Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation:	Spring / Autumn semester, 20..... Open/Confidential
Author:	…… (signature of author)
Programme coordinator: Supervisor(s):	
Title of master's thesis:	
Credits:	
Keywords:	Number of pages: …… + supplemental material/other: …… Stavanger,……. date/year

Automated collection of multi-source spatial information for emergency management

Tracking the influenza seasons

Sandra Moen

A thesis presented for the degree of
Master of Science in Computer Science

University of Stavanger

Department of Electrical Engineering and
Computer Science
University of Stavanger
Norway
Spring 2018

Automated collection of multi-source spatial information for emergency management

Tracking the influenza seasons

Sandra Moen

Abstract

Influenza epidemics costs both lives and a tremendous amount of resources for any country. Citizens that become sick are less productive and the overall quality of life is drastically reduced for the amount of the individuals period of illness as well as the community during a flu season. The ability to reduce the spread of infectious diseases saves both lives and resources as well as an improvement of the quality of life.

This project aims to explore the possibilities to detect influenza outbreaks as soon as they are happening with the use of relevant datasets available. Information about different aspects of a citizens life on a grand scale reveals patterns and trends that could be linked to an epidemic outbreak, and thus prove useful for active measurements against further spread on a early début.

The results show ...

Possible solutions to ...

Acknowledgements

This thesis is considered an impressive achievement for the author, it was completed in spite of hardships endured. Under no circumstance should this thesis be considered a Norwegian accomplishment, for the oppression suffered they are deemed unworthy.

This thesis was written for the Department of Electrical Engineering and Computer Science at the University of Stavanger. Creating a means to solve problems that limit peoples lives have always been a real motivator. Predicting the flu season and hindering it in early stages would save an enormous amount of resources and improve life quality, this would be very rewarding. A special thanks to the supervisor for this project from the University of Stavanger Professor Erlend Tøssebro for his enthusiastic guidance and involvement, and the initiator who inspired incentive to the creation of this project as well as his continuous helpful guidance and involvement Phd fellow Lars Ole Grottenberg.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The power to obtain enough information to detect possible trends of influenza seasons depends on successful integration between a multitude of different participants. Automatic extraction and processing of data is paramount for efficient analysis and gives a solid basis for an autonomous pathological detection system. Scalability is important in merging new relevant datasets as they become available in an ever-growing societal infrastructure. This proposed technology would become an influential part of a bigger foundation intertwined with a robust knowledgeable and organizational means to mobilize assets in order to respond to possible outbreaks as or even before they start.

Influenza is an exceedingly contagious viral infection which gives high fever, general pain, and respiratory symptoms. An estimated five to ten percent of the population becomes infected during a yearly winter season.
The virus is especially dangerous to the elderly and to pregnant people from the second-trimester [1].

1.2 Objectives

This paper describes a plausible examination of the viability of monitoring, collecting and analyzing obtainable relevant data for a self-sufficient influenza seasonal recognition system. The management of seasonal influenza outbreaks is handled by public health officials and epidemiologists with the use of the national surveillance system provided by the Norwegian Institute of Public Health (NIPH)[2]. The Norwegian Syndromic Surveillance System (NorSySS) collects influenza-like illnesses (ILI) from general practitioners (GPs)[3]. These provide the means to monitor current influenza seasons with delay and as a basis to survey urban real-time datasets. The main thesis of this project is as influenza develops this reveals subtle patterns in societal behaviour that is detectable through a variety of mediums, e.g urban datasets from sewage, public transportation, medicinal purchases, recreational habits, social media and other such sources of public information.

1.3 Outline

The thesis is structured into seven chapters.

Chapter 2 describes related works

Chapter 3 mark out in detail the datasets used by this project, describes and give explanation to relevance, challenges, limitation and rewards.

Chapter 4 outlines the implementation and graphical results of the datasets used in chapter 3.

Chapter 5 shows the results.

Chapter 6 discusses the results.

Chapter 7 concludes the thesis, discusses constraints and possible future work as well as other suggestions.

Chapter 2

Related Works

Several research studies have been conducted on the practices that this thesis involves. In this chapter related works will be acknowledged.

2.1 Spatiotemporal information from urban systems

Grottenberg writes ... [4]

2.2 Twitter

A number of studies have been created on the information users on Twitter generate in providing valuable insights into the population by analysing millions of twitter messages (tweets). Researchers have studied tweets to reveal political opinions[5], measure public health[6], linguistic sentiments[7] and even environmental phenomena such as earth quakes[8]. Achrekar et al.[6] examines tweet flu trends and compares them with actual influenza data. The results show a high correlation between self-reported instances of flu-like illnesses (ILI) and reported ILI by public health providers. Achrekar references claims that early prevention limits the spread of infectious diseases and that twitter data is an 'untapped data source' that actually is quite reliable. This demonstrates how social media can be used to predict real-world consequences, and gives credibility to usage in this thesis.
Michal J. Paul and Mark Dredze [9] also conducted research on the usage of twitter data to measure population characteristics. In their conclusion twitter data from many users divulges reliable information about a certain topic of interest and in particular public health. They further discuss the pros and cons namely that self-reported is low cost and rapid transmission, whereas on the other side this is a 'blind authorship, lack of source citation and presentation of opinion as fact'. Certainly twitter messages may be false on an individual level, but however when taking into account thousands or even millions of messages this seems not plausible on a bigger scale.

Chapter 3

Experimental

In this chapter the different datasets used will be introduced. The goal of this project is to use as many datasets possible and then later evaluate them according to relevant results.

3.1 The Norwegian Institute of Public Health

The Norwegian Institute of Public Health (NIPH) have weekly updates[10] on the development on the current influenza season as well as previous ones. The reports include numbers of diagnoses from general practitioners (GPs) considering influenza-like illness (ILI), and hospitalized virus observations with graphs of both. No numbers are appended to the ILI but upon further request this was provided. Exact numbers are only included for the three last years, therefore the project only uses the seasons of the years 2015/2016, 2016/2017 and 2017/2018. The reports covers how many Norwegians seek treatment for ILI and what kind of influenza viruses are circulating in the country and where, vaccine status and recommendations, as well as the overall prognosis of this season. GPs report ILI based on these characteristics: muscle pain, coughing, fever and the feeling of being sick.

3.2 The Norwegian Public Roads Administration

The Norwegian Public Roads Administration (NPRA) have several different collections of data available for a number of different purposes. The motivation of this project requires traffic data of how many cars pass a certain registration station at a given time at a given position, the hypothesis for this that when people are ill they commute less and thus this shows on statistical data. Freely on their website [11] there are a few interesting options. They have traffic information in the standard traffic management exchange data structure (DATEX) application programming interface (API), statistics in an extensible markup language (XLM) and traffic index data relevant to the years before. It is important that the data collected is on a weekly basis at least in order to compare it to the influenza data. The data on their website does not suffice for this purpose, traffic data is only registered on a yearly and monthly basis. Luckily upon further investigation and help from the NPRA better data was granted upon request, hidden from that available on their website. The data given contained a set of traffic registration stations throughout Norway.

With this statistics of the daily traffic amount and spatial bounds can be derived showing the possible correlation influenza can have on traffic. The regions of interest are the whole of Norway and the three cities of Stavanger, Bergen and Oslo.

3.3 Twitter

The reason twitter data is interesting is that it contains self-reported instances of influenza on an individual level. These self-reported cases may even occur without the patient visiting a doctor, and so capture otherwise non-reported instances of ILI. The advantages are an instant notification about possible ILI and its spread, against the disadvantages of it being self-reported and thus somewhat unreliable. Twitter has several APIs available for public use, the one used in this project is the REST or search API which allows for searching against a set of keywords. The representational state transfer (REST) API is limited though, data accessible is roughly only maximum 10 days old and the search limit is on a maximum of one hundred messages called 'tweets'. The other API of interest is the stream API which continually gets the latest tweets. In order to only get Norwegian tweets, a set of geographical locations needs to be defined. The reason the stream API was not used is firstly that it requires a computer running on the internet continuously in order to get all the desired tweets. Secondly, the data collected could become large slowing down other post-processing algorithms and taking up unnecessary storage. Lastly, the stream API only provides a small set of the actual tweets tweeted, this means when searching for a specific term using the stream API some relevant tweets could go unnoticed and thus a search API is more appropriate for this task.

3.4 Kolumbus

Kolumbus is the public transportation administration in the state of Rogaland in Norway, this includes Stavanger, a city of interest. Unfortunately, Kolumbus provides no API, but on further request data of monthly passenger travel was provided from the years of 2015-2017.

3.5 Ruter

Ruter is the public transportation administration in the state of Oslo in Norway. Unfortunately Ruter's API does not include passenger or tickets sold information, this was provided on request for the years 2015, 2016, 2017 and up till 27 of February for the year 2018.

Chapter 4

Implementation

This chapter describes how the use of the different datasets were implemented and the structure and functions provided by the backend.

4.1 The Backend

The thesis is divided into two: The backend and the frontend. The backend is responsible for providing the frontend all the data and deeper functions it needs. It is partitioned into modules based on each data source available. Each module may also be run individually for testing and easy collection purposes. The twitter module is unique as it requires 4 application programming interface (api) keys. The instructions for this set-up is found in the README.md file in the twitter module's directory.

4.1.1 The Norwegian Institute of Public Health

The data is provided two different sets, which is divided into their separate modules, it was a simple job to plot them in a graph using pythons matplotlib library. Figure 4.1 show the three last seasons of influenza in regards of observed virus infections. The plotting was done manually as NPIH only provides the data in pdf format on their official website[10].

Figure 4.2 shows the influenza-like illnesses (ILI) of the year 2016/2017. This was not done manually as data was provided in a simple .xlsx file which was read by python's openpyxl module, processed and then drawn as a graph.

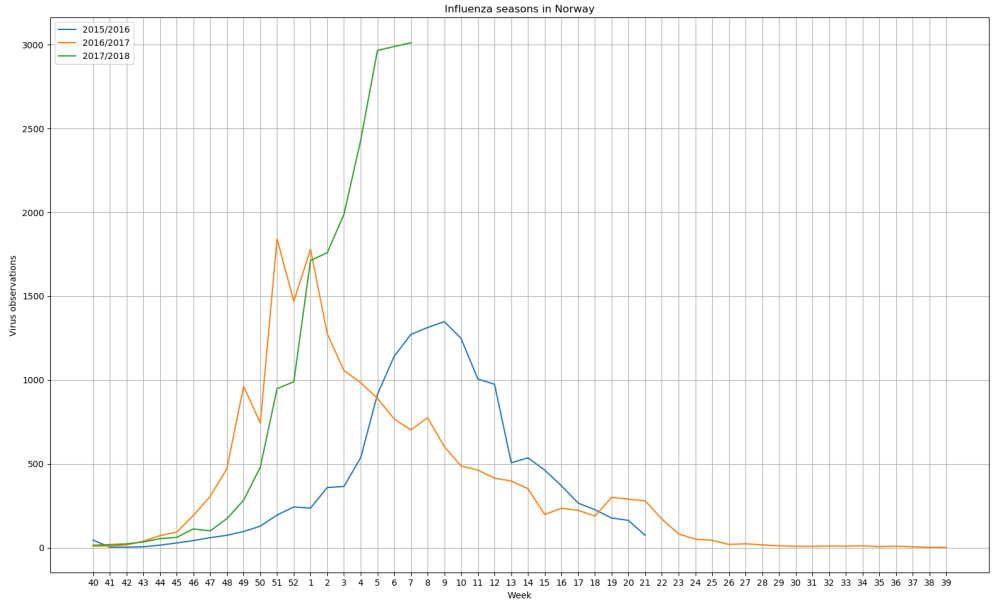

Figure 4.1: Influenza virus observation

4.1.2 The Norwegian Public Roads Administration

From the XLM statistics, some simple graphs were created in python showing the total annual traffic on Norwegian roads from 2002 to 2015 as seen in figure 4.3.

Also derived from this the annual traffic of the two cities Bergen and Oslo, which are towns of interest. Figure 4.4 shows the traffic in Bergen, and figure 4.5 show the traffic in Oslo.

The dataset is an XLM file structure that is downloaded from the NPRA manually. A python program was created that reads through all rows and collects the relevant columns into an array and then draws a graph. For the annual graph, every month of every year was collected. For the towns of Bergen and Oslo the correct roads were identified and loaded from a separate text file, then every year of every month of those roads was collected, loaded into an array and the drawn as a graph. The separate text file is to make it easy to edit should these roads change in the future. The problem of using these datasets is that the data is an average calculation of monthly traffic, this is too coarse for comparison against the influenza data as they are on a weekly basis. A set of traffic registration stations was needed to define the temporal bounds of each area of interest. Defined are the towns of Oslo, Stavanger and Bergen, as well as the whole of Norway on a level 1 basis. The level 1 registrations ensure continually registration throughout the year and is exactly what this project requires.

Figures 4.6, 4.7 and 4.8 shows the traffic on a weekly basis. This provides a better resolution for better analysis.

Figure 4.9, 4.10 and 4.11 shows the different geospatial bounds used to define the cities. The green circles with numbers inside show where and how many traffic registration stations there are.

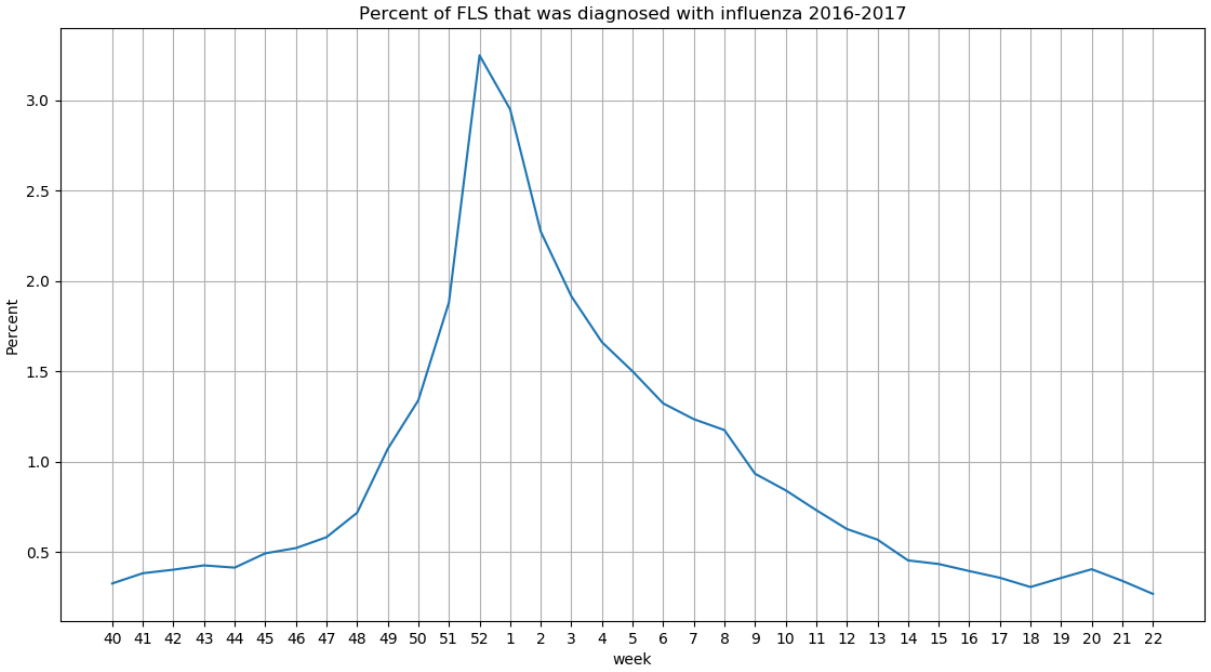

Figure 4.2: Influenza-like illnesses season 2016/2017

4.1.3 Twitter

Using the REST search API it was paramount that in order to build a sufficient dataset acquiring and collecting data had to begin as soon as possible in order to collect enough data for this project. A simple python program was created that takes the input of the API keys and the keywords to be searched upon. The program ensures that no duplicate messages are recorded, and the limit of a hundred tweets was overcome simply by searching for yet another hundred from the last date of the previous hundred until the date limit was reached. The output is appended to a file in this format: id, date, location, tweet.

A simple analysis tool for the Twitter data was created by simply counting how many messages there are. The idea is that during influenza seasons numbers of tweets will rise and vice versa when off the season. Figure 4.12 shows the results.

4.1.4 Kolumbus

The data provided by Kolumbus was in a .png format and had to be converted. From there it was a simple job to plot the data in a python script. Figure 4.13 shows the results.

4.1.5 Ruter

The data provided by Ruter was in a .xlsx file and could easily be read, extracted and plotted by a simple python script. Figure 4.14 shows the results.

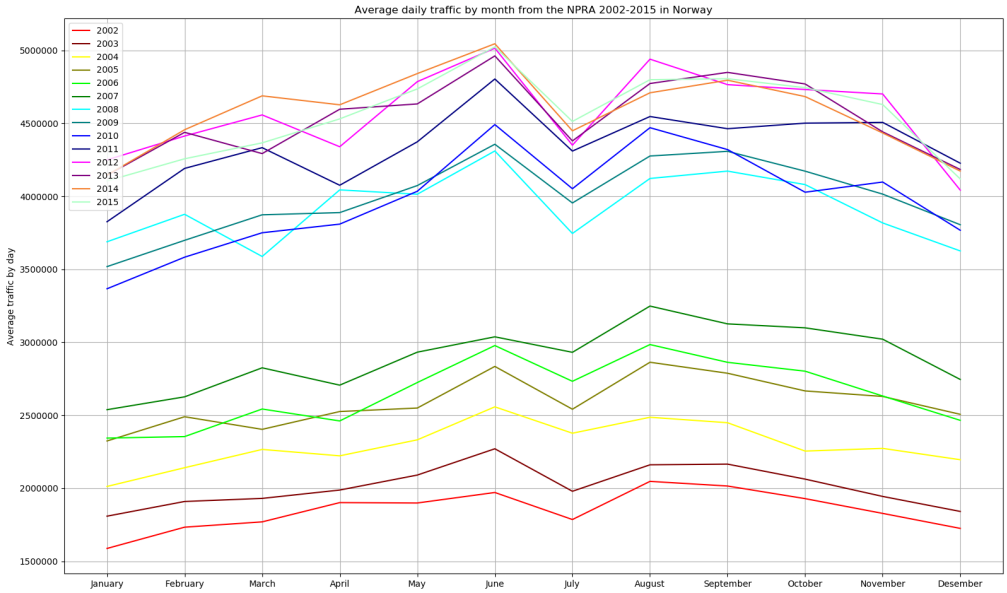

Figure 4.3: Annual traffic 2002-2015

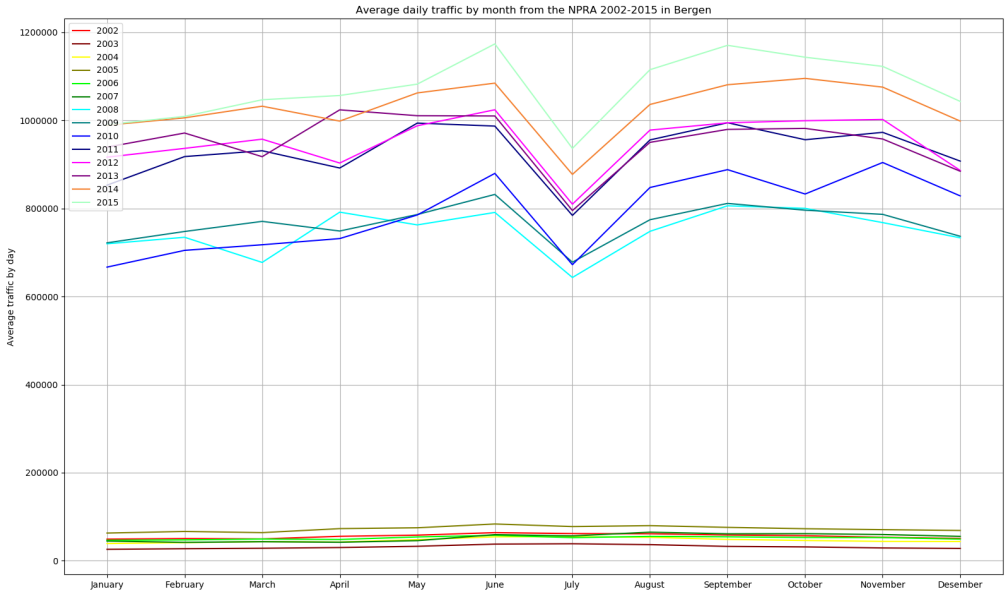

Figure 4.4: Bergen traffic 2002-2015

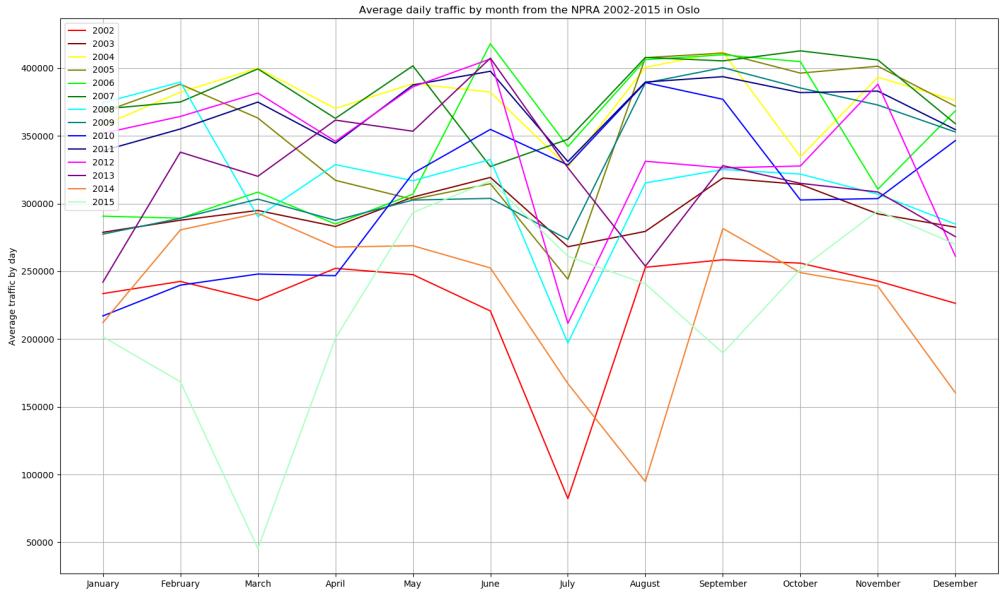

Figure 4.5: Oslo traffic 2002-2015

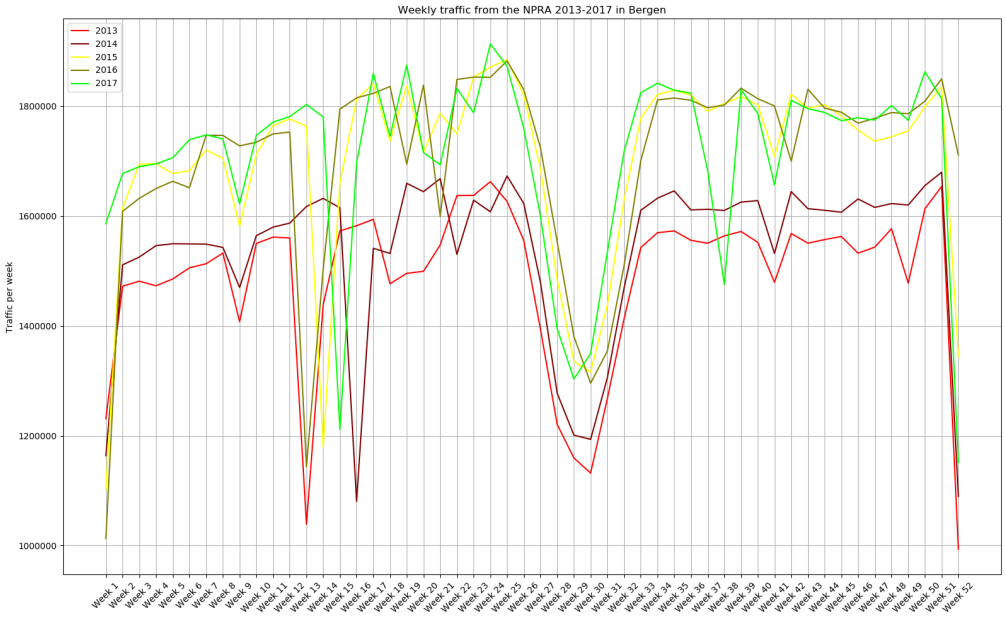

Figure 4.6: Weekly data of the city of Bergen

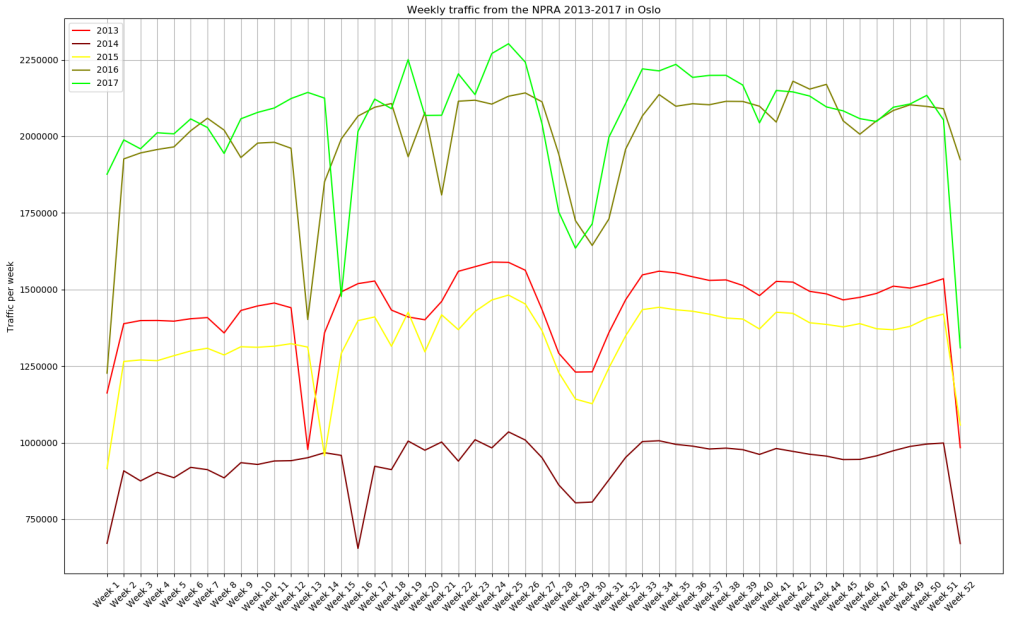

Figure 4.7: Weekly data of the city of Oslo

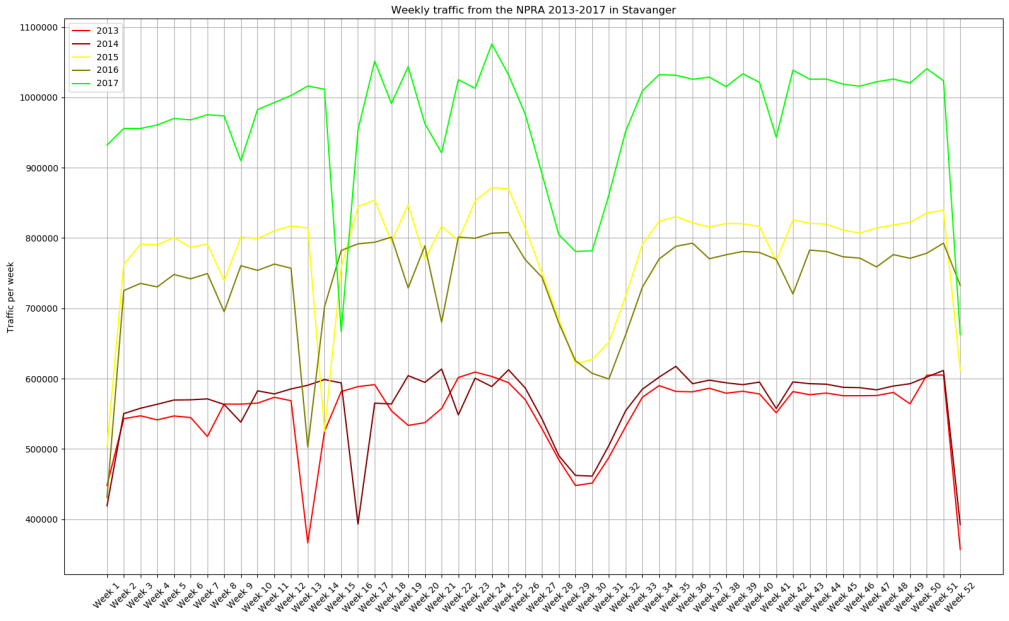

Figure 4.8: Weekly data of the city of Stavanger

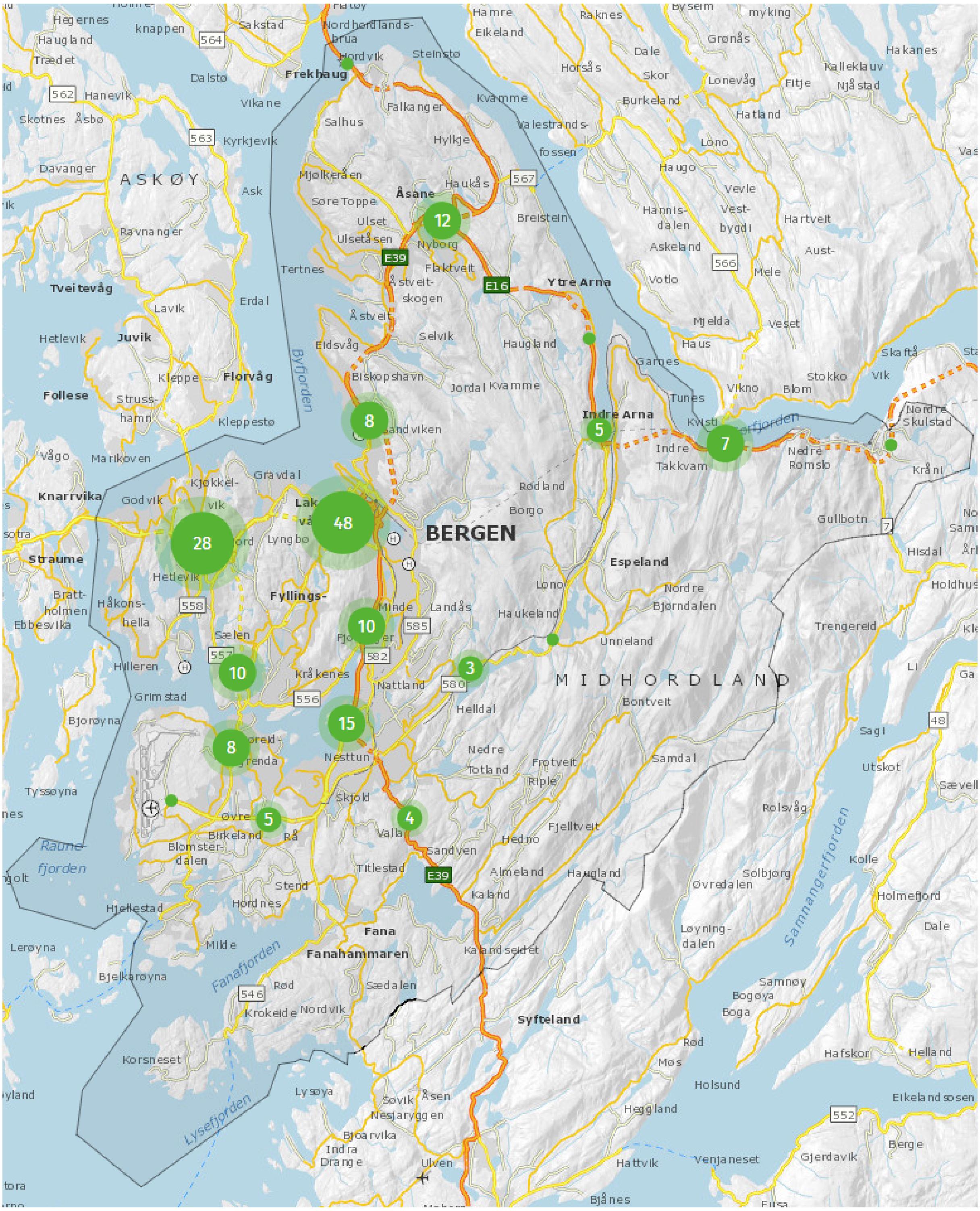

Figure 4.9: Geospatial bounds of Bergen

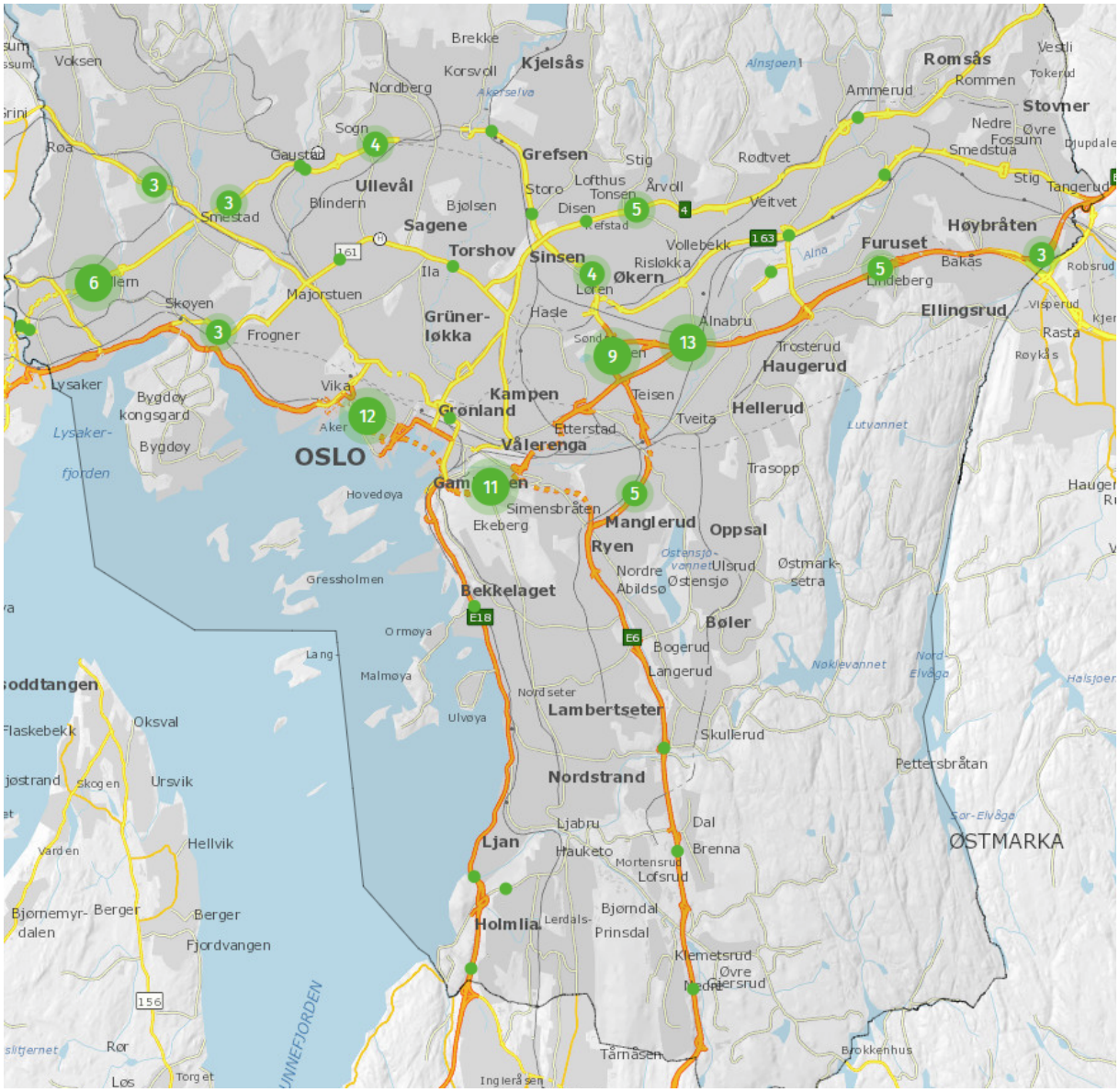

Figure 4.10: Geospatial bounds of Oslo

Figure 4.11: Geospatial bounds of Stavanger

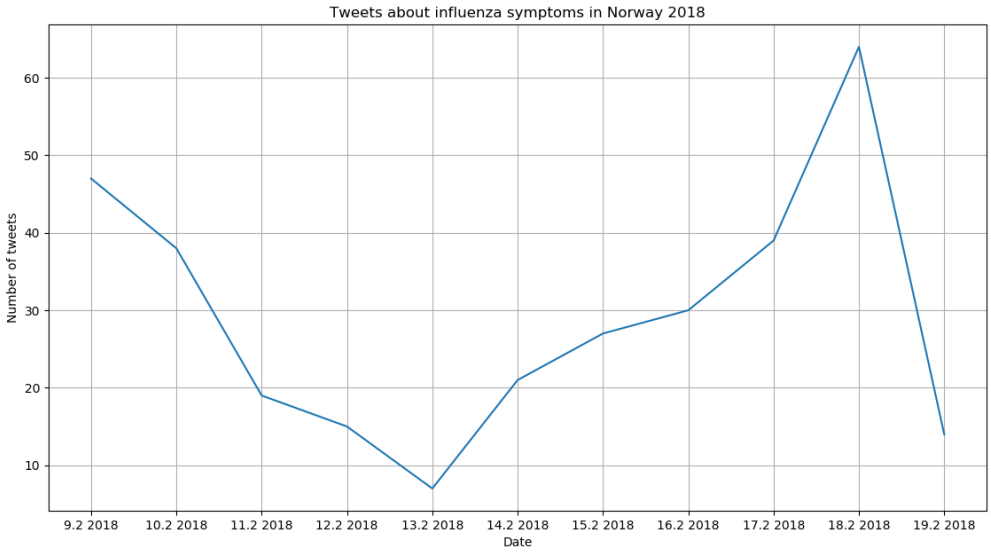

Figure 4.12: Tweets concerning ILS of 2018

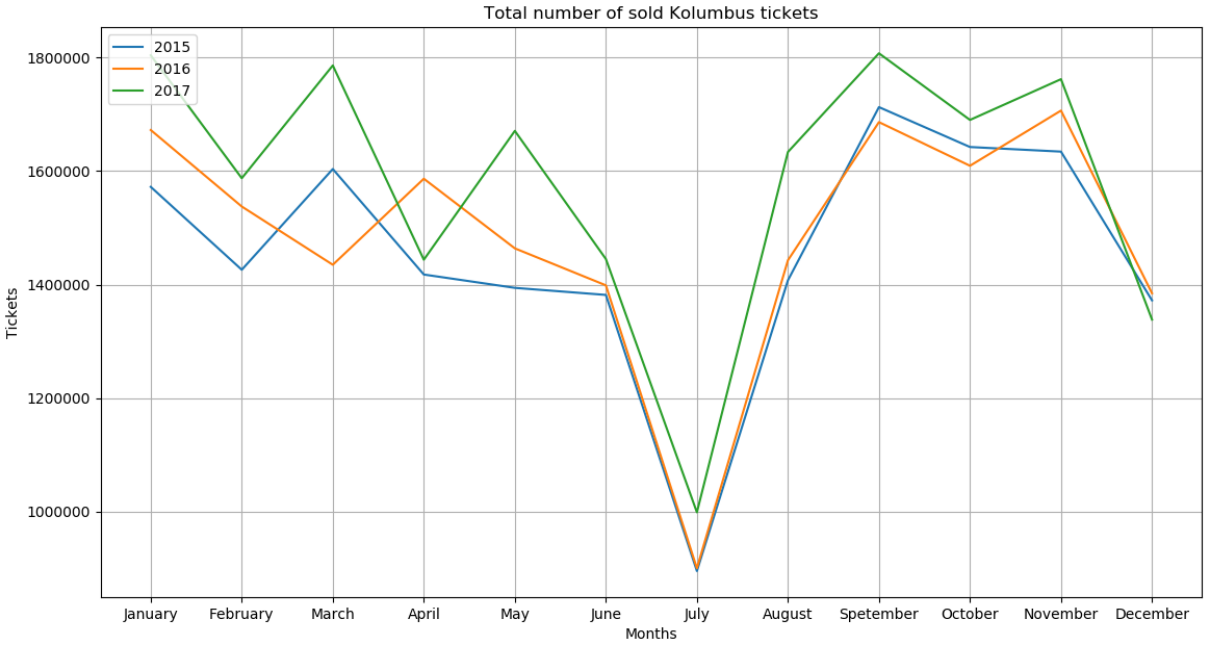

Figure 4.13: Monthly passenger travel with Kolumbus

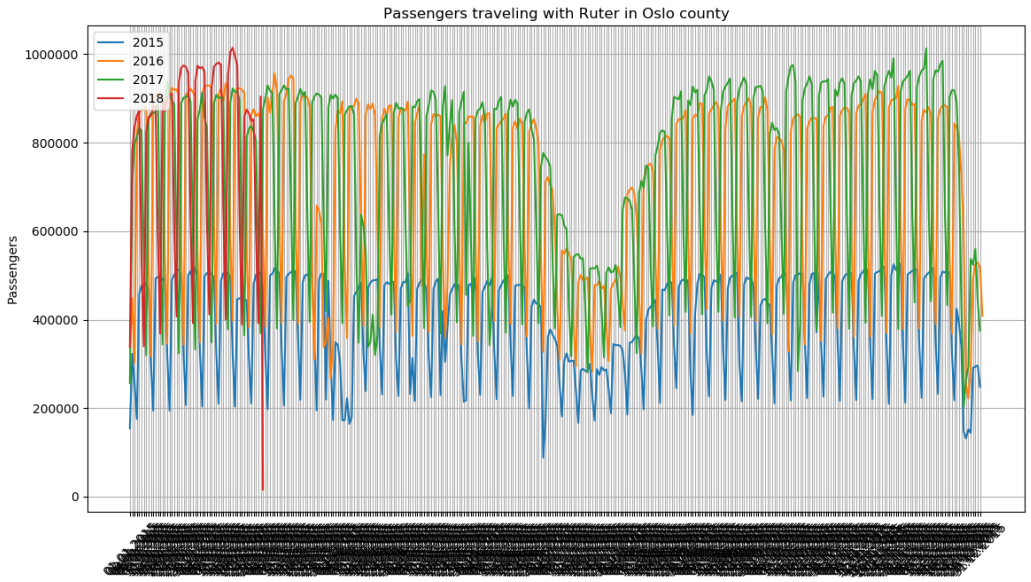

Figure 4.14: Daily tickets sold with Ruter, the year of 2015 does not contain metro services

Chapter 5

Results

5.1 TODO

Chapter 6

Discussion

6.1 TODO

Chapter 7

Conclusion

7.1 TODO

Appendix A

Appendix Title

Bibliography

[1] "The norwegian institute of public health: Influenza symptoms." https://www.fhi.no/en/id/influensa/seasonal-influenza/influenza—fact-sheet-about-season/. Accessed: 2018-06-11.

[2] "The norwegian institute of public health." https://www.fhi.no/en/. Accessed: 2018-06-11.

[3] "The norwegian institute of public health: About the norwegian syndromic surveillance system." https://www.fhi.no/en/hn/statistics/NorSySS/about-the-norwegian-syndromic-surveillance-system/. Accessed: 2018-06-11.

[4] L. O. Grottenberg, O. Njå, E. Tøssebro, G. Braut, R. Tønnessen, and G. M. Grøneng, "Detecting flu outbreaks based on spatiotemporal information from urban systems – designing a novel study," *Icwsm*, vol. 20, pp. 1–7, 2017.

[5] S. B. Elson, D. Yeung, P. Roshan, S. R. Bohandy, and A. Nader, *Using social media to gauge Iranian public opinion and mood after the 2009 election*. Rand Corporation, 2012.

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

[7] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, Association for Computational Linguistics, 2010.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, pp. 851–860, ACM, 2010.

[9] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health.," *Icwsm*, vol. 20, pp. 265–272, 2011.

[10] "The norwegian institute of public health: Influenza information." https://fhi.no/en/id/influensa/seasonal-influenza/. Accessed: 2018-06-11.

[11] "The norwegian public roads administration: Open data, api for developers." https://www.vegvesen.no/en/the+npra/about-the-npra/open-data. Accessed: 2018-06-11.