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 ...

Preface

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

Influenza is a highly 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.

1.2 Objectives

1.3 Structure

The thesis is structured into ... chapters.
Chapter 1, ...
Chapter 2, ...
Chapter 3, ...
Chapter 4, ...
Chapter 5, ...
Chapter 6, ...
Chapter 7, ...

Chapter 2

Related Works

2.1 TODO

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 Folkehelseinstituttet

The Institute of Public Health or Folkehelseinstituttet (fhi) have weekly updates[**fhi**] on the development on the current influenza season as well as previous ones. The reports include numbers of diagnoses from general practitioners (GPs) considering flue-like symptoms (FLS), and hospitalized virus observations with graphs of both. No numbers are appended to the FLS 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 FLS 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 flue-like symptoms based on these characteristics: muscle pain, coughing, fever and the feeling of being sick.

3.2 Vegvesenet

The Norwegian Public Roads Administration (NPRA), or Vegvesenet as it is called in Norwegian, 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 [veg18] there are a few interesting options. They have traffic information in a DATEX API, statistics in XLM and traffic index data relevant to the years before. It is important that the data collected is on a weekly basis atleast 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 in interest is 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 before the patient or even if the patient visits a doctor for diagnosis and treatment. The advantages are instant notification about possible influenza like illness and its spread, against the disadvantages of it being self reported and thus somewhat unreliable. Twitter have 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 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 because 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.

Chapter 4

Implementation

This chapter describes how the use of the different datasets were implemented.

4.1 Folkehelseinstituttet

The data contained two different sets, and it was a simple job to plot them in a graph. Figure 4.1 show the three last seasons of influenza. The plotting was done manually as fhi only provides the data in pdf format.

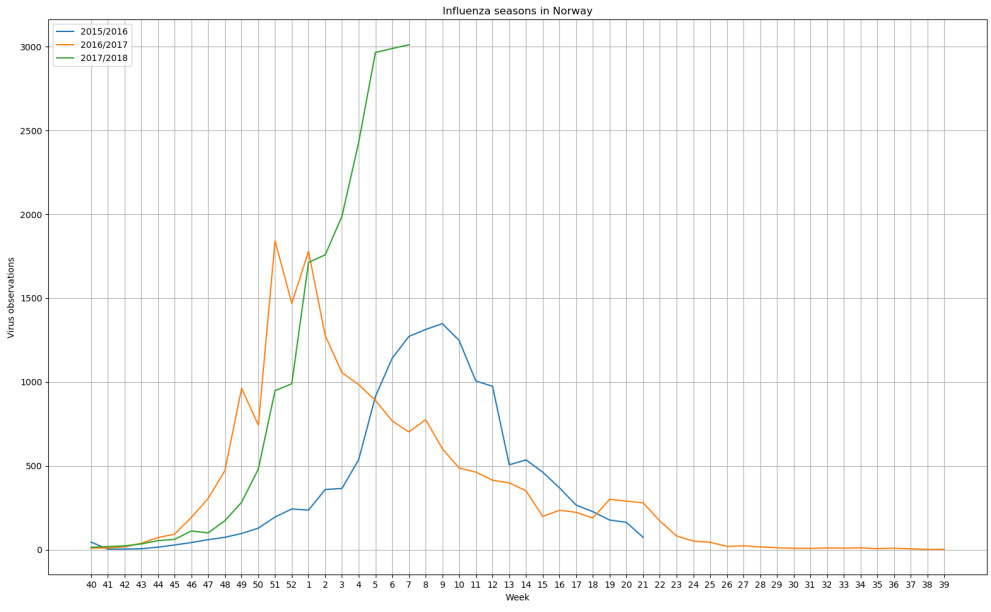

Figure 4.1: Influenza seasons

Figure 4.2 shows the ILS of the year 2016/2017. This was not done manually as data was provided in a simple .xlsx file

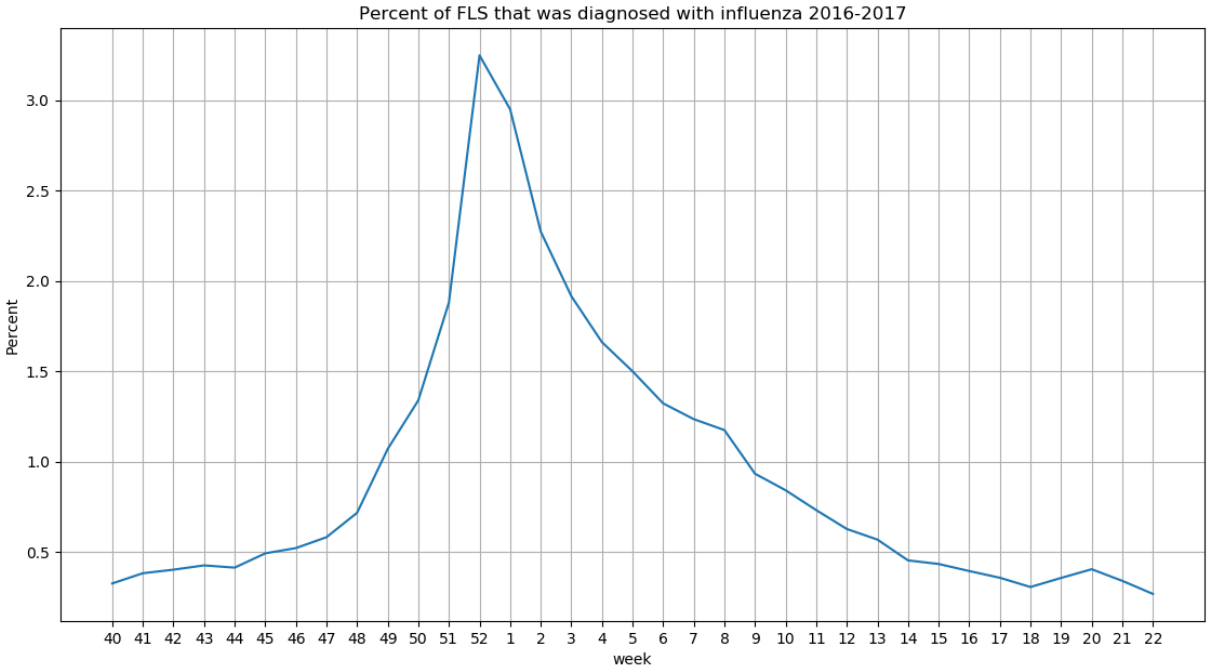

Figure 4.2: Influenza like symptoms season 2016/2017

4.2 Vegvesenet

From the XML 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 in 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 were 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. Luckily upon further investigation and help from the NPRA better data was accessible upon request, hidden from that available on their website. A set of traffic registration stations was needed to define the temporal bounds of each area of interest. Defined are the towns of Oslo and Bergen, as well as the whole of Norway on a level 1 basis. The level 1 registrations ensures 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 cirlces with numbers inside show where and how many traffic registration stations there are.

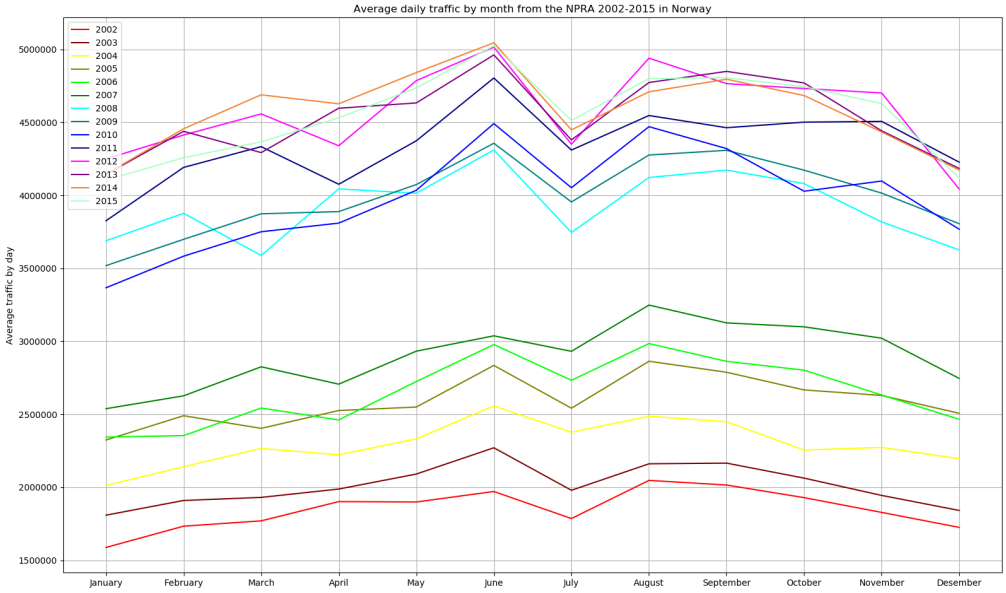

Figure 4.3: Annual traffic 2002-2015

4.3 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. 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.

Analysis of the output then need to be divided into categories based on relevant content.

4.4 Kolumbus

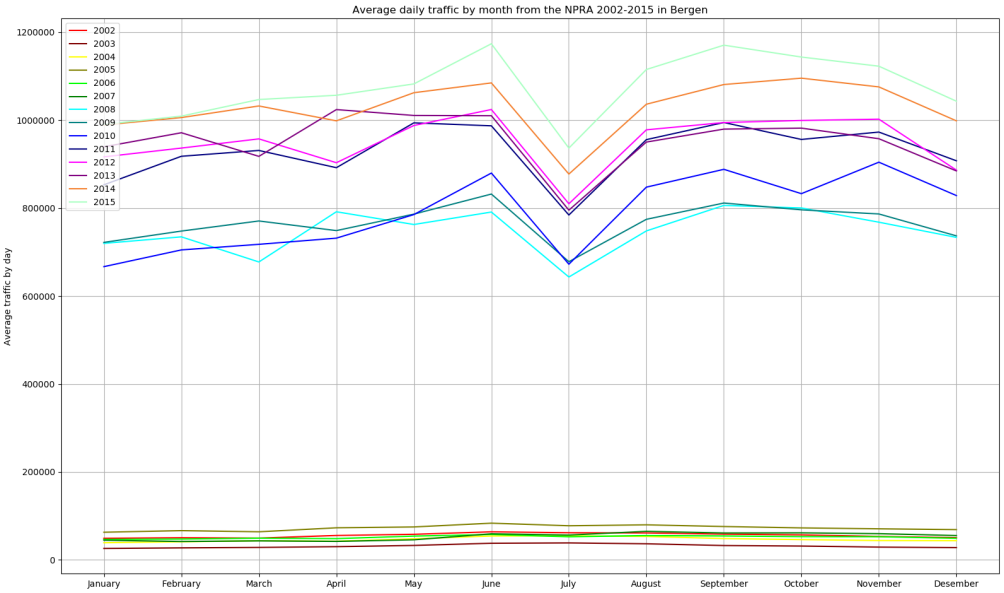

Figure 4.4: Bergen traffic 2002-2015

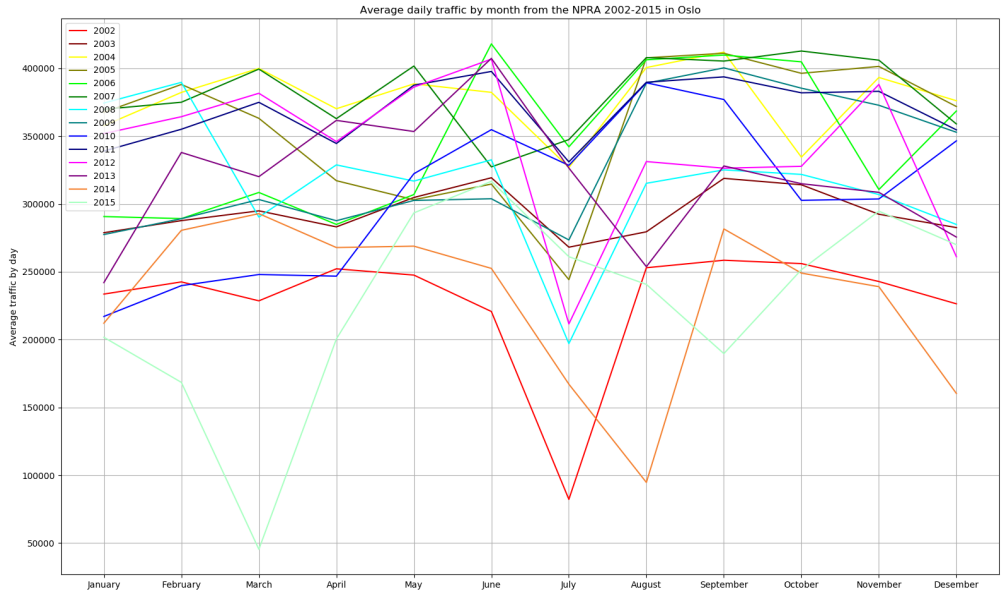

Figure 4.5: Oslo traffic 2002-2015

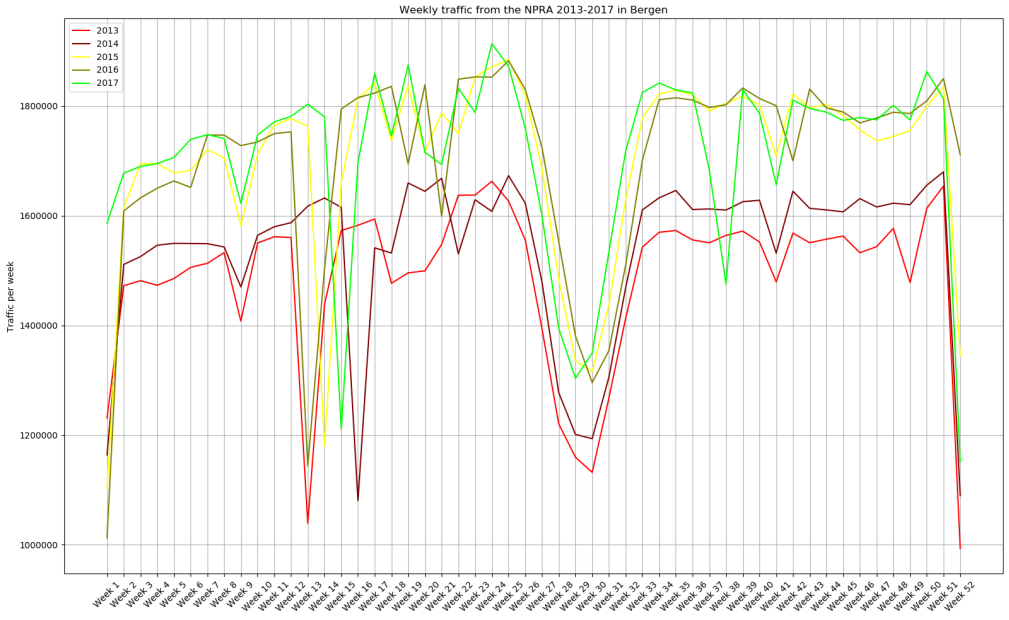

Figure 4.6: Weekly data of the city of Bergen

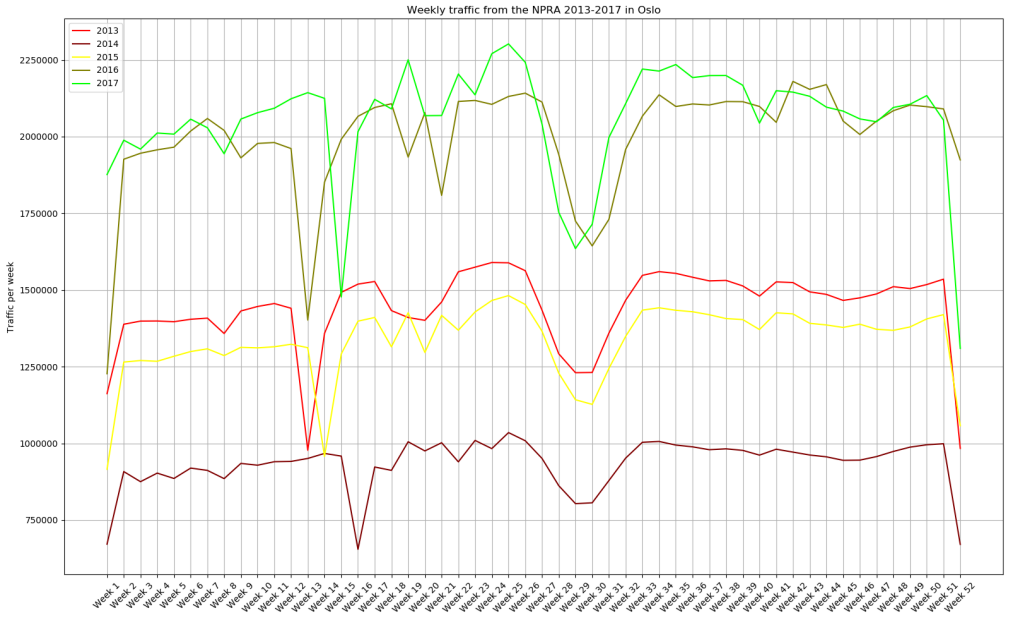

Figure 4.7: Weekly data of the city of Oslo

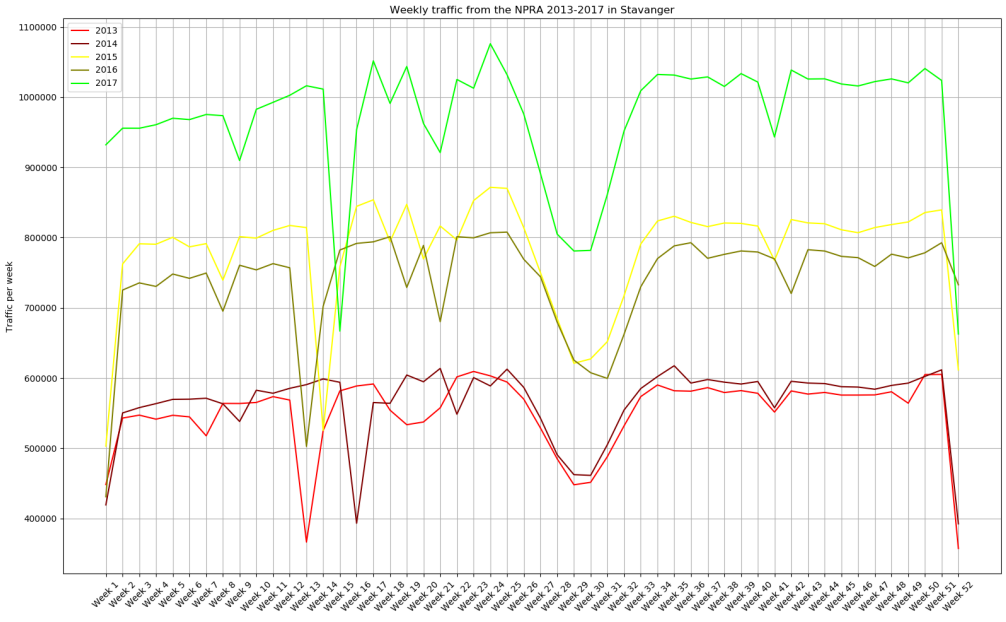

Figure 4.8: Weekly data of the city of Stavanger

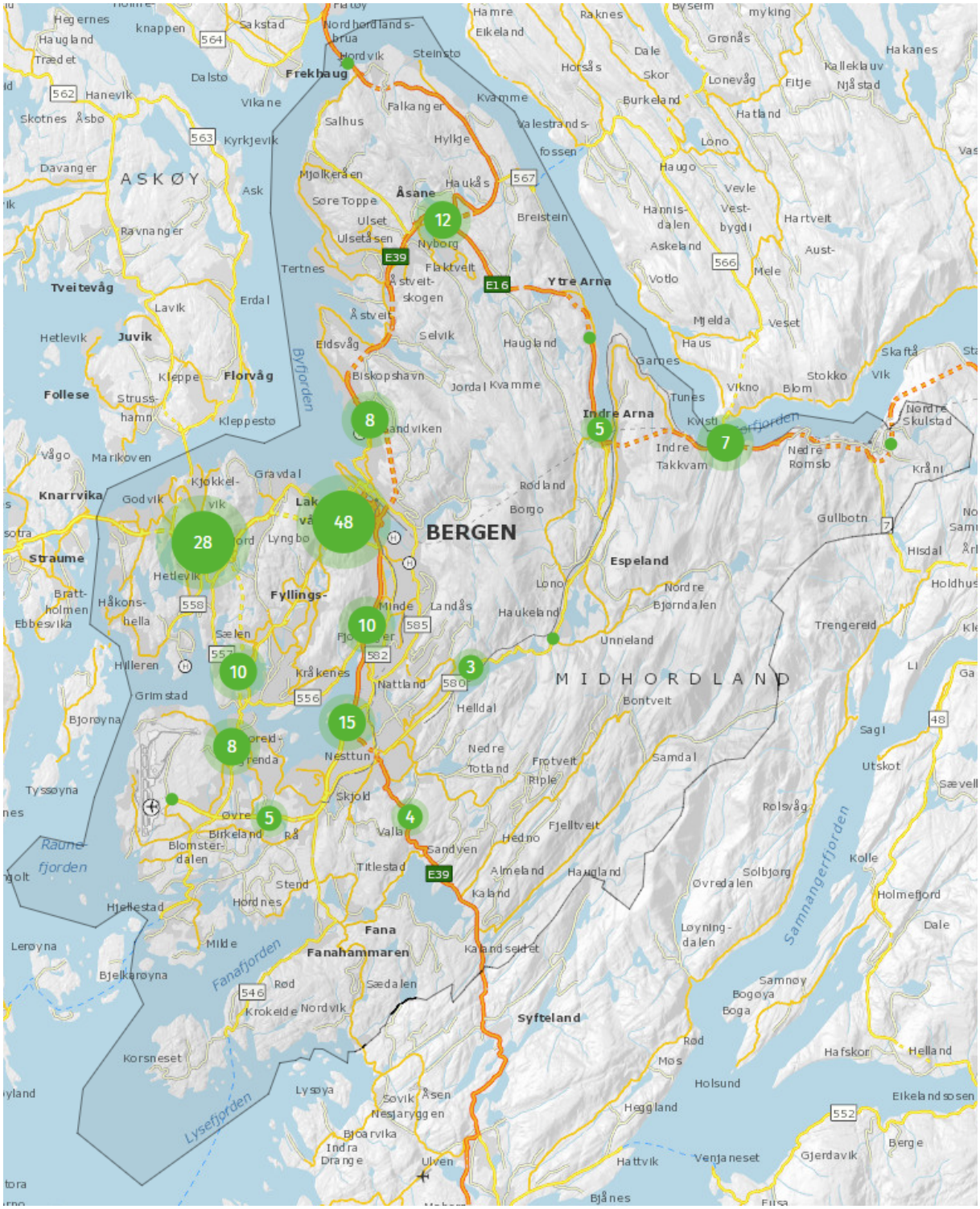

Figure 4.9: Geospatial bounds of Bergen

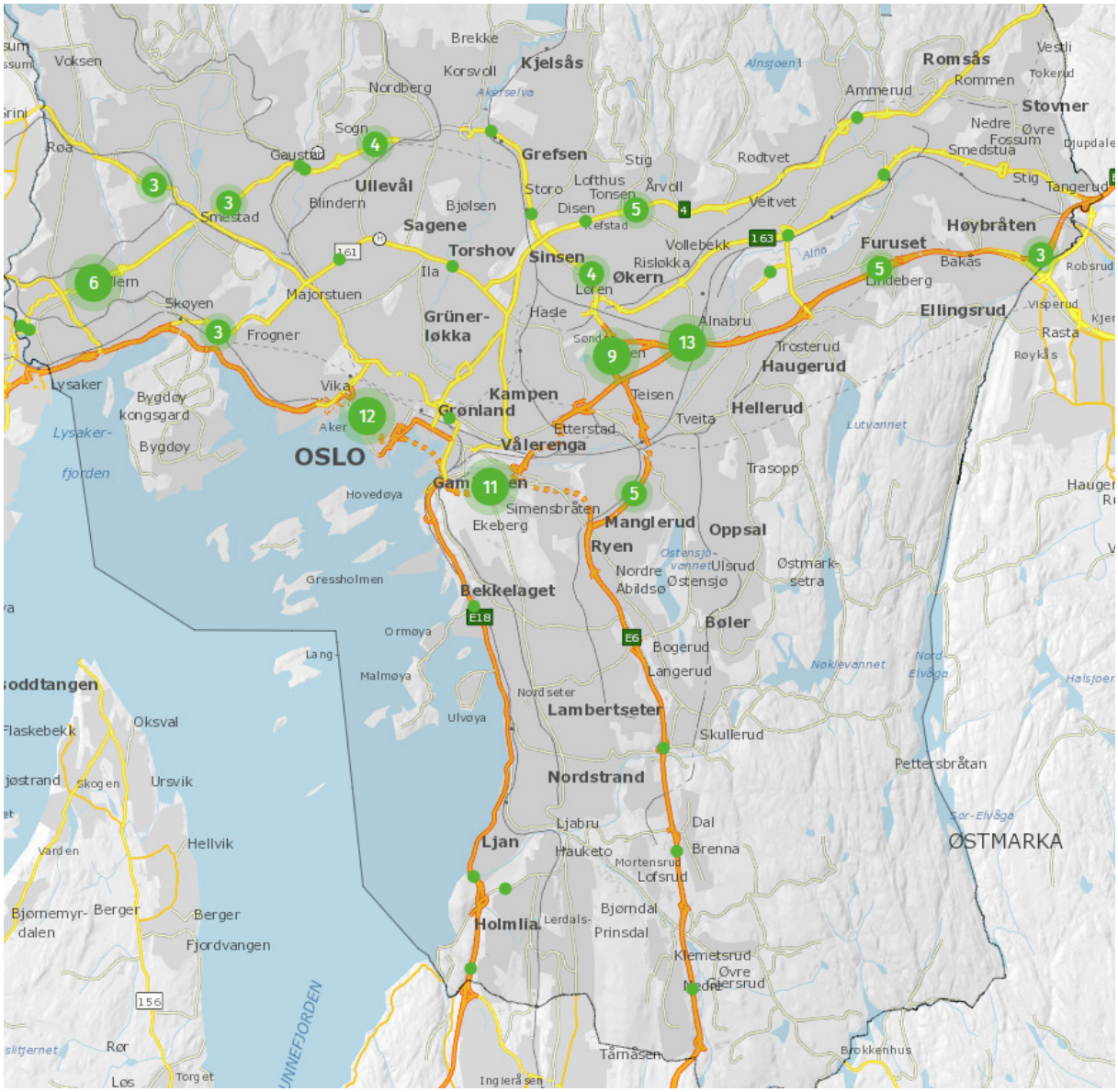

Figure 4.10: Geospatial bounds of Oslo

Figure 4.11: Geospatial bounds of Stavanger

Chapter 5

Results

5.1 TODO

Chapter 6

Discussion

6.1 TODO

Chapter 7

Conclusion

7.1 TODO

Appendix A

Appendix Title

Bibliography

[veg18] Statens vegvesen. *For utviklere/API*. 2018.