cdot \left(1 - \cos\left(\frac{2\pi t}{104} - \frac{\pi}{5}\right)\right) \tag{4.4}$$

4.1 Northern and Southern Hemisphere

In this simulation, the northern hemisphere and the southern hemisphere were simulated separately by using a phase shift in β, which corresponds to a difference in season. Defining the equation above, Eq. 4.4 to be for the Northern hemisphere, gives us therefore that the infection rate in the Southern hemisphere to be given as:

$$\beta = 0.1 \cdot 0.17 \cdot \left(1 - \cos\left(\frac{2\pi t}{104} - \frac{\pi}{5} + \frac{\pi}{2}\right)\right) \tag{4.5}$$

From these two simulated data, one can simulate E-SE Asia, by taking the superposition of these two simulations. Since E-SE Asia is a part of the tropics, adding these two together well-describes the influenza incidence in the region, at least with respect to seasonality.

4.2 The Simulation

In this simulation, we have included a random process to describe viral evolution. This is modelled as a part of the birth rate, since viral evolution represents recruitment of susceptible in the same way as births. The waiting times between random jumps in the birth rate is chosen to be exponentially distributed with parameter $\lambda = \frac{1}{50}$ weeks^{-1}, and the jump sizes are drawn randomly from the numbers 5, 10, 15 or 20 (which must be seen in relation to the total population size in the model $n = 200$). We also have a deterministic birth rate in our model.

To describe the E-SE Asia leading pattern, we model the transmission of the influenza virus from Asia to Europa as a result of human movement. We introduce a threshold, so that there has to be a certain number of infected humans travelling from Asia to Europa to produce a European influenza epidemic. If we assume that human movement is approximately constant in time, this implies a threshold on the incidence in E-SE Asia. Once the threshold is exceeded, the virus type that is currently causing influenza cases in E-SE Asia is established in Europe. A consequence of this modelling approach is that viral evolution in Europe appears more abrupt than in E-SE Asia.

4.3 The Aim of Doing This

One of the aims of this thesis is to explore how this proposed mechanism fits with observations, and to see if European influenza epidemics can be predicted from E-SE Asian influenza incidence. Or alternatively, that despite a causal mechanism, the dynamics is chaotic, and hence unpredictable. To investigate the latter question, we will run whole the model repeatedly with slight perturbations in the initial conditions, and the exact same realizations of the stochastic birth rates.

5

Community Structure

As we know, many different systems takes a network form, a set of nodes or vertices that are joined together in pairs by links or edges. Some examples of this is social network such as acquaintance networks and also collaboration networks. Other examples could be technological network such as the Internet, and power grids. Other networks are neural networks, food webs and metabolic networks [Girvan and Newman, 2002]. Example of different networks is shown in Fig. 5.1 and in Fig. 5.2

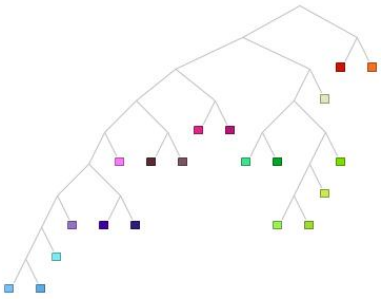


Figure 5.1: A small example of a clustering tree

5.1 Networks

The most recent research on network has focused on a number of distinct statistical properties that most of the network seems to share. One network that has this property is the *small world effect*. In this network the finding of average distance between vertices in a network is short, most usually scaling it logarithmically with the total number n of vertices. Another important property that many networks have in common is clustering, or network transitivity. This is a property where two vertices are neighbors with the same third vertex have a high probability of also being neighbors. This effect of clustering is quantified by Eq. 5.1 [Girvan and Newman, 2002]

$$C = \frac{3 \times \text{number of triangles on the graph}}{\text{number of connected triples of vertices}} \tag{5.1}$$

And this number, C, is exactly the probability of two of one's friends are friends themselves. If it is 1, then the graph is fully connected, everyone knows everyone. In many real-world networks, is it somewhere between 0.1 and 0.5 [Girvan and Newman, 2002].

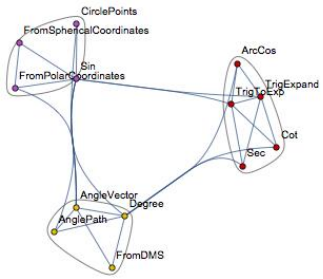


Figure 5.2: An example of a community structure of Cosine

In the paper, *Community Structure in Social and Biological Networks* from 2002, they look at the property that appears to be common in many networks, which is the property of community structure [Girvan and Newman, 2002]. Regarding a social network, a network with friendship or other acquaintances between individuals in this network. In this community there will be communities within it. Subset of groups within the network which are more dense, but between

which connections are less dense [Newman and Girvan, 2004], which is exactly what we see in Fig. 5.2.

Having the ability to detect these community structure in a network would clearly give practical applications. Different communities could represent real social groupings, where this could be background or interest. In this paper they are giving a way of detecting community structure and to apply it to the study of a number of different social and biological structures. The result from this paper is that when it is applied to networks that the information about the community is given, it gives promising results that may help to better understand the relationship between network structure and function [Girvan and Newman, 2002].

5.2 The Traditional Methods

The traditional method to detect community structure which is seen in Fig. 5.2, is hierarchical clustering. What is done first it to calculate the weight, W_{ij}, for every pair i, j of vertices in the network, which to some extent represent the distance between the vertices. After this, one takes the n vertices in the network, with no edges between them, and adds edges between pairs one by one, in order of their weights, starting with the pair that has the strongest weight, and then continuing to the weakest. As the edges are being added to the structure, the resulting graph will show a nested set of increasingly large components which are taken to be communities. There are many different weights that have been proposed to be used in this hierarchical clustering, which in some cases they give reasonable results, and in others where they are less successful [Girvan and Newman, 2002].

5.3 Edge "Betweenness" and Community Structure

In this paper from 2002, they avoid the shortcomings of the hierarchical clustering method, and they proposed another method. Instead of measuring the edges, and which is most central to the community, they focus on the edges that are least central, the edges that are most "between" communities. In this method they rather than construct communities by adding the strongest edges to an initially empty vortex set, they construct them by progressively removing edges from the original graph [Girvan and Newman, 2002].

Vertex betweenness has been studied in the past as measure of the centrality and influence of nodes in networks. The first definition of betweenness centrality of a vertex i is the number of shortest paths between pairs of other vertices through i. This is a measure of the influence of a node over the flow of information between other nodes, and especially in cases where information flow over other a network primary follows the shortest available path [Girvan and Newman, 2002].

To find the edges of a network that are most between other pair of vertices, they have generalized the definition stated above, and define the edge betweenness of an edge as the number of shortest paths between pairs of vertices that run along it. If there is more than one of the shortest way between a pair of vertices, each path is given equal weight such that the total weight of all of the paths are unity. If a network contains communities or groups that are only loosely connected by a few intergroup edges, then all shortest paths between different communities must go along one of these few edges, and therefore the edges that are connected to communities will have high edge betweenness. When these are removed, the groups will be separated groups form each other and will therefore reveal the underlying community structure of the graph [Girvan and Newman, 2002, Newman and Girvan, 2004].

From this we get the following algorithm:

1. Calculate the betweenness for all the edges in the network.

2. Remove the edge with the highest betweenness.

3. Recalculate betweenness for all the edges affected by the removal

4. Repeat from step 2 until no edges remain.

5.4 Application

The authors of this paper, tested their algorithm, and the result of their testing indicated that this method is sensitive and accurate method for extracting community structure from both real and artificial networks [Girvan and Newman, 2002].

5.4.1 Collaboration Network

They applied their application of the community structure on a collaboration network of scientist at the Santa Fe Institute. This network includes all journal and book publications, and along all papers that where published, and they looked at if they had any co-authors. Their algorithms split their network into a few strong communities. Their algorithm found two different communities, the scientist that grouped together because of similar research topics, and grouped together because of methodology. The last group, methodology, is more interesting, and it may be the mark of truly interdisciplinary work. One example is the grouping of those who are working on economics and those who are working with traffic models. That these are collected in the same groups, might be surprising, but when one realizes that these have quite the same technical approaches, it is not so surprising [Girvan and Newman, 2002].

5.4.2 Food Web

Applying their algorithm into the food web, which contains 33 vertices representing the ecosystem's most prominent taxa. Their algorithm found two well defined communities of roughly the same size, plus some vertices. This

network is divided into those who dwell near the surface or in the middle, and those who dwell near the bottom [Girvan and Newman, 2002].

5.5 Community Structures in This Thesis

5.5.1 Influenza Data

The first thing that need to be done so that we will be able to analyze the data is that one need to make the *tweets*, into a time series of each of the cities of where the Twitter data were collected. Which in this case, is days after 1st of January on the x-axis, and how many *tweets* that were collected on the y-axis at the specific time. The data need to be smoothed by using a moving average, by using an averaging run of 7.

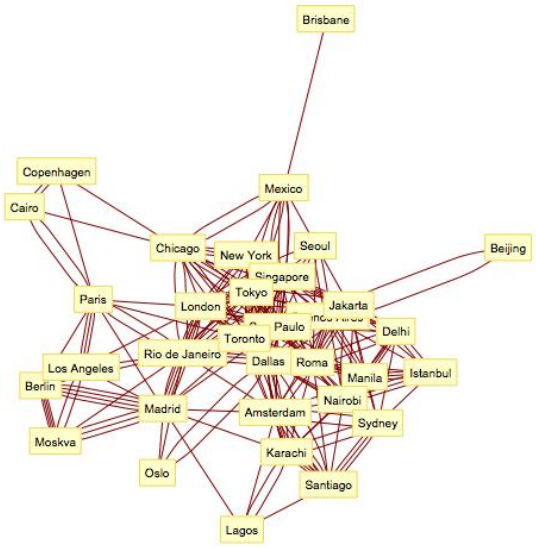


Figure 5.3: Lines between cities using correlation

In this thesis we are using the Twitter data about influenza to make community structures. To analyze this data, different methods have been used. One way to analyze it and to make it into a community structure, is to look at the correlation between two cities and when the influenza *tweets* starts. In this

thesis and this case, we make one line between two cities if the correlation is over 0.65, two lines if it is over 0.75, three if it is over 0.85 and four if the correlation between two cities are over 0.95. Which is done by first collecting all of the cities that have a correlation that is over 0.65, and then make a line between them. After this, the cities that has a correlation over 0.75 get collected from this group, and a new line is drawn between them. This continues until only cities with over 0.95 in correlation is collected and gets the fourth line. This is done so that the cities with a stronger correlation will have something more to say in the community structure.

Another way to analyze it is to look at the peak of the *tweets*, or in other words, analysis that is based on the timing of maximum of each city. In this way to analyze it, one looks for how much the different days of the peaks differ from city to city. If it is the same day, we see more lines between them, with a maximum of 5, and if it differs with many days, than we only see one line between cities. The way to do is, or at least here, is that we take the maximum of each city from the data and take the difference with it and another city. After this we divide these into different groups depending on what the difference is. A difference that is 0 gives 5 lines, while a difference with over 5 days, give 1 line between them. And after this, the Community Structure is made.

Figure 5.4: Lines between cities using timing of maximum

How the arrangement of these two different analysis look like, can be seen in Fig. 5.3 and Fig. 5.4. Where Fig. 5.3 is where the analysis is based on correlation, while in Fig. 5.4 is based on the timing of the maximum.

In this thesis, real life influenza data from over the whole world in a community structure plot is of interest. In which case we are looking at where countries coincide with the influenza season for 3 or more years. The season is defined such that we first calculate the onset, $\lambda = \frac{x_{t+1}}{x_t}$ in 20 weeks window. And when λ change it sign from negative to positive the season has started. One line is made between two countries if the onset is within the same three weeks.

After this, put arrows from a country to another if the influenza season for the different countries are following each other in 3 years or more. Which tell us more about how the influenza virus is moving. In this analysis empty values of infected persons, that is, where there is no data, these dates of that country get removed. Not every country started at the same time, and some of the data is missing. In Fig. 5.5 one can see the countries where the influenza season is following another country for some time.

The community structure of this issue, is divided into three groups. Where one can see which countries are more related with each other than with others.

To be able to analyze the influenza data better for all cases, the following groups will be color coded, and each city or in the influenza data set case, country, are marked on the world map with their respectively group color. Such that the city, or countries, is marked on the map with the color of which group from the community graph it belongs. A city which belongs in one group that is, for example red, will be seen as a red dot on the world map. Therefore, making in much easier to analyze.

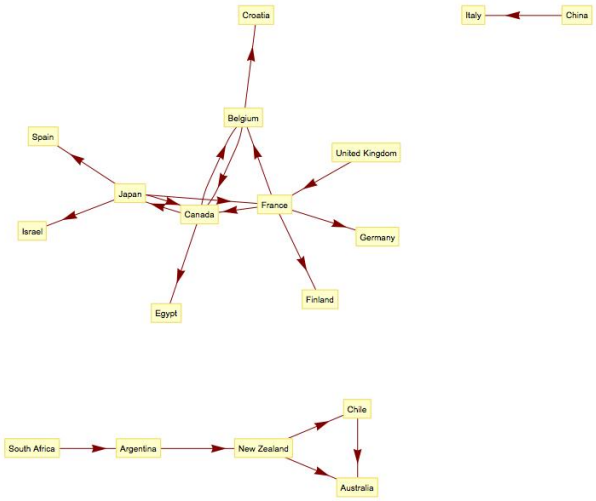


Figure 5.5: Influneza data with lines between the countries

From the Fig. 5.5, we see that the pattern gives us three different groups. One

big, one that is small with only 2 countries, and then a group in between them, with 5 countries. This figure could indicate what type of pattern influenza has.

5.5.2 Flight Data

When making the community structure for the flight data, we take as many lines between each city as there is direct flights. Meaning that where there are more direct flights, it will lead to a closer connection, and therefore it will, with a higher probability be in the same group in the community plot.

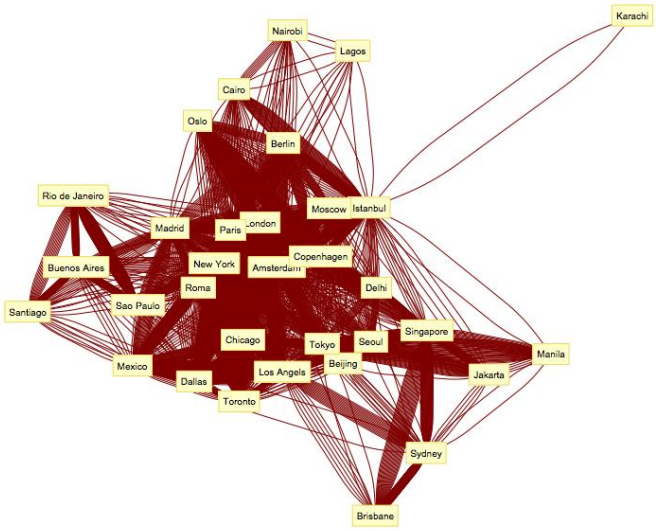


Figure 5.6: Flight data with lines between the countries

To get this plot, Fig. 5.6, the flight-matrix we made is used. This is used to get the number of direct flights between cities. From this figure, we see that each line corresponds to one direct flight. From this plot, we can see reading the number of direct flights is much easier from the matrix. Since the cities are clustering in the middle.

The flight matrix is found in the appendix, Fig. C.1.

6

Results

6.1 Time Series

The time series of the Twitter influenza is in the appendix, Fig. A.1

As we would suspect there are big differences in the number of *tweets* from each city. Los Angeles has a maximum of 600 *tweets*, while Lagos, and Cairo has a maximum of 1. Nairobi is the city with the highest number of *tweets* at one day with 1000 *tweets*. While, Paris is a city that has a peak of 150 *tweets*. These differences continue with all of the time series.

In some of the time series of the *tweets*, it is difficult to see the pattern, as it has one or two days with high peaks, compared to other days. Like Berlin, it has one day where it peaks at 600, and a small peak some days before. Because of this, it is difficult to see the even smaller peaks, so that it looks like zero, even though it is not. Whereas at cities with not as high peaks, it is much easier to see the difference day to day, as they are not too different from each other.

Nairobi only have one peak at this time period, while Oslo only have two peaks in this time period. And we can see, that Sydney has three. There are a few

other cities that also have a few number of *tweets*, see Fig. A.1.

What we can see from these time series of the number of *tweets*, it that most of the peaks happens around 60 days after January 1st, which in 2018, is March 1st. Some before, and some after. The reason for this, is because of when downloading began. All of the *tweets* ended a short time after 80 days, when downloading was no longer available to do, for some unknown reason.

The cities with the most *tweets* are Nairobi, Los Angeles, Berlin, as mention above, Paris, New York, and Tokyo.

6.2 The Community Structures

From the Community structure plots, we get the following:

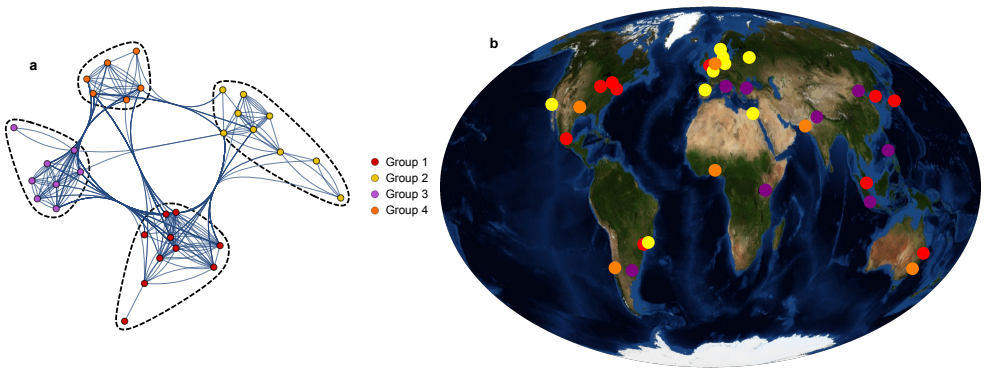


Figure 6.1: Community Structure based on the correlation

In Fig. 6.1, we see what community structure we get when we are looking at the different correlation between *tweets* from different cities. Where, as stated in the previous section, there are four lines the correlation is over 0.95. Three lines are a correlation for over 0.85, two lines for 0.75. And finally, where there is one line, the correlation is over 0.65. The number of lines between cities, can be seen in Fig. 6.1a. What one may see in Fig. 6.1b is that group 1, the red dots, are placed in the eastern part of Northern and Southern America, eastern part of Asia and Oceania, and with one city in Europe. Group 2, yellow, is placed mostly in central Europe, with Cairo, Los Angeles and one city in

South America. Group 3, is placed mostly in Asia and South-East Europe has two cities. The group also have one city in the South of America. Finally, group 4, the orange dots, are placed in some way randomly over the whole world, in every continent.

In Fig. 6.1a we see that every group is dependent on each other. While group 2 and 3 are barely connected, as it looks like only two lines between these groups. The other groups are more connected with each other. Which of the groups that are most connected, is difficult to see from the plot. But group 2 and 4, may seem to have a bit weaker connection than with the others.

In Fig. 6.2 we see the community structure if we are looking at the time of the maximum of the *tweets*. And which is also stated above, the lines between each city, are determined on how close the maximum of *tweets* are to each other.

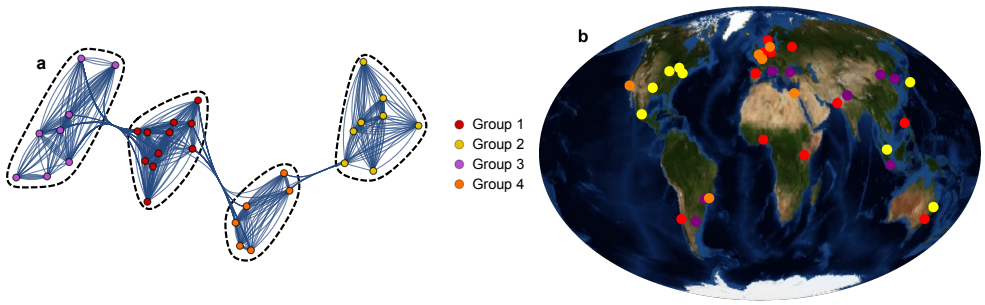


Figure 6.2: Community Structure based on the time of the maximum

In Fig. 6.2b group 1, the red, is placed mostly in Europe, but also cities in Africa, Asia, and Oceania. This group also have one city in South America. Group 2, yellow, is placed in the eastern part of Northern America, and east in Asia and Oceania, this group does not either have a city in South America, which is the only group in this case that does not. Group 3, purple, is mostly in Asia, with two cities in South America. The last group, is placed in Europe, with one city in Northern and Southern America, one, and one in the northern part of Africa, Cairo.

When one look at Fig. 6.2a, one can see that group 2 and group 3 are not directly connected. Indirectly since group 3 is directly connected with group

1, which is directly connected to group 4, which in turn, is directly connected with group 2. It does look like group 1 and 4 are directly connected with two groups, while group 2 and 3, are only directly connected with one other group. It does not look like other groups are directly connected with each other, as we do not see any lines between these groups.

In Fig. 6.3 we are looking at the community structure of real influenza data by country from over the whole world. In Fig. 6.3b one can see that the countries in the Southern Hemisphere, and with some countries in Africa, are connected, since they have the same color. Which is what we could expect, to some extent, as the southern hemisphere, typically has the influenza season on a different time than the northern hemisphere. The countries in Africa that is not a part of the southern hemisphere, is close to the border. What one further can see is that, Asia and eastern part of Europe is connected from the influenza data, and that west Europe and Northern America is connected.

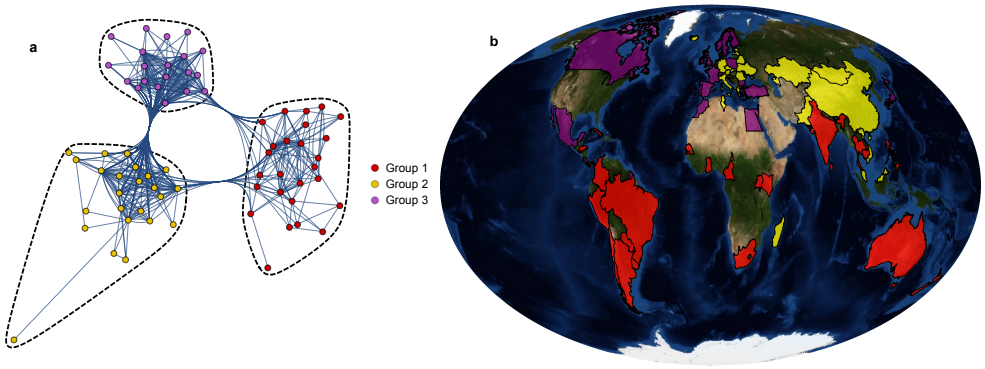


Figure 6.3: Community Structure based on real influenza data

What is very clear in Fig. 6.3b, is that there are, in some way, two separate groups, for the Northern and Southern Hemisphere. Which, is of course, what one sees in real life. As one sees in Fig. 6.3a is that all of the groups are connected, where group 2 and 3 have a stronger bond between them, then what these groups have to group 1. Which is impossible to decide by just looking at the community structure plot Fig. 6.3a.

In Fig. 6.1, Fig. 6.2 and in Fig. 6.3, the community structure is the one to the left, part a, whereas the one to the right is the color coded map of the world, part b, so that we can see which area belongs to which group.

6.3 Flights and Community Structure

As we can see from the matrix containing all the flight data there are not surprisingly many zeros, Fig. C.1. Some cities almost do not have any direct flights, as Brisbane and Karachi. Which is not surprising. Both if these cities only had direct flights from and to Sydney and Istanbul receptively. Which were the only cities which has direct flights in this group of cities. The number of direct flights between cities can also be seen in Fig. 5.6.

Another thing that we see from the matrix is that the cities that are closer in distance are more connected, as we would suspect. The big cities have more direct flights than the smaller cities, as we would suspect. And from the matrix we see that there is a high probability that there are more direct flights within the same continent. As mention, the matrix for the flight data is in the appendix, Fig. C.1.

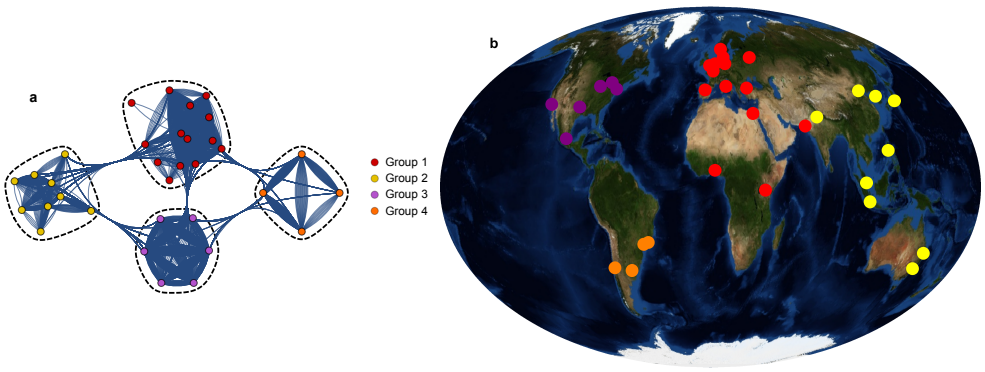


Figure 6.4: Community Structure based on flight data

In the community structure Fig. 6.4, we see that, and not surprisingly, that each city in the same continent is in its on group. Or at least for almost every group. As Africa is in the same group as Europe and the Middle-East. All in all, this is not a surprising plot considering how it is divided, as we would guess that there will be more flights in the same continent. And thus, more connected. This was something we already saw in the matrix of direct flights. As we see in Fig. 6.4a, the groups a very close connected with each other. There is also a strong bond between the group, but of course not as strong.

In group 1, which is red, which is consisting of Europe, Africa and the Middle-

East, is the group which is mostly surprising together. As there is quite a distance from Europe to the two cities in the middle of Africa. Group 2, the yellow group, is the group that consist of cities in Asia and Oceania. Group 3, the purple group, consist of cities in North America. And finally, group 4, which is orange is South America.

Group 1 is the group with the most cities, where group 2 follows, then group 3, and at last group 4. But, as we can see in the community structure plot in Fig. 6.4a, is that all of the groups are connected directly with each other, except group 3 and 4. As we could not find any direct flights between these regions.

6.4 Flight Data with Correlated Twitter signals

To get a better look at how the flight and Twitter data are connected, a fitted linear model with these data sets, direct flights and the correlation between cities, was made, and it gave the following results, with the best fit parameters:

$$\text{lm} = 0.462294 + 0.005131x \tag{6.1}$$

With the following Parameter Table:

	Estimate	Standard Error	t-Statistics	P-value
1	0.462294	0.0122046	37.8788	$2.07872x10^{-152}$
x	0.00513135	0.00129942	3.94896	0.000089178

Table 6.1: The Parameter Table for the fitted linear model

Fig. 6.5 is the plotted fitted linear model, where Table 6.1 tell us the parameter values. In the plot, the red line is the fitted linear model, whereas the red dots are the mean between ten direct flights. Where the first dot, is the mean for the ten first, the second, for the next ten etcetera. The small black dots are the actually number of direct flights between each city. Which is why there are

many small black dots at zero direct flights. The p in the plot, indicates that the slope has a p-value that is less than 10^{-4}

As we can read form the figure Fig. 6.5, we see that on the x-axis, is the number of direct flights between the chosen city. And the y-axis is the correlation between the influenza Twitter data, or signals.

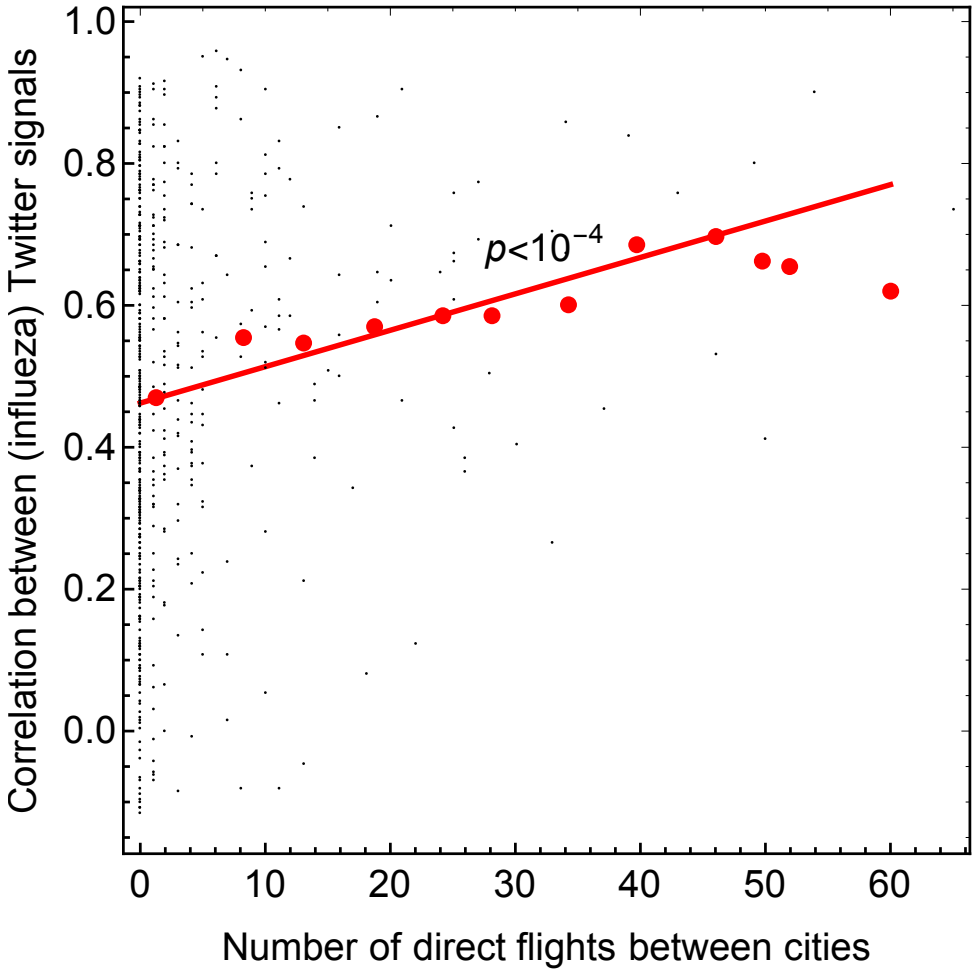


Figure 6.5: Fitted linear model of direct flights and influenza Twitter signals

6.5 The SIR Model

What we would like to see from this model, is that the new simulated data from Europe, would follow the simulated data from Asia. Which we do not see in Fig. 6.9. Trying different values from the parameters, does not either give the result that we would like to see. The new simulated data of what should represent Europe, Fig. 6.8 is simulated using what we would like to call E-SE Asia, Fig. 6.7, does not either follow the old simulated data from Europe, Fig. 6.11, the one that E-SE Asia is simulated on, even though it does so closely in the beginning. That is for a small period of time, the original and the new simulation of Europe follow each other. After this we do not see a distinct pattern in either of the two cases, everything looks random.

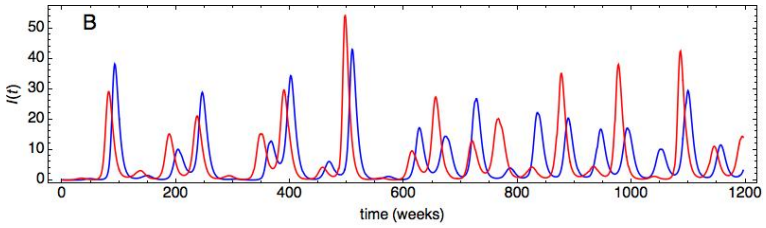


Figure 6.6: Simulated influenza data of Northern (blue) and Southern (red) Hemisphere with $\mu = 0.4$

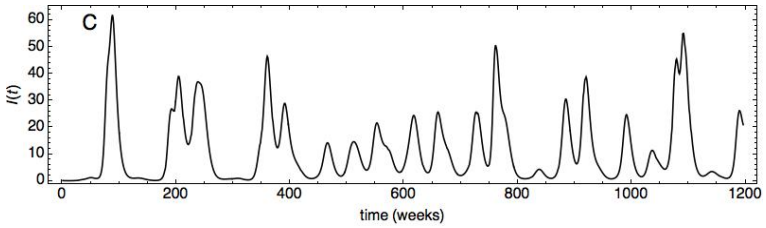


Figure 6.7: Simulated influenza data of Asia with $\mu = 0.4$

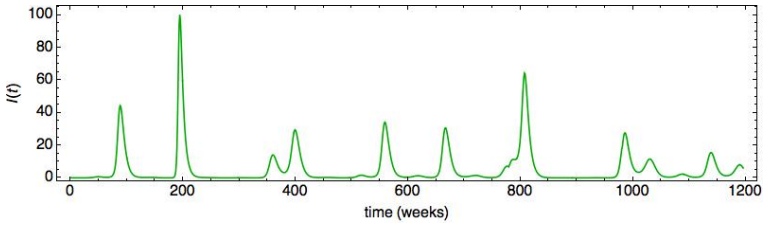


Figure 6.8: Simulated influenza data of Europa with $\mu = 0.4$, using the simulated data from Asia

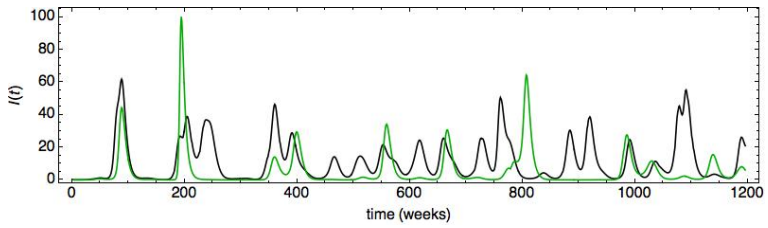


Figure 6.9: Simulated influenza data of Europa (green) and Asia(black) with $\mu = 0.4$, using the simulated data from Asia

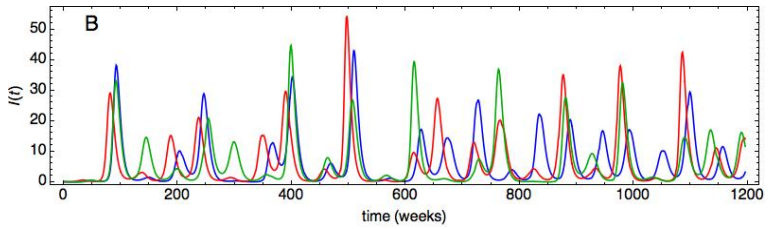


Figure 6.10: Simulated influenza data of "Old-Europa", "New-Europe" and a region in the Southern Hemsiphere and Asia(black) with $\mu = 0.4$, using the simulated data from Asia

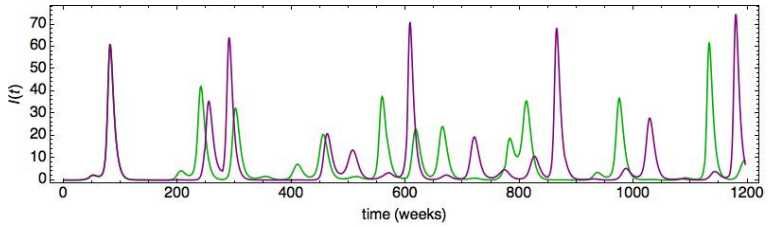


Figure 6.11: Two simulated data from the same region with a small perturbation in the initial condition in the purple curve

All of this could be reminded of chaos in the system. And therefore, this is also simulated to see Fig. 6.11 and Fig. 6.12. From this simulation we can see that we do get the characteristic pattern of a system with chaos. It does follow each other for some time, and then not for other times. It might follow each other again, but it will always diverge from each other. In Fig. 6.12, the model is ran with μ=0.7 for all cases, for both simulated Asia and simulated Europe. One can see in this figure that these two runs follows each other for some time, then differ, and then follow each other again, as we saw in the previous plot as well Fig. 6.11 From this we get the characteristic pattern of a chaotic system, and we thus have a chaotic system for this model.

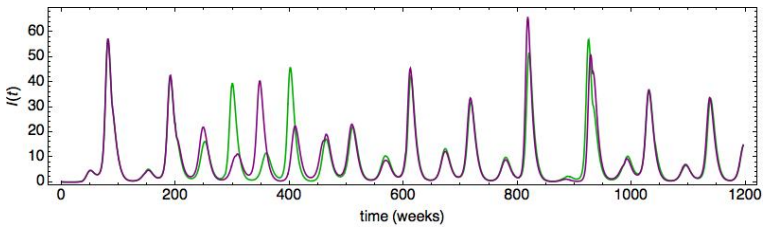


Figure 6.12: Two simulated data from the same region with a small perturbation in the initial condition in the purple curve

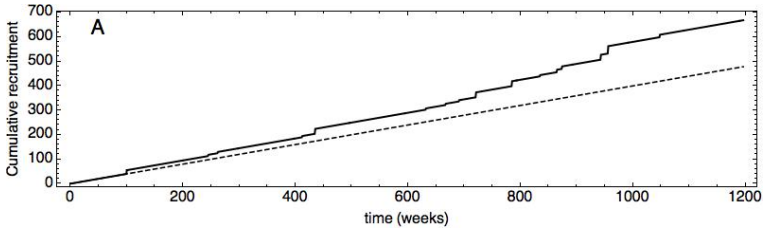


Figure 6.13: Birth/recruitment rate

In Fig. 6.13 one can see the affect in having a birth rate that has these jumps. The dashed line will be where there is a constant birth rate, whereas the solid line is the birth rated with the jumps, as we can see. And this is the one that is used later on in this model.

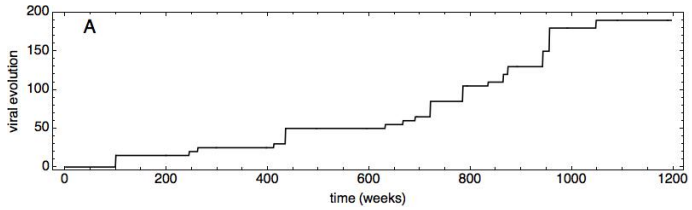


Figure 6.14: Birth/recruitment rate for Asia.

In Fig. 6.14 the viral evolution is found by subtracting the birth rate and the birth rate with jumps with each other.

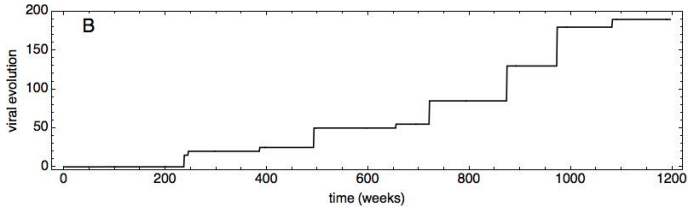


Figure 6.15: Birth/recruitment rate for Europe

In Fig. 6.14 and Fig. 6.15 one can see the difference a threshold of people coming from Asia with the virus to infect the people in Europe has to say for the birth/recruitment rate in Europe. The threshold makes so that we get less and larger jumps as we can see in Fig 6.14 and Fig. 6.15. A larger threshold would make less jumps, and thus a smaller threshold would make more jumps in the birth/recruitment rate.

7

Discussion

7.1 Twitter-Data

The reason for the small amount of data, is that we were not able to download it the way we did it before, for some unknown reason. Although the downloading stopped, the ILI-*tweets* that was downloaded, was enough to analyze it in some way. The data analyzed where approximately from February 12th to March 27th. Which means that we were able to download data for about 6 weeks.

One of the risk of using Google Translate as a translator is that it is known to make mistakes. It is not given that the input will produce the right answer, and as not every language is known, one cannot guarantee that every word is correct. But, as there were only made searches for simple words, we can assume these are the real words. In addition to this, it is known that Mathematica managed to understand Japanese *tweets* who said something about influenza, even though it was not written in English. This did not happen in Norwegian, so the choice to translate words to get more of the *tweets* were made. The reason for this, that Mathematica understood the Japanese *tweets*, could be due to the fact that it is a more used language than Norwegian, and that Mathematica has their own edition in Japanese. That is, Mathematica do know Japanese.

It was not tested, if Mathematica could understand other languages than Japanese, since we think that it would be much easier to translate the words into the first language in every country. Than to go through all the *tweets* to see if this are being collected as well.

Another thing that was discovered, is that Mathematica was able to understand a *tweet* that were written in another language, if it mentioned an article that contained the English word that we searched for, if it was in its title. So that we were able to download *tweets*, that were non-English, even though we did searches in English.

As Twitter is most used by young people, rather than the oldest and the youngest, we get data from the population where there is the smallest proportion of complications among them. As well as the proportion where it is less likely will get the influenza virus confirmed. This is a large group, and as time goes by, more and more people use social media. Which can lead Twitter to become even more representative for influenza data in the future.

In Fig. 6.1, we see that most of group 2 is in central Europe which is in all, 6 cities. The three others in this group are Los Angeles, Cairo and Rio De Janeiro. That the 6 cities in Europe are grouped together, is no surprise, that these cities are more correlated. It is stranger that the other three are in the same group. If we look at the flight matrix, Fig. C.1, and flights from and to Los Angeles, we see that there are some flights going from and to this city with Europe, and in one day, there are 148 flights. This might explain the grouping. Which also is a bit interesting, is that even though we only have few weeks of data, one can see that west Europe and North American is connected in both Fig. 6.1 and Fig. 6.3. This do we also see in Fig. 6.2.

The other groups in North America, group 1 and 4, do have a direct connection with group 2. Where it is a stronger connection between 1 and 2, because of the darker color we can see between them. From this figure, Fig. 6.1, we therefore can see some interesting tendencies that we also see in real influenza.

In Fig. 6.2, where we are looking at the maximum of peaks. It is safe to say that the small amount of data collected, affects the results. Which could indicate

the result we get when looking at the time of the maximum, as these looks a bit random.

What we really would like to see, would be that E-SE Asia would be grouped together. And then North America, Europe and Oceania in another group. And then South America alone in the last group. If this was the case, we would be able to see the migration pattern as it possible is. But, this is not the case. And getting this grouping could potentially take several years.

In the figure, Fig. 6.2, and the fact that group 2 and 3 are not directly connected, gives an interesting result. We do also believe that group 1 and 2 are not directly connected, with the same property for group 3 and 4. This tell us that some of the cities in Europe, is not directly connected to most of the cities in North America. But, as we can see, is that these cities in Europe, is directly in contact with the other cities in Europe. Which in turn is in directly contact with North America.

Another thing we see in this figure, is that Asia is divided into three. As the groups in Asia, is group 1, 2 and 3. As we can see in Fig. 6.2a, is that group 2 is not directly in contact with the two others. Group 2 is closer to group 1, than to group 3. This is not what we see in 6.3, as Asia in not divided into different groups.

A thing that is interesting, is that Japan, is in the same group as North America in both of these figures Fig. 6.2 and Fig. 6.3. That is that Tokyo/Japan is separated from Asia in both of these. Since in Fig. 6.2, Japan is in group 2, which is not in directly contact with group 1 or 3, as stated above. And what we see in Fig. 6.3, is that Japan is in the group with Western Europe and North America. This is something we also see in Fig. 6.1, but in that structure, Tokyo is not alone in Asia. As stated previously, Japan is not a part of the E-SE Asian migration pattern. Thus, that Japan/Tokyo is not a part of the Asian group, could be explained by this.

7.2 Influenza Data

In Fig. 6.3, we see that Oceania and South America is in the same group. And thinking about the seasonality, this makes sense. It often occurs in the winter season, and since they are in the same hemisphere, the season of influenza in these two continents will coincide. The cities in Africa, that is not in the Southern Hemisphere, are quite south and close to the border. This makes it more reasonable to be in this group.

If we think about the leading pattern and antigenic part of influenza, it says that it will first come to Oceania and then to South America, which it might be. But, which of these two continents it will first hit, is impossible to see from this plot, as this plot does not tell us anything about time. Only which of the groups that have more in common when the influenza season hits. But something that is interesting in this picture, is that Asia, with some countries in Europe and Africa, is alone in one group. If we do not look at group 1, we see that Asia and some countries in Europe, is alone, and the rest of Europe with North America is in another group. This is something we would like to see. This could mean that we, somehow, can see that the virus evolves and occurs as a new virus, in Asia. It would be even better to see if we could have some information about the time of the peak in these groups. Which we also see in this, is that group 2 and 3 are strongly connected with each other. More than with group 1.

As we can see from the same figure, Fig. 6.3, is that the countries in group 2 are mostly countries in Asia with some countries in the East of Europe. Since Europe is divided into two, and most of the east part is connected with the group from Asia, we can assume that these countries get the influenza virus before, as they are closer to Asia, than the others. This does not sound to unlikely, that something that are closer to the origin, get infected first. Maybe most flights have a stop in these countries, and from there it spread.

7.3 Flight Traffic

Considering the fact that only a certain date was used to collect the flight data, it is not a guarantee that all of the direct flights between cities were discovered. That is, another date could have more direct flights, or less flights. And the fact that it was not a date that is normal in the influenza season, since past flights are not available at Expedia, and the fact that next season will not hit for a long time, at least for most of the chosen cities. The direct flights still make a good representation of flight traffic between cities, as it still say us something about it. That is, which cities that have more visits from different cities, and the connections between them.

It is safe to assume that if there is one direct flight one way from a city to another, it will return to the origin city after this, as this is something that occurs frequently. But, of course since this is an assumption there could be an error. But, we are thinking this would be small, as it makes sense that the flights will go the other way as well. Another error that could occur, is that it might be that the flight will return the next day. But, this will not be for absolutely all the flights, only a small portion of them, if it even happens. Even though it flies back the next day, which can have a different number of direct flights, it still goes back. Thus, we can make this assumption.

Taking a random day and looking at direct flight does not need to show the whole picture of direct flights. The number of flights change over the year, and in the summer, where more people take flights and travels, we could make the assumption that there will be more flights. One could make the assumption that there will be more flights in other seasons as well, that could be for example, Christmas. As more people travelling, the need for available seats increases.

Expedia[1] is a web page that use an algorithm to go through all available flights between cities. This was chosen to use, so that we did not need to go through all the airline companies to see if there was direct flights between cities. As it only shows available flights, there could be more direct flights between cities. It

1. https://www.expedia.no/

may as well be commercial flights flying with cargo, rather than people. But as these flights do not transport as many people, the probability of infecting many people, are much less, and thus it will not have much to say in the circulation of influenza compared to commercial flights. These planes do not have the same contact with people in the same extent as planes with people.

As we were only interested in flights between those cities where we had managed to collect Twitter data, we did not look for flights that stopped in different cities. This is much easier and therefore less to collect. If there was a stopover in one of the cities that we were interested in, it would be picked up later on, when we looked at those two cities. Because of this choice, of only looking at direct flights, leads to the fact that there will be many elements in the flight matrix that is zero.

The flight matrix tells us how many flights one can assume flies directly from one city to another. There is no surprise that the biggest airports are the ones with the most flights, and that most of the flights between airports are in the same continent. Which we also see in the community structure of flight transport, where each group is in the same group. Expect Africa, which is included in Europe. This makes sense as we can see in the matrix. The direct flights that goes from Africa goes mostly to Europe and the other two countries in Africa.

One surprise, in some way, is that in the community structure, is that Karachi is in the Europe group as it is much closer to the Asian group. But, if we look at the flight matrix again, Fig. 5.6, we see that the only direct flight to Karachi, is from Istanbul. If we had chosen different cities in Asia, we might have seen a different picture as there might be other direct flights into Asia.

What we see in the community structure, Fig. 6.4a, is that South America and Asia is not connected, that is, no direct flights between them. This was discussed in [Russell et al., 2008a], where they believed that the reason for South America's late new influenza introduction, was because of little travel and trade between these continents.

7.4 Comparing Flight Traffic and the Influenza Data

If we look at the different community structures of the Influenza data, Twitter data and flight data, we can see that the groups are different. And that even though there are similarities, they are not that easy to see in every case.

7.4.1 Correlation Twitter Data

Comparing these two structures Fig. 6.1 and Fig. 6.4, we may see that in these figures, they do not completely show the same picture. As we can see in Fig. 6.1a, that even though they are in different groups they are all connected. But as mentioned in the previous section, group 2 and 3 are poorly related compared to the others. Using the fact that the groups are related, one sees that almost all of the cities in Asia and Oceania are related with each other, with some correlation. This can we also see in the flight data, Fig. 6.4, where all of the cities in Asia and Oceania are related. Another thing we see, is that group 2, mostly in Europe, which in Fig. 6.4, is alone in one group.

7.4.2 Time of Maximum Twitter Data

In these two different structures, we can see that Africa is connected with cities in Europe in both of these. Even though we can see that not all of the cities in Europe is in the same group. Another thing that also is similar in these two structures, is that North and South America is separated in both of these.

Another thing that these structures have in common, is that, Karachi and New Delhi is in two different groups, even though they are close to each other. This is actually something we see in all of the figures. Expect, of course Fig. 6.3, as this one looks at the countries and not cities.

7.4.3 Actual Influenza Data

One problem comparing these two figures, Fig. 6.3 and Fig. 6.4, is that we have only three large groups in the Influenza data structure, and four groups with a small portion of cities in the flight community structure.

If we look at these groups in these two plots, Fig. 6.3 and Fig. 6.4, one can see that the groups are quite different. As Africa, South America and Oceania are in one group in Fig. 6.4, while they are in three different groups in Fig. 6.3. As we see from these two figures, is that some cities are in the same group in both of them, but not all, which does not come as a surprise. In 6.3, Europe is divided into two groups, while in Fig. 6.4, it is only one group.

In the community structure of the flight data Fig. 6.4a, we see that South America and Asia are not connected as mention. In Fig. 6.3a we see that these continents are not as closely connected as Asia is with Europe. But in this, there are big areas in the same groups. This could affect it.

The flight data, which in some way could be seen as an average, which it is clearly not. As there is only data from one day. There could be changes in the number of flights over the years. And comparing cities with countries, and with the fact that the influenza data is for over many years, is not the best thing to do.

7.4.4 Linear Model of Direct Flights and Twitter signals

As we can see from Table 6.1, and Fig. 6.5 the p-value for the slope, is quite small. This tell us that it is statistically significant for the model. Which we also have for the y-intercept, that tell us again that it is statistically significant for the model.

As we can see in this figure, Fig. 6.5, there are especially three mean dots, that are further away from the linear regression line, called outliers, with the last dot being the mean with the furthest apart from the regression line. But, still with the p-value being so little, the regression model is a good one.

What we can see from the same figure, Fig. 6.5, is that the higher number of direct flights between cities, the higher number of correlation of influenza between those same cities. This tell us that number of flight do have an impact on influenza cases. And that it can be a part of the influenza migration, and how it spreads.

Even though it can be difficult to see it, from looking only at the community structure in Fig. 6.1 and Fig. 6.4, we see that there is a dependence between them from Fig. 6.5.

As this is a fitted model, we made some assumptions in the beginning, which the validity depends on. That is, constant variance, where the points should be constant distributed around the mean, and we have that the residuals should be normal distributed. But, as the p-value is so small that it is in this thesis, it is a good indication that the slope is a good fit for the model.

As the data is so small, the black dots in Fig. 6.5, it will be easier to look at the red dots, the average over them, to say something about the residuals. From this we see that the average dots, does not differ to much from the red line. This is a good thing. But as the p-value is so small, as we can see in Table 6.1, it is safe to say that the assumptions are being held.

This tell us that even though we have only a few weeks of Twitter data, it is possible to see a pattern with the flight data. This indicates that a bigger collection of data, should be even better. And that it is possible to use Twitter to get influenza data, but that it needs to be downloaded for a much longer time span, so that we might be able to get something like Fig. 6.3, actual influenza proven data. Which we have seen other studies have managed to show.

7.5 Migration Pattern Continued

7.5.1 The SIR Model

This model was done to see if it is possible to model the pattern of influenza.

From the results from the different plots from the SIR model, Fig. 6.6-6.12, and the fact that it is a chaotic system, it would be impossible to predict when and to what extent an epidemic would be. Even though there is evidence, from antigenic analysis of the virus, that new viruses first emerge in E-SE Asia and then travels to the rest of the world, first in Europe, Oceania and the North of America, and finally the South America, this model could not predict this result, because of chaos in the model. For some cases it might be possible to predict it for some small-time lag, but we would not know when these two would diverge from each other again. It would also be impossible to say something about when and if they would meet again. And therefore, it would be impossible to predict influenza epidemics in Europe using influenza data from E-SE Asia.

Even though the virus itself comes from E-SE Asia, predicting when this certain virus reach Europe is difficult to say something about. Sometimes there is even an epidemic in Europe, that is not in E-SE Asia and vice versa, Fig. 6.9. One could also see that between 400 and 600 weeks in Fig. 6.9, there are two incidents of small influenza epidemics in Asia, one could interpret these two epidemics for being to small, such that it will not come to Europe. But we still see a small one after these two in Europe, that hits approximately at the same time as the third epidemic hits Asia. As we can see in these plots, they do not either tell us anything about how the virus travels between the different regions, since everything seems to be independent of each other Fig. 6.9.

7.5.2 The Community Structures

As there in only a few weeks of data, it is suspected that we will not see the influenza pattern, because of the short time. Even though if we were able to download more data, it is not sure that we would see this. As it might take several years to see it. It might be easier to see it, if we have started to download data from before this year influenza season.

From the community structures, Fig, 6.1 and Fig. 6.2, which is of the downloaded Twitter data. There is no clear evidence of a leading pattern.

In Fig. 6.1, we see that Group 1 and 3, are the ones that are in the E-SE Asia. Both of these two groups have cities over the whole world. And they are strongly connected to each other, as we see in Fig. 6.1a. From this figure, we also see that group 4 is strongly connected with these two groups. As mentioned in the results. This indicates that all of the cities in these groups are more connected with each other, than with group 2. This tell us that that most of the cities over the whole world in connected, expect one city in North America, and many cities in Europe. That group 1 and 3 are strongly connected is a good sign. Since it is in E-SE Asia is believed to be the origin area of influenza.

As in the we can see in the same figure, is that all of the groups are represented in South America. And as it is not typically the influenza season in this region, and by looking at it as noise, we can disregard this.

In Fig. 6.2b, as group 1 is strongly connected with both 2 and 3, which tell us that all of Europe, with most of Asia is strongly connected with, as the leading pattern tell us. As we can see from group 2, is that it is in East of Asia and in Northern America. This could indicate that these are from the same origin. Group 2 is connected with group 4, and therefore it could be possible that it origins from E-SE Asia. As it is not possible to know the time, and the fact that it is only from a few weeks, every peak from every time series, is close to each other. And to be able to see a much better and more believable pattern, would take several years.

8 Conclusion

8.1 Summary

Influenza has a high mortality and morbidity rate across the world, and a decrease in these numbers are something that is wanted. To get these numbers down, prevention need to get better, and the understanding on how it circulates needs to be better understood and known. As young persons that get infected most likely will not go to a doctor and get the influenza virus confirmed. It is difficult to get the total number of the influenza infected population. A confirmed infected person will further take some time, so that other sources of influenza data may be a good idea.

There are known ways of collecting data, and there are many influenza data set on the Internet, some are from health organization but another source of data, is social media. Which is more and more used, and especially Twitter. Twitter gives in time data, where people write about their sickness. These data will be ILI, and it has been showed that it has a good correlation with the confirmed number of infected people [Paul et al., 2014, Doshi, 2008, Agogo and Hess, 2018].

Even though it is difficult to say exactly when an influenza epidemic could hit, we know to some extent when an epidemic could hit, as it often starts in the coldest months as previously stated. But, as always, there are exceptions, and epidemics could start in the summer months, with one example being the swine flu in 2009 [WHO, 2014].

Social media is growing, and more and more people get a public profile where they publish their thoughts and meanings. Which makes it a great and easy platform to get and collect ILI-data from over the whole world, with especially Twitter, because of its open API. The are many studies that have used social media for research, where some of them have been mention in this thesis.

In the future when everything is much better understood, preventing a pandemic would be much easier. This means that the people in the risk groups, would be more capable to prevent it. And with it, less deaths and hospitalizations.

Trying to figure out more about the migration pattern of influenza scientist have tried to determine the origin of different virus over the years. And they figured out that, for the most part, the influenza virus origin form E-SE Asia.

As there is so much deaths and hospitalizations because of the influenza virus each year, and there is also a big cost for the community, the study of the influenza cannot stop. It is important to discover new things and to understand the virus better. How it moves, and changes. How the virus likes the different climate factors, as this could help understanding when the epidemic could hit.

8.2 Conclusion Remarks

Even though we only got a few weeks of data, there is possible to describe and conclude with something. As there is possible to see small tendencies, in the different groups in the community structure. The SIR model gave chaos, which will not be able to use to describe the pattern.

From the *tweets* collected about ILI, there were different numbers of them, which is not a surprise. As the use of Twitter is different for every city. We were still able to see something for every city. As we got data from every city, where none of them had much less data than others, we may assume that the translation is right, and for cities with different languages, the right languages was used.

The community structures are good to use to see patterns more clearly. As it is a way to divide data into groups with common properties. Where it can be difficult to see, only by looking at the data.

The community structures in this thesis, from the Twitter-data, look like they are random for the most part. But we are able to see some patterns that might be right. If we take a look at Fig. 6.1 and Fig. 6.4, one can see that Europe and North America are connected, and in the same group. Since all of the groups in North America, is close with groups in Europe in both of the community structures.

Looking to see if we can find a pattern in twitter-data versus real influenza data is difficult. But we can see small tendencies. If we got a bigger data-set, it might have been possible to see a better picture. That we can see this tendencies after only a few weeks of data, is quite interesting, and it could indicate that this should be continued, and that we will be able to see something after a while.

The flight data in Fig. 6.4, were not the biggest surprise, as most of the cities within the same group is in the same continent. With the exception of Africa, which is in the same group as Europe.

From the results in this thesis, we can see that it indicated that Twitter can be used to say something about Influenza. Even though we only have a few weeks of data, because of downloading problems. Especially looking at Fig. 6.5, where we see that the correlation of two cities is higher if there are more direct flight between them. This tell us that flight traffic could be a part of the spreading of influenza, and that countries that have more with each other to do, more often have the influenza season approximately at the same time.

The difficulty of finding the influenza pattern in the community structures, is for a big certainty because of the few weeks of data, or at least it could be one of the reason. A few weeks make many limitations, and a whole influenza season do last for several months. And even if we had data for the whole season, it might have not shown a clear and distinct pattern.

Even though it was not clear and that it did not tell that much, we know that Twitter can be used to say something about the influenza season for this year, there is evidence that it can by other studies. But as mention, there are tendencies that are quite interesting from only a few weeks. That could be the fact, that Tokyo was not in group with other cities in almost all of the community structures. This could be the tendency of the E-SE Asian circulation pattern, because of what is known about the origin of the influenza viruses and that Japan is not a part of this.

The result from the SIR model told us that it was not possible to predict or to see the migration pattern of influenza, if it comes from Asia because there was chaos in the model. As this makes it impossible to make predictions, and thus not possible to predict the influenza pattern or the origin of it. Chaos in mathematical SIR model have occurred in other studies as well [Glendinning and Perry, 1997]. This does not mean that this is not true. Only that this model was not able to predict it. As from other research studies have shown the pattern, using antigenic and genetics of the influenza virus to show it.

8.3 Further Work

For a better data set, downloading more Twitter data is needed. A bigger data set would make a better understanding of these question that we may have. To get a bigger data set, downloading need to be done over a larger time scale. It is also possible to download data from a higher number of cities in the world.

As problem with downloading Twitter data arose, different ways of downloading these need to be looked at to get a bigger data set. Or other ways to evaluate

the location of the *tweets* could be done. A bigger data set would produce a more convincing result, and the result would thus be more believable and more legitimate.

It might be possible to model this pattern using another SIR model with different variables, or another model completely, so this could be something to work further with. A new model must be more advanced than the SIR model used here, such that we get a much better understanding on how it works. And a more complex model might be able to see the pattern which antigenic analysis have been able to see.

A

Cities and Their Time Series Used in This Thesis

Buenos Aires, Amsterdam, Beijing, Berlin, Brisbane, Cairo, Chicago, Copenhagen, Dallas, Delhi, Istanbul, Jakarta, Karachi, Los Angels, Lagos, London, Madrid, Manila, Mexico, Moscow, Nairobi, New York, Oslo, Paris, Rio de Janeiro, Roma, Santiago, São Paulo, Seoul, Singapore, Sydney, Tokyo and Toronto.

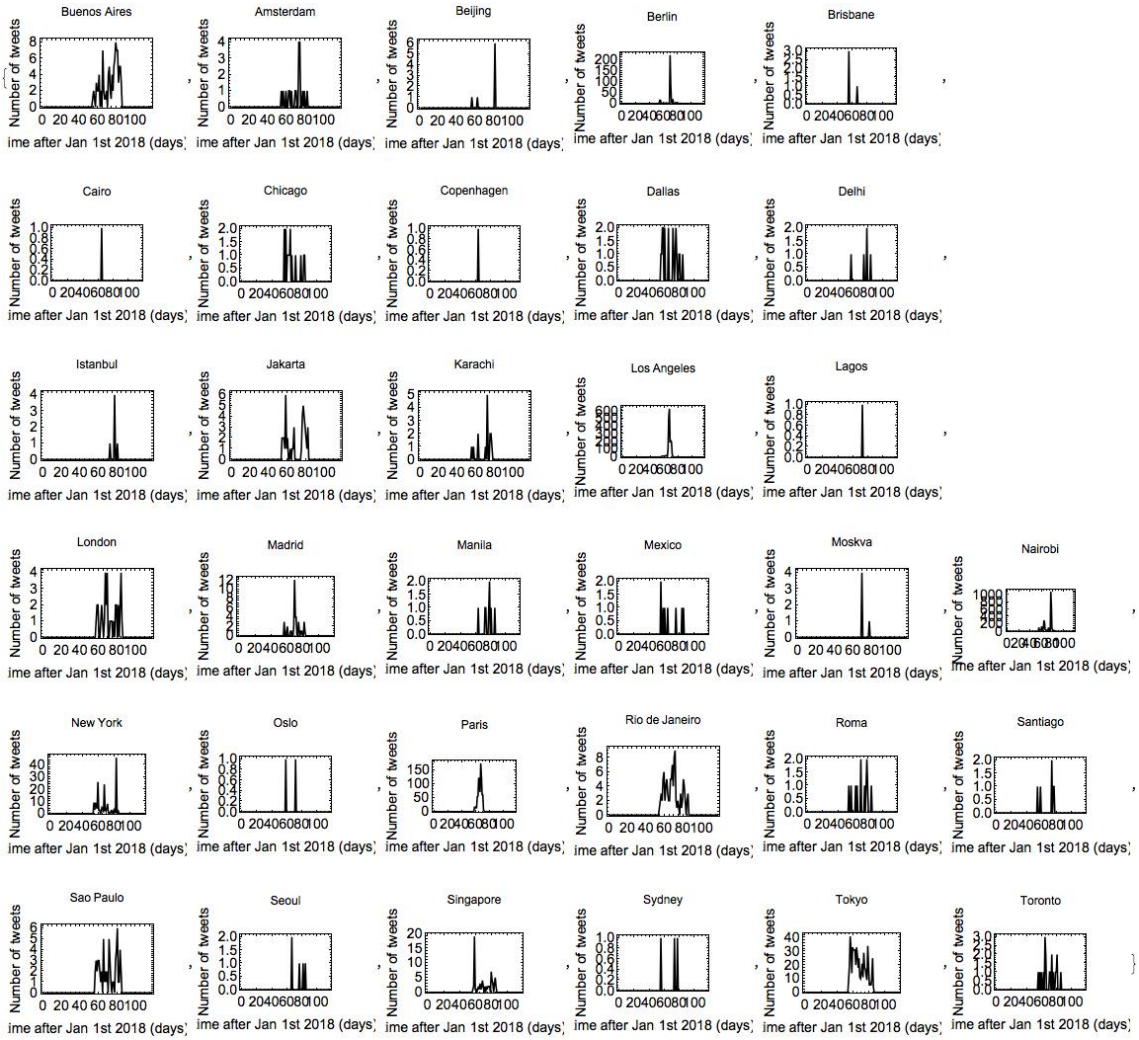


Figure A.1: The Time Series of all the cities.

B

The Mathematica codes used in this thesis

B.1 For the SIR model

```
  jumptimes = {};
Do[
  jumptimes =
    Append[jumptimes,
     Round[RandomReal[ExponentialDistribution[1/50.]]]];
  , {100}];
jumptimes = FoldList[Plus, 0, jumptimes];

rand = RandomChoice[{5, 10, 15, 20}, 100];

μ = 0.4;
β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5]);

δ = 0.06;
```

```
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
     {tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
   &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
```

```
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];

PL2=ListPlot[B[[All,3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle->Directive[16,Black,FontFamily -> Times],
Axes->False,Frame->True,
FrameLabel->{"Time (t)","S},
PlotStyle->Black,Epilog->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt = Partition[B[[All, 3]], 2][[All, 2]];
synt = synt[[1 ;; 1196]];
synt1 = synt;



μ = 0.4;
β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5] + +π/2]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50;
δ2 = δ;
```

```
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
    {tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle ->Black,Epilog->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
```

```
PL2=ListPlot[B[[All,3]],Joined->True,
AspectRatio->1/5,
ImageSize->800,
FrameStyle->Directive[16,␣Black,FontFamily->Times],
Axes->False,Frame->True,
FrameLabel␣->␣{"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt = Partition[B[[All, 3]], 2][[All, 2]];
synt = synt[[1 ;; 1196]];
synt2 = synt;


figb = ListPlot[{synt1, synt2},
PlotStyle -> {{Blue}, {Red}},
  Joined -> True, AspectRatio -> 1/4, Axes -> False,
Frame -> True,
  PlotStyle -> {Black, Thick},
  FrameStyle -> Directive[16,
  FontFamily -> "Helvetica", Black],
  FrameLabel -> {"time(weeks)",I(t)"}ImageSize␣->␣800,
Epilog->
Inset[Text[
Style["B",22,FontFamily␣->"Helvetica",Background->White]],
Scaled[{.06,.9}]],PlotRange->All]

Δj= Drop[jumptimes, 1] - Drop[jumptimes, -1];
jumps = Flatten[Table[Table[FoldList[Plus, 0, rand][[k]],
{Δ j[[k]]}], {k, 1, Length[rand]}]];
deter = Δ + μ *Range[Length[jumps]] + jumps;
deter = deter[[1 ;; Length[synt]]];

figa = ListPlot[{deter, μ*Range[1196]}, Joined -> True,
AspectRatio -> 1/4, Axes -> False, Frame -> True,
```

```
PlotStyle -> {{Black, Thick}, {Black, Dashed}},
FrameStyle -> Directive[16,
FontFamily -> "Helvetica",Black],
FrameLabel->{"time (weeks)","Cumulative recruitment"},
ImageSize->800,Epilog->Inset[Text[Style["A",22,
FontFamily->"Helvetica",Background->White]],
Scaled[{.06,.9}]]]

figc=ListPlot[synt1+synt2,PlotStyle->{Black},Joined->True,
AspectRatio->1/4,Axes->False,Frame->True,
PlotStyle->{Black,Thick},
FrameStyle->Directive[16,FontFamily->"Helvetica",Black],
FrameLabel->{"time (weeks)","I(t)"},ImageSize->800,
Epilog->Inset[Text[Style["C",22,FontFamily->"Helvetica",
Background->White]],Scaled[{.06,.9}]]]



FIGA=ListPlot[deter-μ*Range[1196],PlotStyle->{Black},
Joined->True,AspectRatio->1/4,Axes->False,Frame->True,
PlotStyle->{Black, Thick},
FrameStyle-> Directive[16,FontFamily->"Helvetica",
Black],FrameLabel->{"time (weeks)","viral evolution"},
ImageSize->800,Epilog->Inset[Text[Style["A",22,
FontFamily->"Helvetica",Background->White]],
Scaled[{.06,.9}]]]

z=(UnitStep[synt1+synt2-30])*(deter-μ*Range[1196]);
Z = Split[z];
Do[
        If[First[Z[[k]]] == 0,
                Z[[k]] = Z[[k]] + First[Z[[k - 1]]];
        ];
        , {k, 3, Length[Z]}];
Z = Flatten[Z];
```

```
ListPlot[Z]

FIGB = ListPlot[Z, PlotStyle -> {Black}, Joined -> True,
AspectRatio -> 1/4, Axes -> False, Frame -> True,
PlotStyle -> {Black, Thick}, FrameStyle ->
Directive[16, FontFamily -> "Helvetica",Black],
FrameLabel->{"time (weeks)","viral evolution"},
ImageSize->800,
Epilog->Inset[Text[Style["B",22,FontFamily->"Helvetica",
Background->
White]],Scaled[{.06,.9}]]]

a=Map[Length[#]&,Split[Z,#2-#1<10&]];
a=Flatten[Table[a,{20}]];
b=Drop[Map[First[#]&,Split[Z,#2-#1<10&]],1]-
Drop[Map[First[#]&,Split[Z,#2-#1<10&]],-1];
rand=Flatten[Table[b,{20}]];
jumptimes=FoldList[Plus,0,a]+3;


μ = 0.4;
β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50;
δ2 = δ;
δ3 = 0.;
```

```
Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
    {tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];

PL2 = ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
```

```
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle->Black, Epilog ->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All,3]],2][[All,2]];
synt=synt[[1;;1196]];
synt1=synt;
```

```
figbb =ListPlot[synt1, PlotStyle->{Darker[Green]},
Joined->True,
AspectRatio->1/4,Axes->False, Frame->True,
PlotStyle->Black, Thick},
FrameStyle->Directive[16,
FontFamily->"Helvetica",Black],
FrameLabel->{"time (weeks)","I(t)"},ImageSize->800,
PlotRange->All]

Show[figb, figbb, PlotRange->All]
Show[{figc, figbb},Epilog->Inset[""],PlotRange->All]
```

EXPLORING CHAOS IN THE MODEL:

```
 jumptimes = {};
Do[
  jumptimes =
    Append[jumptimes,
     Round[RandomReal[ExponentialDistribution[1/50.]]]];
  , {100}];
jumptimes = FoldList[Plus, 0, jumptimes];
```

```
rand = RandomChoice[{5, 10, 15, 20}, 100];

μ = 0.4;
β = 0.1 + 0.17*(1 − cos[2πt/(104) − π/5]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == −β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n − γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol], {
    tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
```

```
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel ->{"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
PL2 = ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel->{"Time (t)","S},
PlotStyle->Black,Epilog ->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All, 3]],2][[All,2]];
synt=synt[[1;;1196]];
synt1=synt;

β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5] + +π/2]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
```

```
δ1 = 50;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
    {tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black,
FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
```

```
Scaled[{.06,.9}]]];


PL2=ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle->Black,Epilog->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All,3]],2][[All,2]];
synt=synt[[1;;1196]];
synt2=synt;


Δj= Drop[jumptimes, 1] - Drop[jumptimes, -1];
jumps = Flatten[Table[Table[FoldList[Plus, 0, rand][[k]],
{Δ j[[k]]}], {k, 1, Length[rand]}]];
deter = Δ + μ *Range[Length[jumps]] + jumps;
deter = deter[[1 ;; Length[synt]]];


z=(UnitStep[synt1+synt2-30])*(deter-μ*Range[1196]);
Z = Split[z];
Do[
        If[First[Z[[k]]] == 0,
                Z[[k]] = Z[[k]] + First[Z[[k - 1]]];
        ];
        , {k, 3, Length[Z]}];
Z = Flatten[Z];
ListPlot[Z]


a=Map[Length[#] &,Split[Z,#2-#1<10&]];
```

```
a=Flatten[Table[a,{20}]];
b=Drop[Map[First[#] &,Split[Z,#2-#1<10&]],1]-
Drop[Map[First[#] &,Split[Z,#2-#1<10&]],-1];
rand=Flatten[Table[b, {20}]];
jumptimes=FoldList[Plus,0,a]+3;

μ = 0.4;
β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
    {tt, t1, t2}];
```

```
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];

PL2 = ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle->Black, Epilog ->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All,3]],2][[All,2]];
synt=synt[[1;;1196]];
synt1=synt;

F1 = ListPlot[synt1, PlotStyle->{Darker[Green]},
Joined->True,
AspectRatio->1/4,Axes->False,Frame->True,
PlotStyle->{Black,Thick},
```

```
Sperturb=10;
μ = 0.4;
β = 0.1 + 0.17*(1 − cos[2πt/(104) − π/5]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50 + Sperturb;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == −β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n − γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
    {tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
```

```
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel ->{"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
PL2 = ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel->{"Time (t)","S},
PlotStyle->Black,Epilog ->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All, 3]],2][[All,2]];
synt=synt[[1;;1196]];
synt1=synt;

β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5] + +π/2]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
```

```
δ1 = 50;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol], {
    tt, t1, t2}];
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black,
FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
```

```
Scaled[{.06,.9}]]];


PL2=ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle->Black,Epilog->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All,3]],2][[All,2]];
synt=synt[[1;;1196]];
synt2=synt;


Δj= Drop[jumptimes, 1] - Drop[jumptimes, -1];
jumps = Flatten[Table[Table[FoldList[Plus, 0, rand][[k]],
{Δ j[[k]]}], {k, 1, Length[rand]}]];
deter = Δ + μ *Range[Length[jumps]] + jumps;
deter = deter[[1 ;; Length[synt]]];


z=(UnitStep[synt1+synt2-30])*(deter-μ*Range[1196]);
Z = Split[z];
Do[
        If[First[Z[[k]]] == 0,
                Z[[k]] = Z[[k]] + First[Z[[k - 1]]];
        ];
        , {k, 3, Length[Z]}];
Z = Flatten[Z];
ListPlot[Z]


a=Map[Length[#] &,Split[Z,#2-#1<10&]];
```

```
a=Flatten[Table[a,{20}]];
b=Drop[Map[First[#] &,Split[Z,#2-#1<10&]],1]-
Drop[Map[First[#] &,Split[Z,#2-#1<10&]],-1];
rand=Flatten[Table[b, {20}]];
jumptimes=FoldList[Plus,0,a]+3;

μ = 0.4;
β = 0.1 + 0.17*(1 - cos[2πt/(104) - π/5]);

δ = 0.06;
m = 5;
σ = 5.;
γ = 0.10;
n = 200;
Clear[R, S, i]
L = {};
LL = {};
δ1 = 50;
δ2 = δ;
δ3 = 0.;

Do[
  t1 = jumptimes[[kk]];
  t2 = jumptimes[[kk + 1]];
  r0 = rand[[kk]];
  slutt = {δ1 + r0, δ2, δ3};
  eq = {S'[t] == -β S[t]*i[t]/n + μ,
    i'[t] == β S[t]*i[t]/n - γ i[t],
    R'[t] == γ i[t]};
  sol = NDSolve[{eq, S[t1] == slutt[[1]],
  i[t1] == slutt[[2]],
     R[t1] == slutt[[3]]}, {S, i, R}, {t, t1, t2}];
  L1 = Table[
    Flatten[{tt, S[tt], i[tt], R[tt]} /. sol],
    {tt, t1, t2}];
```

```
 δ1 = Last[L1][[2]];
  δ2 = Last[L1][[3]];
  δ3 = Last[L1][[4]];
  L = Join[L, L1[[All, 2 ;; 3]]];
  LL = Append[LL, L1];
  , {kk, 1, 50}];
  B = Partition[Flatten[Thread[{#1, #2}
  &[Range[Length[L]], L]]], 3];
PL1 = ListPlot[B[[All, 2]], Joined -> True,
AspectRatio -> 1/5,
ImageSize -> 800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle -> Black, Epilog ->
Inset[Text[Style["(C)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];

PL2 = ListPlot[B[[All, 3]],Joined->True,
AspectRatio->1/5,ImageSize->800,
FrameStyle -> Directive[16, Black, FontFamily -> Times],
Axes -> False, Frame -> True,
FrameLabel -> {"Time t","S},
PlotStyle->Black, Epilog ->
Inset[Text[Style["(D)",22,FontFamily->Times]],
Scaled[{.06,.9}]]];
synt=Partition[B[[All,3]],2][[All,2]];
synt=synt[[1;;1196]];
synt1=synt;

F2=ListPlot[synt1,PlotStyle->{Purple},Joined->True,
AspectRatio->1/4,Axes->False,Frame->True,
PlotStyle->{Black,Thick},
FrameStyle -> Directive[16,
FontFamily ->["Helvetica",Black],
```

```
FrameLabel->{"time (weeks)","I(t)"},ImageSize->800,
PlotRange->All]

Show[F1,F2,PlotRange->All]
```

B.2 Community Structures

B.2.1 Based on correlation

```
smooths =
  Table[N[MovingAverage[rekker[[i]], 7]],
  {i, 1, Length[rekker]}];

cortable =
  Table[Correlation[smooths[[i]], smooths[[j]]],
  {i, 1,
    Length[smooths]}, {j, 1, Length[smooths]}];
Do[
  cortable[[i, j]] = 0;,
  {i, 1, Length[smooths]}, {j, 1, i}];

par = Position[cortable, _?(# > 0.65 &)];
gr1 = Table[
   navn[[par[[k]][[1]]]] -> navn[[par[[k]][[2]]]],
   {k, 1, Length[par]}];
par = Position[cortable, _?(# > 0.75 &)];
gr2 = Table[
   navn[[par[[k]][[1]]]] -> navn[[par[[k]][[2]]]],
   {k, 1, Length[par]}];
par = Position[cortable, _?(# > 0.85 &)];
gr3 = Table[
   navn[[par[[k]][[1]]]] -> navn[[par[[k]][[2]]]],
   {k, 1,Length[par]}];
par = Position[cortable, _?(# > 0.95 &)];
gr4 = Table[
   navn[[par[[k]][[1]]]] -> navn[[par[[k]][[2]]]],
```

```
   {k, 1,  Length[par]}];
gr = Join[gr1, gr2, gr3, gr4];

GraphPlot[gr, VertexLabeling -> True]

figa = CommunityGraphPlot[gr, FindGraphCommunities[gr],
  PlotLegends -> {"Group_1", "Group_2", "Group_3",
  "Group_4"}, Epilog -> Inset[Style["\!\(\*
StyleBox[\"a\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400,
  CommunityBoundaryStyle -> {{Thick, Dashed},
  {Thick, s Dashed}, {Thick, Dashed}, {Thick, Dashed}}]

clusters = FindGraphCommunities[gr]

{{"London", "New_York", "Sao_Paulo", "Seoul",
"Singapore", "Tokyo", "Toronto", "Brisbane",
"Mexico", "Chicago"}, {"Rio_de_Janeiro",
  "Madrid", "Berlin", "Los_Angeles",
  "Moskva", "Paris", "Cairo",
  "Copenhagen", "Oslo"},
   {"Buenos_Aires", "Beijing", "Delhi",
  "Istanbul", "Jakarta", "Manila",
  "Nairobi", "Roma"}, {"Amsterdam",
  "Dallas", "Karachi", "Lagos", "Santiago",
  "Sydney"}}

r1 = Graphics[{Red, Disk[{0, 0}, 0.2]}];
r2 = Graphics[{Yellow, Disk[{0, 0}, 0.2]}];
r3 = Graphics[{Purple, Disk[{0, 0}, 0.2]}];
r4 = Graphics[{Orange, Disk[{0, 0}, 0.2]}];

figb = GeoGraphics[{GeoMarker[g1, r1, "Scale" -> 0.15],
   GeoMarker[g2, r2, "Scale" -> 0.15],
```

```
    GeoMarker[g3, r3, "Scale" -> 0.15],
    GeoMarker[g4, r4, "Scale" -> 0.15]}, GeoRange -> All,
  GeoBackground -> "Satellite",
  GeoProjection -> "Mollweide",
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"b\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 600, AspectRatio -> 2/3]

Grid[{{figa, figb}}]
```

B.2.2 Based on Time of Maximum

```
figa = CommunityGraphPlot[gr, FindGraphCommunities[gr],
  PlotLegends -> {"Group␣1", "Group␣2", "Group␣3",
  "Group␣4"},
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"a\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400,
  CommunityBoundaryStyle -> {{Thick, Dashed},
  {Thick,  Dashed}, {Thick, Dashed}, {Thick, Dashed}}]

clusters = FindGraphCommunities[gr]

{{"London", "New␣York", "Sao␣Paulo", "Seoul", "Singapore",
 "Tokyo", "Toronto", "Brisbane", "Mexico", "Chicago"},
 {"Rio␣de␣Janeiro", "Madrid", "Berlin", "Los␣Angeles",
  "Moskva", "Paris", "Cairo",  "Copenhagen", "Oslo"},
  {"Buenos␣Aires", "Beijing", "Delhi",  "Istanbul",
  "Jakarta", "Manila", "Nairobi", "Roma"},
  {"Amsterdam", "Dallas", "Karachi", "Lagos",
  "Santiago", "Sydney"}}

r1 = Graphics[{Red, Disk[{0, 0}, 0.2]}];
r2 = Graphics[{Yellow, Disk[{0, 0}, 0.2]}];
```

```
r3 = Graphics[{Purple, Disk[{0, 0}, 0.2]}];
r4 = Graphics[{Orange, Disk[{0, 0}, 0.2]}];

figb = GeoGraphics[{GeoMarker[g1, r1, "Scale" -> 0.15],
   GeoMarker[g2, r2, "Scale" -> 0.15],
   GeoMarker[g3, r3, "Scale" -> 0.15],
   GeoMarker[g4, r4, "Scale" -> 0.15]},
   GeoRange -> All,
  GeoBackground -> "Satellite",
  GeoProjection -> "Mollweide",
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"b\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 600, AspectRatio -> 2/3]

Grid[{{figa, figb}}]

maks = Table[Last[Ordering[smooths[[i]]]],
{i, 1, Length[smooths]}];
matrise =
  Table[Abs[maks[[i]] - maks[[j]]], {i, 1, Length[maks]},
  {j, 1,
    Length[maks]}];
Do[
  Do[
    matrise[[i, j]] = 999
    , {i, 1, j}];
  ,
  {j, 1, Length[matrise]}];
par = Join[Position[matrise, _?(# == 0 &)],
   Position[matrise, _?(# <= 1 &)],
   Position[matrise, _?(# <= 2 &)],
   Position[matrise, _?(# <= 3 &)],
   Position[matrise, _?(# <= 4 &)],
   Position[matrise, _?(# <= 5 &)]];
```

```
gr = Table[
   navn[[par[[k]][[1]]]] -> navn[[par[[k]][[2]]]],
   {k, 1, Length[par]}];
GraphPlot[gr, VertexLabeling -> True]

CommunityGraphPlot[gr, FindGraphCommunities[gr],
ImageSize -> 400]

figa = CommunityGraphPlot[gr, FindGraphCommunities[gr],
  PlotLegends -> {"Group_1", "Group_2", "Group_3",
  "Group_4"},
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"a\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400,
  CommunityBoundaryStyle -> {{Thick, Dashed},
  {Thick, Dashed}, {Thick, Dashed}, {Thick, Dashed}}]

clusters = FindGraphCommunities[gr]

{{"Lagos", "Karachi", "Manila", "Moskva", "Berlin",
"Nairobi",  "Santiago", "Amsterdam", "Sydney",
"Madrid", "Oslo"}, {"Dallas",  "Brisbane", "Mexico",
"New_York", "Singapore", "Chicago", "Tokyo",
"Toronto"}, {"Delhi", "Buenos_Aires", "Sao_Paulo",
"Seoul", "Beijing", "Istanbul", "Jakarta", "Roma"},
{"London", "Copenhagen", "Paris", "Cairo",
"Los_Angeles", "Rio_de_Janeiro"}}

r1 = Graphics[{Purple, Disk[{0, 0}, 0.2]}];
r2 = Graphics[{Red, Disk[{0, 0}, 0.2]}];
r3 = Graphics[{Orange, Disk[{0, 0}, 0.2]}];
r4 = Graphics[{Yellow, Disk[{0, 0}, 0.2]}];
```

```
g2 = Map[Interpreter["City"][#] &, clusters[[2]]]
g3 = Map[Interpreter["City"][#] &, clusters[[3]]]
g4 = Map[Interpreter["City"][#] &, clusters[[4]]]

figb = GeoGraphics[{GeoMarker[g1, r1, "Scale" -> 0.15],
   GeoMarker[g2, r2, "Scale" -> 0.15],
   GeoMarker[g3, r3, "Scale" -> 0.15],
   GeoMarker[g4, r4, "Scale" -> 0.15]}, GeoRange -> All,
  GeoBackground -> "Satellite",
  GeoProjection -> "Mollweide",
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"b\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400, AspectRatio -> 2/3]
```

B.2.3 Real Influenza Data

```
worldonsets = {};
Do[
  land = liste[[i]];
  S = Extract[s,
    Position[
     Map[StringSplit[#, ","] &, s][[All, 1]],
     _?(# == land &)]];
  S = Map[StringSplit[#, ","] &, S];
  flu = S[[Flatten[Position[S[[All, 2]],
  _?(# == "flu" &)]]]];
  f = flu[[All, 6]];
  dates = flu[[All, 3 ;; 5]];
  n = Length[f];
  F = Thread[{#1, #2} &[Range[n], f]];
  pos = Position[f, _?(# == "NaN" &)];
  F = ToExpression[Delete[F, pos]];
  dates = Delete[dates, pos];
  times = F[[All, 1]];
  \[CapitalDelta] = Drop[times, 1] - Drop[times, -1];
```

```
pos = Position[\[CapitalDelta], _?(# == 1 &)];
a = F[[All, 2]][[Flatten[pos]]];
b = F[[All, 2]][[Flatten[pos] + 1]];
pairs = Thread[{#1, #2} &[a, b]];
times = Extract[times, pos];
win = 20;
L = {};
Do[
 ppos = Position[times, _?(t - win < # < t &)];
 If[Length[ppos] > 0,
  par = Extract[pairs, ppos];
  \[Lambda] = Fit[par, {zz}, zz]/zz;
  L = Append[L, {t, \[Lambda]}];
  ];
 , {t, 1, Last[times]}];
segs = Split[L[[All, 2]] - 1, #1*#2 > 0 &];
sign = Map[Mean[#] &, Map[Sign[#] &, segs]];
hv = {};
Do[
 If[sign[[i]] == 1 && sign[[i - 1]] != 1 &&
    Length[segs[[i]]] >= 3, (* condtions for onsets*)

   hv = Append[hv, i];
   ];
 , {i, 2, Length[sign]}];
grids0 =
 Table[Length[Flatten[segs[[1 ;; hv[[k]] - 1]]]],
 {k, 1, Length[hv]}];
grids = L[[All, 1]][[grids0]];
i = 1;
poss = {};
While[i <= Length[grids],
 mid = grids[[i]];
 pos = Position[grids, _?(mid < # < mid + 12 &)];
 poss = Join[poss, pos];
```

```
    If[Length[pos] > 0,
     i = Last[pos][[1]];
     ,
     i = i + 1;
     ];
    ];
   grids = Delete[grids, poss];
   QL1 = ListPlot[F, PlotRange -> All, Joined -> True,
     PlotRange -> All, AspectRatio -> 1/5,
     ImageSize -> 800,
     Axes -> False, Frame -> True,
     FrameStyle -> Directive[16,
     FontFamily -> Times, Black],
     PlotStyle -> Black,
     FrameLabel -> {"time (weeks)", "Incidence"},
     GridLines -> {grids, None}, GridLinesStyle -> Blue];
   onsets = grids;
   onsetdates =
    ToExpression[
     dates[[Flatten[
       Table[Position[F[[All, 1]], _?(# == onsets[[k]] &)],
       {k, 1,  Length[onsets]}]]]];
   worldonsets = Append[worldonsets, {land, onsetdates}];
   , {i, 1, Length[liste]}];

func = DateDifference[{1995, 1, 1}, #][[1]] &;

worldonsets2 =
   Table[{worldonsets[[i]][[1]],
   Map[func, worldonsets[[i]][[2]]]},
   {i,1, Length[worldonsets]}];

Z = Union[Flatten[worldonsets2[[All, 2]]]];
landZ = Table[
    liste[[Position[worldonsets2,
```

```
   _?(# == Z[[k]] &)][[1]][[1]]],
   {k, 1, Length[Z]}];

ZZ = Table[{Z[[k]],
    Extract[Z, Position[Z,
    _?(Z[[k]] < # < Z[[k]] + 28 &)]]},
    {k, 1, Length[Z]}];

piler = Flatten[
   DeleteCases[
    Table[Table[
      liste[[Position[worldonsets2,
      _?(# == ZZ[[k]][[1]] &)][[1]][[
         1]]]] ->
       liste[[Position[worldonsets2,
       _?(# == ZZ[[k]][[2]][[j]] &)][[
          1]][[1]]]], {j, 1, Length[ZZ[[k]][[2]]]}],
        {k, 1,  Length[ZZ]}], _?(# == {} &)]];

piler2 = Extract[Normal[Counts[piler]],
  Position[Normal[Counts[piler]][[All, 2]],
  _?(# >= 3 &)]]

{("United_Kingdom" -> "France") -> 3,
("France" -> "Finland") ->
  3, ("Japan" -> "France") -> 3,
  ("Japan" -> "Spain") ->
  3, ("France" -> "Belgium") -> 4,
  ("France" -> "Canada") -> 3,
  ("Belgium" -> "Canada") -> 3,
  ("South_Africa" -> "Argentina") -> 3,
  ("Japan" -> "Israel") -> 3,
  ("Japan" -> "Canada") -> 3,
  ("Canada" -> "Japan") -> 3,
  ("Argentina" -> "New_Zealand") -> 6,
```

```
  ("New_Zealand" -> "Chile") -> 3, (
  "Canada" -> "Belgium") -> 3,
  ("China" -> "Italy") -> 3,
  ("France" -> "Germany") -> 4,
  ("New_Zealand" -> "Australia") -> 3,
  ("Chile" -> "Australia") -> 3,
  ("Canada" -> "Egypt") -> 3,
  ("Belgium" -> "Croatia") -> 3}

GraphPlot[piler2[[All, 1]], VertexLabeling -> True,
 DirectedEdges -> True, ImageSize -> 800]

GraphPlot[piler]

figa = CommunityGraphPlot[piler,
FindGraphCommunities[piler],
  PlotLegends -> {"Group_1", "Group_2", "Group_3"},
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"a\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400,
  CommunityBoundaryStyle -> {{Thick, Dashed}, {Thick,
     Dashed}, {Thick, Dashed}}]

clusters = FindGraphCommunities[piler]

gruppe1 = Map[Interpreter["Country"][#] &,
clusters[[1]]]

gruppe2 = Map[Interpreter["Country"][#] &,
clusters[[2]]]

gruppe3 = Map[Interpreter["Country"][#] &,
clusters[[3]]]
```

```
figb = GeoGraphics[{
    GeoStyling[Opacity[.7]], EdgeForm[Black],
    Red, Polygon /@ gruppe1,
    GeoStyling[Opacity[.7]], EdgeForm[Black],
    Yellow, Polygon /@ gruppe2,
    GeoStyling[Opacity[.7]], EdgeForm[Black],
    Purple, Polygon /@ gruppe3
    }, GeoRange -> All, GeoBackground -> "Satellite",
  GeoProjection -> "Mollweide",
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"b\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 600, AspectRatio -> 2/3]

Grid[{{figa, figb}}]
```

B.2.4 Flight Traffic

```
gr = Flatten[
    Table[Table[par[[i, j]], {B[[i, j]]}],
    {i, 1, Length[A]}, {j, 1,
      Length[A]}]];

GraphPlot[gr, ImageSize -> 800, VertexLabeling -> True]

figa = CommunityGraphPlot[gr, FindGraphCommunities[gr],
  PlotLegends -> {"Group_1", "Group_2", "Group_3",
  "Group_4"},
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"a\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400,
  CommunityBoundaryStyle -> {{Thick, Dashed}, {Thick,
     Dashed}, {Thick, Dashed}, {Thick, Dashed}},
  Method -> "Hierarchical"]
```

```
clusters = FindGraphCommunities[gr]

r1 = Graphics[{Purple, Disk[{0, 0}, 0.2]}];
r2 = Graphics[{Red, Disk[{0, 0}, 0.2]}];
r3 = Graphics[{Orange, Disk[{0, 0}, 0.2]}];
r4 = Graphics[{Yellow, Disk[{0, 0}, 0.2]}];

figb = GeoGraphics[{GeoMarker[g1, r1, "Scale" -> 0.15],
   GeoMarker[g2, r2, "Scale" -> 0.15],
   GeoMarker[g3, r3, "Scale" -> 0.15],
   GeoMarker[g4, r4, "Scale" -> 0.15]}, GeoRange -> All,
  GeoBackground -> "Satellite",
  GeoProjection -> "Mollweide",
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"b\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 600, AspectRatio -> 2/3]

Grid[{{figa, figb}}]

R = Table[
   N[Mean[Extract[X,
      Position[X[[All, 1]], _?(i - 5 <= # < i + 5 &)]]]],
      {i, 5, 60,
    5}];
```

This is the code for the fitted linear model:

```
gr = Flatten[
   Table[Table[par[[i, j]], {B[[i, j]]}], {i, 1, Length[A]},
    {j, 1, Length[A]}]];

GraphPlot[gr, ImageSize -> 800, VertexLabeling -> True]

figa = CommunityGraphPlot[gr, FindGraphCommunities[gr],
  PlotLegends -> {"Group_1", "Group_2", "Group_3",
```

```
  "Group_4"},
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"a\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 400,
  CommunityBoundaryStyle -> {{Thick, Dashed}, {Thick,
     Dashed}, {Thick, Dashed}, {Thick, Dashed}},
  Method -> "Hierarchical"]

clusters = FindGraphCommunities[gr]

r1 = Graphics[{Purple, Disk[{0, 0}, 0.2]}];
r2 = Graphics[{Red, Disk[{0, 0}, 0.2]}];
r3 = Graphics[{Orange, Disk[{0, 0}, 0.2]}];
r4 = Graphics[{Yellow, Disk[{0, 0}, 0.2]}];

figb = GeoGraphics[{GeoMarker[g1, r1, "Scale" -> 0.15],
   GeoMarker[g2, r2, "Scale" -> 0.15],
   GeoMarker[g3, r3, "Scale" -> 0.15],
   GeoMarker[g4, r4, "Scale" -> 0.15]}, GeoRange -> All,
  GeoBackground -> "Satellite",
  GeoProjection -> "Mollweide",
  Epilog -> Inset[Style["\!\(\*
StyleBox[\"b\",\nFontWeight->\"Bold\"]\)", 16],
Scaled[{0.1, 0.9}]],
  ImageSize -> 600, AspectRatio -> 2/3]

Grid[{{figa, figb}}]

X = Partition[
   Flatten[Table[
     Table[{AA[[i, j]], cortable[[j, i]]}, {i, j + 1, 33}],
     {j, 1, 32}]], 2];

lm = LinearModelFit[X, x, x]
```

```
lm["ParameterTable"]

gg = Fit[X, {zz, 1}, zz];
PL1 = Plot[gg, {zz, 0, 60}, PlotStyle -> {Red, Thick}];
PL2 = ListPlot[X, PlotRange -> All, Axes -> False,
Frame -> True,
   AspectRatio -> 1, PlotStyle -> {Black,
   PointSize[0.002]}];
R = Table[
   N[Mean[Extract[X,
      Position[X[[All, 1]], _?(i - 5 <= # < i + 5 &)]]]],
      {i, 5, 60,
    5}];
PL3 = ListPlot[R, PlotRange -> All, Axes -> False,
Frame -> True,
   AspectRatio -> 1, PlotStyle -> {PointSize[0.02], Red}];
Show[{PL1, PL2, PL3}, PlotRange -> All, Axes -> False,
Frame -> True,
 AspectRatio -> 1, FrameStyle -> Directive[Black, 14],
 FrameLabel -> {"Number␣of␣direct␣flights␣between␣cities",
   "Correlation␣between␣(influeza)␣Twitter␣signals"},
 Epilog -> Inset[Style["\!\(\*
StyleBox[\"p\",\n\
FontSlant->\"Italic\"]\)<\!\(\*SuperscriptBox[\(10\),
\(-4\)]\)",
    14], Scaled[{0.5, 0.72}]]]

R = Table[
   N[Mean[Extract[X,
      Position[X[[All, 1]], _?(i - 5 <= # < i + 5 &)]]]],
      {i, 5, 60, 5}];
```

B.2.5 The Community Structure Examples

```
WolframLanguageData["Cos", "RelationshipCommunityGraph"]

col = RandomColor[20];

ClusteringTree[col]
```

B.3 For Downloading Twitter Data

This is an example for only one city, but it is done precisely the same way, only that where it says London is changed with the other cities name. After the query, other words where used as well.

```
london = twitter["TweetSearch", "Query" -> "flu" ,
   "GeoPosition" -> Interpreter["City"]["London"],
   MaxItems -> 1000];

filnavnlondon =
  StringJoin["/Users/inga/Dropbox/twitter/",
   StringJoin[
    StringJoin[Drop[Drop[StringSplit[DateString[
    Date[]]], 1], -1]],
    "_london.M"]];
Export[filnavnlondon, london];
```

/C

The Flight Matrix

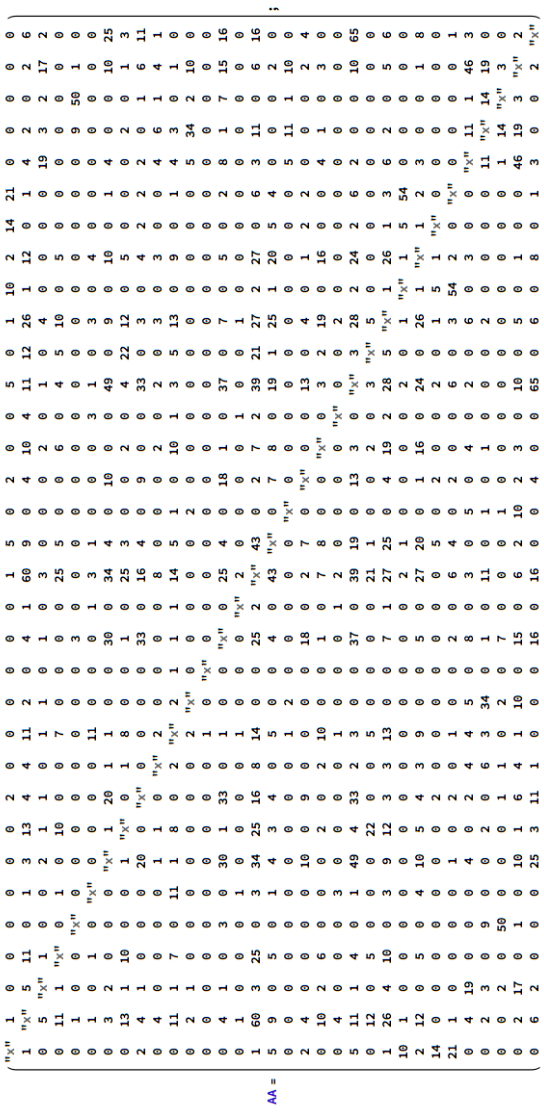


Figure C.1: The Flight Matrix.

Bibliography

[Achrekar et al., 2011] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE.

[Agogo and Hess, 2018] Agogo, D. and Hess, T. J. (2018). Scale development using twitter data: applying contemporary natural language processing methods in is research. In *Analytics and Data Science*, pages 163–178. Springer.

[Azziz Baumgartner et al., 2012] Azziz Baumgartner, E., Dao, C. N., Nasreen, S., Bhuiyan, M. U., Mah-E-Muneer, S., Mamun, A. A., Sharker, M. Y., Zaman, R. U., Cheng, P.-Y., Klimov, A. I., et al. (2012). Seasonality, timing, and climate drivers of influenza activity worldwide. *The Journal of infectious diseases*, 206(6):838–846.

[Bedford et al., 2010] Bedford, T., Cobey, S., Beerli, P., and Pascual, M. (2010). Global migration dynamics underlie evolution and persistence of human influenza a (h3n2). *PLoS pathogens*, 6(5):e1000918.

[Bedford et al., 2015] Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217.

[Broniatowski et al., 2013] Broniatowski, D. A., Paul, M. J., and Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672.

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

[Cooper et al., 2003] Cooper, N. J., Sutton, A. J., Abrams, K. R., Wailoo, A., Turner, D., and Nicholson, K. G. (2003). Effectiveness of neuraminidase inhibitors in treatment and prevention of influenza a and b: systematic review and meta-analyses of randomised controlled trials. *Bmj*, 326(7401):1235.

[Corley et al., 2010] Corley, C. D., Cook, D. J., Mikler, A. R., and Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, 7(2):596–615.

[Dawar et al., 2018] Dawar, A., Purwar, A., Anand, N., and Singla, C. (2018). Tweetrush: A tool for analysis of twitter data.

[Doshi, 2008] Doshi, P. (2008). Trends in recorded influenza mortality: United states, 1900–2004. *American journal of public health*, 98(5):939–945.

[Flahault et al., 2006] Flahault, A., Vergu, E., Coudeville, L., and Grais, R. F. (2006). Strategies for containing a global influenza pandemic. *Vaccine*, 24(44-46):6751–6755.

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

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

[Glendinning and Perry, 1997] Glendinning, P. and Perry, L. P. (1997). Melnikov analysis of chaos in a simple epidemiological model. *Journal of mathematical biology*, 35(3):359–373.

[Hayes, 2008] Hayes, C. E. (2008). Prevention of influenza. *Journal of Midwifery & Women's Health*, 53(3):268–271.

[Hussain et al., 2018] Hussain, H., McGeer, A., McNeil, S., Katz, K., Loeb, M., Simor, A., Powis, J., Langley, J., Muller, M., and Coleman, B. (2018). Factors associated with influenza vaccination among health care workers in acute care hospitals in canada. *Influenza and other respiratory viruses*.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM.

[Kilbourne, 2006] Kilbourne, E. D. (2006). Influenza pandemics of the 20th century. *Emerging infectious diseases*, 12(1):9.

[Kumar et al., 2014] Kumar, S., Morstatter, F., and Liu, H. (2014). *Twitter data analytics*. Springer.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

[Lampos and Cristianini, 2010] Lampos, V. and Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE.

[Leder and Newman, 2005] Leder, K. and Newman, D. (2005). Respiratory infections during air travel. *Internal medicine journal*, 35(1):50–55.

[Leitmeyer and Adlhoch, 2016] Leitmeyer, K. and Adlhoch, C. (2016). Influenza transmission on aircraft: a systematic literature review. *Epidemiology (Cambridge, Mass.)*, 27(5):743.

[Lowen et al., 2007] Lowen, A. C., Mubareka, S., Steel, J., and Palese, P. (2007). Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151.

[Masse, 2011] Masse, M. (2011). *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. " O'Reilly Media, Inc.".

[Meehan et al., 1997] Meehan, T. P., Fine, M. J., Krumholz, H. M., Scinto, J. D., Galusha, D. H., Mockalis, J. T., Weber, G. F., Petrillo, M. K., Houck, P. M., and Fine, J. M. (1997). Quality of care, process, and outcomes in elderly patients with pneumonia. *Jama*, 278(23):2080–2084.

[Newman and Girvan, 2004] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

[Paul et al., 2014] Paul, M. J., Dredze, M., and Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLoS currents*, 6.

[Paules et al., 2018] Paules, C. I., Sullivan, S. G., Subbarao, K., and Fauci, A. S. (2018). Chasing seasonal influenza—the need for a universal influenza vaccine. *New England Journal of Medicine*, 378(1):7–9.

[Polgreen et al., 2008] Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448.

[Rambaut et al., 2008] Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615.

[Rothberg et al., 2008] Rothberg, M. B., Haessler, S. D., and Brown, R. B. (2008). Complications of viral influenza. *The American journal of medicine*, 121(4):258–264.

[Russell et al., 2008a] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., et al. (2008a). The global circulation of seasonal influenza a (h3n2) viruses. *Science*, 320(5874):340–346.

[Russell et al., 2008b] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., et al. (2008b). Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26:D31–D34.

[Rvachev and Longini Jr, 1985] Rvachev, L. A. and Longini Jr, I. M. (1985). A mathematical model for the global spread of influenza. *Mathematical biosciences*, 75(1):3–22.

[Signorini et al., 2011a] Signorini, A., Segre, A. M., and Polgreen, P. M. (2011a). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

[Signorini et al., 2011b] Signorini, A., Segre, A. M., and Polgreen, P. M. (2011b). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

[Sinnenberg et al., 2017] Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., and Merchant, R. M. (2017). Twitter as a tool for health research: a systematic review. *American journal of public health*, 107(1):e1–e8.

[Statista, 2018] Statista (2018). Leading countries based on number of twitter users as of april 2018. *URL: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/*.

[Teitzel, 2018] Teitzel, G. (2018). The moving target of flu. *Cell*, 172(6):1139–1141.

[Tellier, 2006] Tellier, R. (2006). Review of aerosol transmission of influenza a virus. *Emerging infectious diseases*, 12(11):1657.

[Tellier, 2009] Tellier, R. (2009). Aerosol transmission of influenza a virus: a review of new studies. *Journal of the Royal Society Interface*, page rsif20090302.

[Viboud et al., 2006a] Viboud, C., Alonso, W. J., and Simonsen, L. (2006a). Influenza in tropical regions. *PLoS medicine*, 3(4):e89.

[Viboud et al., 2006b] Viboud, C., Miller, M. A., Grenfell, B. T., Bjørnstad, O. N., and Simonsen, L. (2006b). Air travel and the spread of influenza: important caveats. *PLoS Medicine*, 3(11):e503.

[Wagner et al., 2009] Wagner, B. G., Coburn, B. J., and Blower, S. (2009). Calculating the potential for within-flight transmission of influenza a (h1n1). *BMC medicine*, 7(1):81.

[Webster, 1997] Webster, R. (1997). Influenza virus: transmission between species and relevance to emergence of the next human pandemic. In *Viral Zoonoses and Food of Animal Origin*, pages 105–113. Springer.

[Weinstein et al., 2003] Weinstein, R. A., Bridges, C. B., Kuehnert, M. J., and Hall, C. B. (2003). Transmission of influenza: implications for control in health care settings. *Clinical infectious diseases*, 37(8):1094–1101.

[WHO, 2014] WHO (2014). Fact sheet no. 211, march 2014. *URL: http://www. who. int/mediacentre/factsheets/fs211/en*.

[York, 2018] York, A. (2018). Viral infection: Breathing alone may spread the flu. *Nature Reviews Microbiology*, 16(3):123.