Universiteit Leiden
Opleiding Informatica

Public health triangulation:

integrating multiple Ebola-related data sources

Name:	Thomas Helling
Date:	31/07/2015
1st supervisor:	Prof. Dr. H. J. van den Herik
2nd supervisor:	Prof. Dr. A. Plaat
Advisor:	Dr. S. G. R. Nijssen

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Public health triangulation: integrating multiple Ebola-related data sources

Thomas Helling

Preface

Data is rapidly evolving and is becoming one of the main challenges of the 21th century in a large variety of fields. Transforming data into information, information into knowledge, and knowledge into wisdom is where the real power of data science comes from. During my three years of studying Computer Science at Leiden University, data was an interesting topic and it always caught my attention. I was therefore excited to hear about the opening of the Leiden Centre of Data Science in September 2014, and looked forward to participate in the field of data science.

I am grateful for my supervisor Jaap van den Herik, who provided me the opportunity to make a step towards data science and offered me an interesting topic. His enthusiasm and wisdom always motivated me to work on my thesis and also intensified my eager to learn. I would also like to thank my second supervisor, Aske Plaat, who gave me timely and accurate advice whenever I faced challenges during my project. Moreover, I would like to thank Siegfried Nijssen, for providing his knowledge and wisdom about machine learning to improve the results of the investigations.

During my research I experienced difficulties in precisely defining my problem statement and research questions. I would like to thank Meenal Pore, and the health care group of IBM Africa, for sharing their knowledge about the disease outbreak and thus supporting me in defining the proper problem statement and research questions. Moreover, I would like to thank the Leiden Centre for Innovation in The Hague, and in specific Thomas Baar, for providing me with feedback about the intricacies of the problem statement and giving me the opportunity to learn about the public health area.

Finally, I would like to thank my family. My brother who has given me good advice, my sister who was prepared to read my thesis and gave feedback, and my parents for supporting me during the time of writing my thesis.

Thomas Helling,
Leiden, August 21, 2015

Abstract

The Ebola outbreak of 2014 in West Africa struck harder than the 24 previous Ebola outbreaks together. Understanding the spread of the infectious disease can help policymakers and health officials in combating the disease.

In this research, we investigate the possibilities of gaining insights into the disease dissemination by analyzing and combining open data sets. We examine to what extent we can analyze data on sub-national level to gain new insights. We also use machine learning algorithms and a mathematical model to forecast perilous locations.

Surprisingly, we found that combining data sets did not provide new insights into the disease outbreak. The results of the analysis showed that the data sets (1) lack quality or (2) are incomplete and thus cannot help in investigating the disease. Our main recommendation is to improve the data quality since only then data analysis can lead to opportunities to better investigate the disease outbreak and to combat the disease. Then, data science can offer new techniques, that are also used and described in the thesis.

Table of Contents

Preface i

Abstract ii

Table of Contents iii

List of Abbreviations vi

List of Tables vii

List of Figures viii

1 Introduction 1

1.1 Ebola: problem statement . . . 1

1.2 Three research questions . . . 2

1.3 Research aim . . . 3

1.4 Research methodology . . . 3

1.5 Structure of the thesis . . . 3

2 Background 5

2.1 Ebola Hemmorrhagic Fever . . . 5

2.2 Providing insights . . . 6

2.3 Forecasting risk . . . 8

3 Research methodology 9

3.1 Data collection . . . 9

3.2 Data integration . . . 10

3.3 Public health triangulation . . . 10

3.4 Forecasting risk . . . 10

3.5 Data visualization . . . 11

4 Data collection **12**

4.1 Ebola case rates . . . 12
4.2 Sub-national indicators . . . 14
4.3 Weather circumstances . . . 14
4.4 Ebola treatment centers and units . . . 14
4.5 Control efforts . . . 15
4.6 Discussion . . . 15
4.7 Chapter conclusions . . . 16

5 Data integration **17**

5.1 Case rates . . . 17
5.2 Sub-national indicators . . . 18
5.3 Weather circumstances . . . 19
5.4 Target attributes . . . 19
5.5 Result . . . 19
5.6 Chapter conclusions . . . 20

6 Public health triangulation **22**

6.1 Speed of Ebola dissemination . . . 22
6.2 Sub-national indicators . . . 23
6.3 Discussion . . . 24
6.4 Chapter conclusions . . . 25

7 Forecasting risk **26**

7.1 Data preprocessing . . . 26
7.2 Data evaluation . . . 27
7.3 Baseline algorithm . . . 27
7.4 Forecasting by Weka . . . 28
7.5 Forecasting with the IDEA model . . . 30
7.6 Improvements . . . 31
7.7 Discussion . . . 33
7.8 Chapter conclusions . . . 34

8 Conclusions **36**

8.1 Answers to the research questions . . . 36
8.2 Answers to the problem statement . . . 37
8.3 Future work . . . 38

References 39

Appendices 43

A Map of Ebola affected countries 43

B The sub-national indicators 45

C Correlations with the sub-national indicators 46

D Training and test sets 49

E Results of the forecasting models 51

List of Abbreviations

WHO	World Health Organization
IDEA	Incidence Decay and Exponential Adjustment
API	Application Programming Interface
IWI	International Wealth Index
HDX	Humanitarian Data Exchange

List of Tables

1.1 Overview of the research methodology stages and what stages are used to answer the research questions. . . . 3

1.2 Overview of the thesis structure and how each chapter supports in answering the problem statement and research questions . . . 4

5.1 Assigned risk values based on incidence of the upcoming week . . . 19

5.2 The resulting attributes of the data set . . . 20

6.1 Correlation of four of the sub-national indicators with the speed of dissemination and the total number of cases . . . 24

7.1 Forecasting results of the baseline algorithm . . . 28

7.2 Forecasting results of the C4.5 algorithm . . . 29

7.3 Forecasting results of the mathematical model . . . 32

7.4 Forecasting results of the improved mathematical model . . . 33

B.1 Description of the sub-national indicators of the Ebola affected countries . . . 45

C.1 Correlation of the sub-national indicators with the speed of dissemination and the total number of cases before removal of outliers . . . 47

C.2 Correlation of the sub-national indicators with the speed of dissemination and the total number of cases after removal of outliers . . . 48

D.1 Risk distribution of the training sets and test sets without smoothing . . . 49

D.2 Risk distribution of the training sets and test sets with smoothing . . . 50

List of Figures

2.1 The Ebola dashboard that Simon J. Johnson from the British Red Cross created to provide policymakers and health officials more insight into the Ebola dissemination on the situation of 19 July 2015 [16]. 7

4.1 Example of lack of quality control in the collected data sets, where the case rates of two locations have been swapped between November 2014 and January 2015. 13
4.2 Example of two locations where under- and over reporting occurs in the prevalence data set . . 13

5.1 An example that shows the results of the imputation into the data set for the district Kenema . 18

6.1 Example of the take-off date and stabilization date measured by our approach for the district Bo 23
6.2 Scatterplots of the urbanization index with the cumulative number of cases with and without outliers 24

7.1 Example of curve fitting of the mathematical model with overestimation 31
7.2 Example of re-fitting the forecasting curve because the disease stabilized in an earlier stage . . 31
7.3 Incidence of Ebola for Conakry before and after smoothing 32

A.1 The geographical areas on sub-national level of Guinea, Liberia, and Sierra Leone 44

Chapter 1

Introduction

In December 2013, the first case of the Ebola outbreak of 2014 was reported. Although the disease was encountered many times before, 2014 is the year in which the disease spread out at a new level of intensity. This time the Ebola outbreak originated in West-Africa, and the number of Ebola cases became higher than in the 24 previous outbreaks together. With a case amount of around 28,000 and a mortality rate of 41 percent, investigating the disease outbreak could lead to new insights [35].

1.1 Ebola: problem statement

Policymakers and health officials from several organizations have tried to effectively respond to the Ebola outbreak by renewing efforts and developing preventive options, ranging from drug discovery to providing resources [23]. Of course, they analyzed reports by the World Health Organization(WHO) to support decision making. However, they experienced difficulties in deciding what locations are in the highest need of resources and other materials. Therefore, our problem statement reads as follows.

Problem statement: *To what extent is it possible to help policymakers and health officials in reacting effectively to an outbreak of Ebola that is as vigorous as the Ebola outbreak of 2014?*

The main topic of our investigations is to diminish the spreading of the infection over the geographical areas. The research is conducted in the field of data science. More precisely, we will investigate the possibilities of combining disparate data sets from open data initiatives to gain insights into Ebola and the way it is diminished. Two of these open data initiatives are the Humanitarian Data Exchange [14] and the Ebola Open Data Jam[22]. Data sets about health facilities, burial teams, rates of Ebola cases, and more are available on these open data initiatives. As the availability of data is increasing daily, analyzing this data can provide

opportunities to better understand the way to combat and adequately respond to an outbreak of a disease such as Ebola.

1.2 Three research questions

In order to answer the problem statement, we formulate three research questions. They will guide our research.

Analyzing disparate data sets provides us with information about Ebola and its intricacies. Transforming the information into knowledge is necessary to obtain adequate insights into the disparate data sets. A large variety of data sets is supplied on the open data initiatives, but not all of the data sets are useful for the research. Therefore, our first research question reads as follows.

Research question 1: *What kind of disparate Ebola-related open data sets could provide adequate insights into the dissemination of Ebola?*

Disparate data sets provide knowledge. Combining multiple data sets can provide us with new information and thus knowledge about Ebola and the way it disseminates. We will investigate whether certain circumstances at a specific location influence the dissemination of Ebola. For example, we analyze whether districts with better wealth circumstances have a lower spread of Ebola. Therefore, our second research question reads as follows.

Research question 2: *To what extent can we combine the Ebola-related data sets to gain more knowledge about the disease dissemination?*

A better understanding of Ebola and its consequences by combining related data sets can be of critical importance in preventing and/or combating the dissemination of Ebola. However, we can also provide actionable insights into fighting the disease by combining data sets to forecast whether an outbreak may occur or if an outbreak will get more intense at a specific location. Forecasting high risk Ebola locations may intensify the attitude towards providing resources or other preventive options. Hence, our third research question reads as follows.

Research question 3: *To what extent can we forecast perilous Ebola locations with the use of the Ebola-related data sets?*

Answering these research questions will demonstrate how and to what extent we can use Ebola related data sources to provide knowledge about the intricacies of the Ebola outbreak to policymakers and health officials. The knowledge may lead to a better understanding of the disease outbreak and may support policymakers and health officials in decision making in the longer run and even in the short run.

1.3 Research aim

Data science is applied in a large variety of fields, and also in the Ebola outbreak of 2014 organizations started to collaborate by joining open data initiatives to allow data scientists to help combat the disease [25]. The aim of this research is to improve to combat against diseases such as Ebola in countries similar to the West African countries by investigating the outbreak. We hope to provide organizations with insight into a way to act more preventive towards the way the disease disseminates so that they can contain the intensity of a future outbreak.

1.4 Research methodology

To answer the problem statement and research questions, we formulated a research methodology. The research methodology consists of five stages; (1) data collection, (2) data integration, (3) public health triangulation, (4) forecasting risk, and (5) data visualization. Table 1.1 addresses what stages of the research methodology are used to answer the research questions. A description of the stages is given in chapter 3.

Research methodology stage	RQ1	RQ2	RQ3
Data collection	✓	✓	✓
Data integration	✓	✓	✓
Public health triangulation		✓	✓
Forecasting risk			✓
Data visualization	✓	✓	✓

Table 1.1: Overview of the research methodology stages and what stages are used to answer the research questions.

1.5 Structure of the thesis

Chapter 1 formulates the problem statement and research questions. It also states the aim and shows the research methodology. Chapter 2 provides background information about Ebola and the related work in the field of Ebola and data science. Chapter 3 discusses the methods that we will use to analyse the data. Chapter 4 discusses the data sets that have been chosen for the research. Chapter 5 discusses the integration of the data sets. Chapter 6 will compare different data sets to gain new knowledge. Chapter 7 discusses the forecasting possibilities of risky Ebola locations. Chapter 8 concludes by answering the three research questions and problem statement. Table 1.2 provides an overview of which chapters will be used to answer the research questions and problem statement.

Chapter	PS	RQ1	RQ2	RQ3
Chapter 1: Introduction	✓	✓	✓	✓
Chapter 2: Background	✓	✓	✓	✓
Chapter 3: Research methodology				
Chapter 4: Data collection		✓		
Chapter 5: Data integration			✓	
Chapter 6: Public health triangulation			✓	
Chapter 7: Forecasting risk			✓	✓
Chapter 8: Conclusions	✓	✓	✓	✓

Table 1.2: Overview of the thesis structure and how each chapter supports in answering the problem statement and research questions

Chapter 2

Background

In this chapter we provide a background of Ebola and introduce the related work that has been done in the field of Ebola and data science. In section 2.1 we provide information about the disease itself and discuss why it broke out at this level of intensity. Section 2.2 discusses methods that have already been applied in the field of Ebola to provide health officials and policymakers with new insights. In section 2.3 we discuss three mathematical models that have been applied to forecast the intensity of the Ebola outbreak.

2.1 Ebola Hemmorrhagic Fever

The first case of the Ebola Hemorrhagic Fever was administered in 1976. Ever since, small outbreaks of the disease appeared multiple times in Uganda, Sudan, Zaire, and the Democratic Republic of Congo. The virus is transmitted by blood, sweat, urine, vomit, and other bodily fluids. The first symptoms can occur anywhere between 2 to 21 days after exposure to Ebola, but the average is 8 to 10 days [31]. These symptoms include a fever, a severe headache, muscle pain, a sore throat and a weak feeling. More intense symptoms occur in an advanced stage of the infectious period, such as abdominal pain, diarrhea, vomiting, and internal- and external bleeding out of the nose, ears and eyes [3].

In December 2013, the first Ebola case of the 25th known outbreak of Ebola was encountered [35]. Since that moment, the number of cases grew at an unexpected rapid pace. By August 2014, the outbreak was so intense that the WHO declared the epidemic to meet the conditions for a Public Health Emergency of International Concern, meaning that the outbreak is an extraordinary event which is determined to (1) constitute public health risk to other states through the international spread of the disease and (2) to be a situation that is unusual, unexpected, carries implications for public health beyond the affected state's national borders and may require immediate international action [33, 34].

The spread of the outbreak of 2014 has been intensified by several factors [6, 24]. We mention four of them. (1) Years of civil war have led to distrust of the citizens towards authorities. Constant criticism towards the way the public health organizations act against the disease outbreak does not improve the trust of citizens and the citizens actually think the government is the cause of the disease [19]. (2) There is a scarcity of health care workers and materials in the affected countries [6]. It was the first time Ebola struck in West Africa, and therefore health care workers expected the first patients to be infected by Malaria, a disease with similar symptoms. As the health care workers did not expect that the patients were infected by Ebola, they infected more patients by using the same injection needles on multiple patients. (3) Religions groups in West Africa believe in burial ceremonies that include touching and washing the bodies of the dead, while the virus is still active on the dead body. For example, a religion group used sponges to clean the body and believed that they were able to consume the wisdom of the dead body by rubbing the sponge over their own head after cleansing the body. What they did not know, is that they actually infected themselves with Ebola. (4) There is a high mobility of the population across borders in West-Africa, and therefore the disease disseminates faster towards new areas [24].

In Ebola reporting by the WHO, there are three sorts of Ebola cases; (1) A confirmed case, which is a suspected case that tested positive on the virus in the laboratory. (2) A suspected case, which is a person that has one or more of the Ebola symptoms, or had contact with a suspected Ebola case, probable Ebola case, confirmed Ebola case, or a dead or sick animal. (3) A probable case, which is a suspected case that is evaluated by a clinician, or any diseased suspected case that had some association with a confirmed case [32].

2.2 Providing insights

So far, different organizations have tried to empower and support decision making in preventing and containing the spread of Ebola by visualizing the disease dissemination using data sources.

IBM saw an opportunity to help and provided advice to the Sierra Leone government by combining data sources and engaging with citizens. They used Sierra Leones Open Government Initiative in combination with the possibility for citizens to report Ebola-related issues and concerns [29]. The system maps two phenomenons: (1) cities with a low number of resources and (2) daily experiences of citizens affected by Ebola [2]. Mapping these data collections supports the government in managing resources necessary to fight the disease.

Simon J. Johnson, from the British Red Cross, created Ebola-related dashboards to better understand the disease outbreak over time [16]. The goal of the dashboards is to help Ebola responders analyse the data that is available on the open data initiatives. Figure 2.1 provides an example of a dashboard that was created by

Simon. It provides four figures about the Ebola situation until 19 July 2015. The upper left figure provides information about the new confirmed cases in the last two weeks, which is important because it can explain where perilous locations are in the observed situation. The figure on the upper right provides four key statistics of the Ebola outbreak until 19 July 2015; (1) the total number of cases, (2) the total number of deaths, (3) the total population size, and (4) the mortality rate. The two other figures are interactive. It provides an option to hover over a district, county, or prefecture with your mouse on the map. If you click on one of the locations, the right figure on the downside shows the corresponding cumulative number of cases over the outbreak period for that location. Using the dashboard, policymakers and health officials can gain insight into the way the disease disseminates per location and what location is the most risky at this moment.

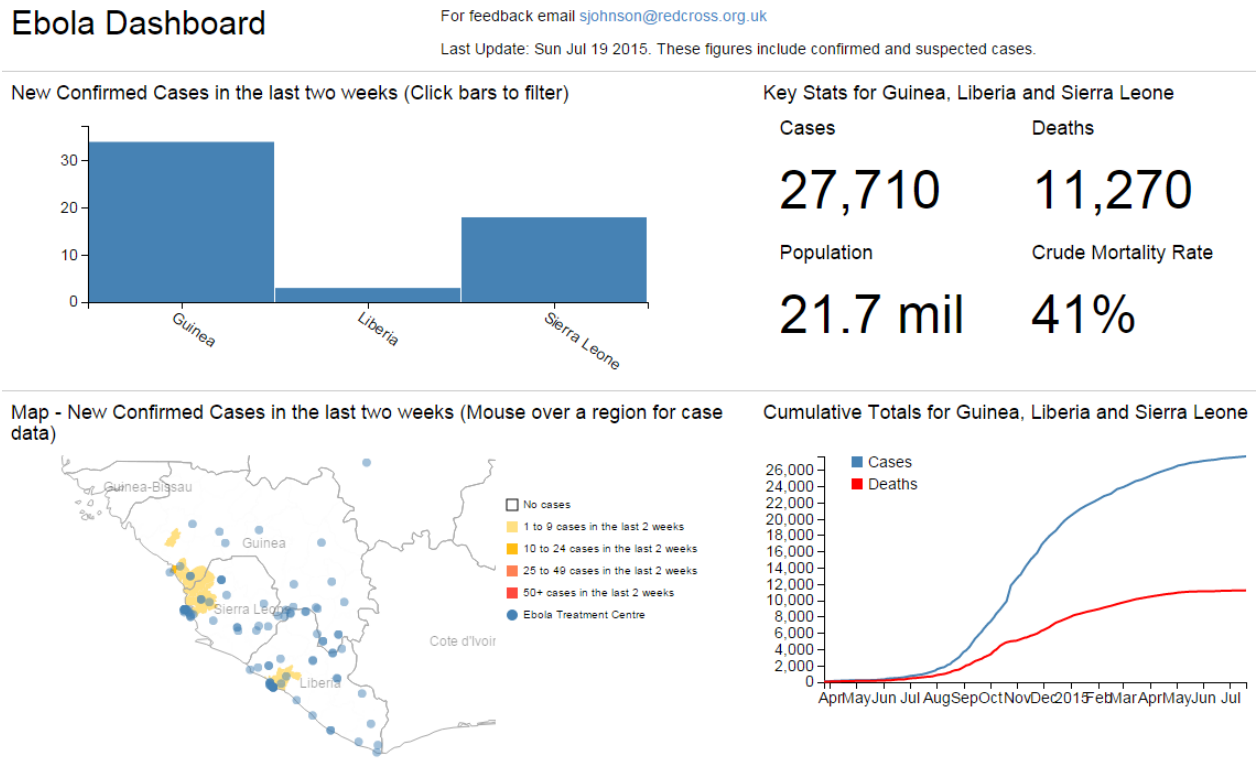


Figure 2.1: The Ebola dashboard that Simon J. Johnson from the British Red Cross created to provide policymakers and health officials more insight into the Ebola dissemination on the situation of 19 July 2015 [16].

A further source of information is the WHO, which releases a weekly situation report about the disease dissemination with the corresponding visualization of that week [35]. These reports include (1) a summary of the outbreak, (2) the number of cases reported in the last week, and (3) how well the response against the disease is doing per country.

Of course, other kinds of data visualization methods are also used to gain knowledge about the disease. For example, HealthMap provided a map with Ebola related news activity over the time of the outbreak [12]. Data science strategies such as analyzing HealthMap may also give new information about the disease outbreak.

2.3 Forecasting risk

Forecasting by using machine learning algorithms is not yet applied to disease outbreaks and to measure the amount of risk at specific locations. However, since the end of the twentieth century mathematical modeling is used more frequently to empower and support decision making in the public health area [17]. In the field of Ebola, there have been multiple successful and unsuccessful attempts in forecasting the future incidence by using mathematical models.

One of the key values in mathematical models for infectious diseases is the reproduction number R_0, which describes the average number of persons infected by a single infected case over the infectious period. R_0 is difficult to compute, because it is computed by the aid of the transmission rate and contact rate of an infected case. These variables are extremely hard to determine without intensive control [17]. Most of the mathematical models attempt to make an estimate by fitting the data on formulas. The reproduction number of Ebola varies between 1.5 and 2.5, which is relatively low in comparison to a disease such as the measles with a reproduction number of between 12 and 18 [1].

Fisman et al. [8] introduced the Incidence Decay with Exponential Adjustment (IDEA) model, which describes the disease dissemination with the previously mentioned reproduction number R_0, and an additional variable which describes the decay of the outbreak factor d. d represents the control efforts against the disease, public behaviour changes, or any other dynamic change that slows disease transmission. They fitted the model on two straightforward parameters: (1) the number of cases and (2) the number of deaths. However, d was computed by using historical data. If policymakers and health officials are starting to respond more intense to the disease outbreak, the value of d would not be sufficiently accurate to forecast the future incidence over a longer time. The same kind of results were found by the WHO Ebola Response Team, but they were also not able to take the control efforts into account as well [31].

Models with more than two parameters have also been used to provide policymakers and health officials with insights into the disease dissemination by mathematical modeling. Drake et al. investigated a model by using more parameters than the IDEA model in Liberia, such as infected cases that are looking for hospitalization and individual behaviour of citizens. Using their model, they showed that the dissemination of the disease depends on both hospital availability and individual behaviour towards Ebola. However, even if the capacities of hospitals tend to increase or decrease, individual behaviour is the most important aspect in the way the dissemination of the disease diminishes and can be taken care of [7].

Chapter 3

Research methodology

This chapter describes our approach to answer the problem statement and research questions. Section 3.1 discusses the identification and selection of relevant data sets for the research. In section 3.2 we discuss our approach to combine the disparate data sets and deal with missing data. In section 3.3 we consider how we investigate the combination of data sets to gain new insights. Section 3.4 provides a method to forecast perilous situations with the use of a data mining tool. In the last section, we consider the methods used to visualize and analyze the data sets.

3.1 Data collection

We analyzed a range of data sets from the HDX [14] and the Ebola Open Data Jam [22]. Since there is a large variety of available data, it is important to notice how we identified data sets that are relevant for the research.

A data set has to meet three conditions to be useful for the research; (1) the data set has to consist of the attribute(s)[1] time and/or location on sub-national level[2]. Which of the two attributes is necessary depends on the type of data set. For instance, a data set that contains information that does not change over time can be considered as a data set in which we only need the location attribute. (2) The data set has to be reliable. A data set is reliable if it is supplied by a trusted source (e.g. the WHO) and if the data is up-to-date. (3) The data set has to be relevant. Relevancy depends on whether we can imagine that the attributes of the data set influence the way Ebola disseminates over time. In most cases these three conditions cannot be met by the open data initiatives. Therefore, availability of the data sets is the most important aspect in finding data sets that are useful for the research.

[1] An attribute is a quality or feature regarded as a characteristic or inherent part of someone or something.
[2] The districts of Sierra Leone, counties of Liberia and the prefectures of Guinea. The geographical areas are given in appendix A

3.2 Data integration

Data integration is defined as combining data from multiple sources to provide an unified view of the disparate data sets [18]. We will combine disparate data sets on the attributes time and location. Therefore, we created a data set with the time and location attribute. Afterwards, we combined each data set on the attributes time and location by developing PHP algorithms and by using a MySQL database. For two data sets, the location attribute is insufficiently location specific. For example, a data set provided information on regional level instead of prefecture level. To integrate these values with other data sets we used (reversed) geocoding of the Application Programming Interface (API) from Google Maps [10].

Missing data can be an issue on two fronts: (1) missing fields in the combined data set and (2) missing information that is not available in any of the available data sets. There are several methods for handling missing values in a data set [21]. The most common method is the use of multiple imputation on missing data values. There are several methods for multiple imputation, ranging from imputing the mean of the attribute to considering a given number of instances that are most similar to the instance of interest. In this research, we imputed data by searching for the value that is most likely to be standing there. Moreover, missing information that is not available in any of the data sets is something we cannot include in the research. Therefore, the focus of the research is on the attributes that are available for analysis.

3.3 Public health triangulation

Denzin defined triangulation as the combination of methodologies in a study of the same phenomenon [5]. However, triangulation has had multiple definitions since then. One of them is public health triangulation. Public health triangulation is defined as combining and interpreting data from multiple sources to empower and support decision making [27]. As we combine the data sets for new information, public health triangulation comes in when we transform the information about the disease dissemination into knowledge. We computed and compared Ebola-related indicators and measure whether the indicators can provide insight into the way the disease disseminates.

3.4 Forecasting risk

There is a fundamental difference in forecasting and prediction. Forecasting relies mainly on data from the past and present and analysis of trends, it is an extrapolation of the past into the future. Prediction is the way things will happen in the future, often but not always based on experience or knowledge. Therefore,

forecasting is a subset of prediction. In our case, we are extrapolating data from the past into the future, and therefore we are forecasting.

To forecast perilous situations, we used two different approaches. (1) Weka, a data mining tool that analyzes the data set and tests multiple classification algorithms on the data set [30]. The tool has a collection of machine learning algorithms for data mining tasks and it provides various methods for preprocessing and evaluating data. (2) We fitted the Ebola case data by using a mathematical model to extrapolate the case data and forecast the risk in the upcoming week with the use of Python's package SciPy. We used the confusion matrix to evaluate the forecasting accuracy of the models and to compare the models to each other.

In section 3.2 we discuss the combination of disparate data sets. In order to forecast and analyze the perilous situations, we defined a target attribute that describes the risk of a situation for the upcoming week. The target attribute can then be used to train the forecasting models.

3.5 Data visualization

To visualize the results we used three methods: (1) a web-based platform programmed in PHP, Javascript and Highcharts to visualize and analyze the prevalence data, (2) Microsoft Excel, for scatterplotting relations between indicators, and (3) SciPy's pylab to visualize the forecasting methods of the mathematical model.

Chapter 4

Data collection

In this chapter we discuss the chosen data sets for the research. The data sets have to meet the three conditions that we described in section 3.1; (1) the attributes time and location, (2) relevancy, and (3) reliability. However, in most cases these three conditions cannot be met by the open data initiatives. The most important aspect for the selection of data sets is therefore availability. The first part of this chapter (section 4.1 to 4.3) discusses the three data sets that can be used to analyze the disease outbreak; (1) the prevalence data, (2) the sub-national indicators, and (3) the weather circumstances. The second part (section 4.4 and 4.5) describes two data sets that could be of critical importance for the analysis but are incomplete. In section 4.6 we discuss two of the problems that we experienced by collecting the data sets. Finally, section 4.7 provides an answer to research question 1.

4.1 Ebola case rates

The prevalence of Ebola is important for the research because it will be the basis of our investigations in the geographical areas. There is a large number of data sets that include the prevalence data over time. Analysis of the data sets showed the data sets consisted of many inaccurate or outdated data points, and therefore the quality of the data sets is lacking. For instance, a data set provided by the WHO (1) swapped the values of two districts for two months, and (2) only provided data until the end of March. An example is provided in Figure 4.1, in which the cumulative number of cases of Western Area Rural (the orange line) and Western Area Urban (the blue line) are swapped between November 2014 and January 2015. This problem is a result of the fact that the WHO updates the data manually.

Simon J. Johnson from the British Red Cross created a data set that includes the number of cases from the 7th of April until the 22th of June. The data set consists of the prevalence data of the counties of Liberia, the

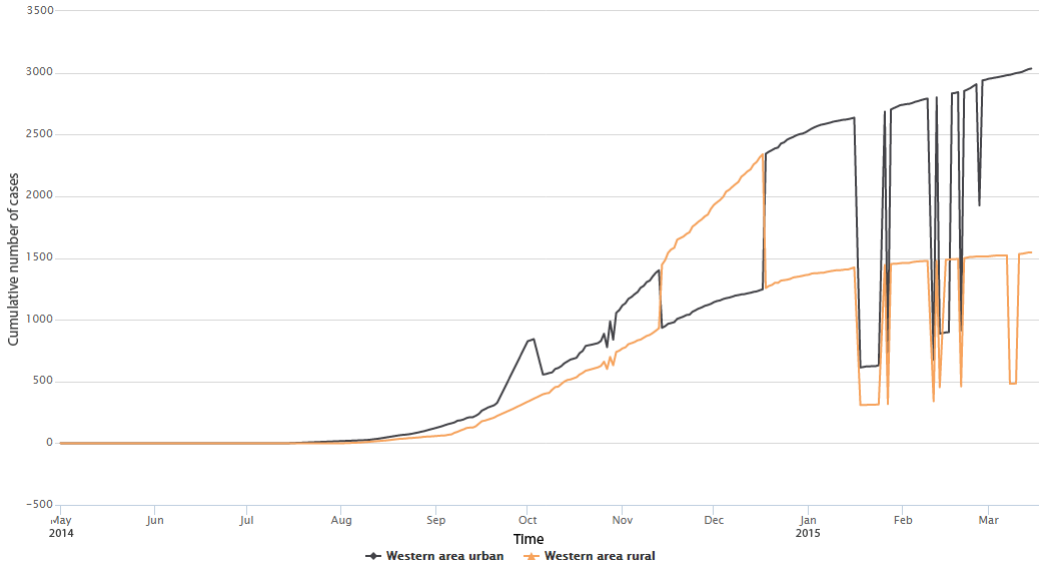
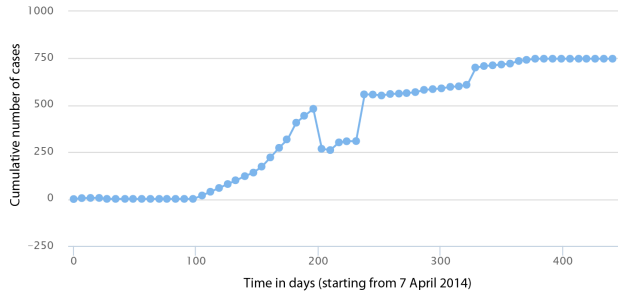


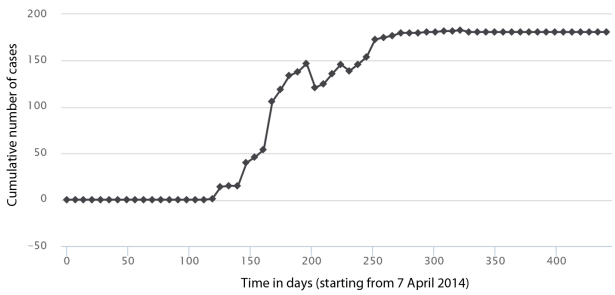
Figure 4.1: Example of lack of quality control in the collected data sets, where the case rates of two locations have been swapped between November 2014 and January 2015.

prefectures of Guinea, and the districts of Sierra Leone. The cases[1], deaths and confirmed cases are given on a weekly and location basis. We verified the data set against the WHO's weekly reports, and the data set is accurate [35].

Since the prevalence data of the British Red Cross meets our conditions, we decided to use the data set to investigate the disease dissemination. However, the data set is inconsistent because the weekly reports of the WHO consist of under- and over reporting. Under- and over reporting is done when the number of cases is lower or higher than the WHO expected on the moment of reporting. Two examples can be seen in figure 4.2, where there is a sudden decrease and increase in the data points for the locations Bong and Grand Bassa.



(a) Bong, between days 200 and 240



(b) Grand Bassa, between days 200 and 250

Figure 4.2: Example of two locations where under- and over reporting occurs in the prevalence data set

[1]Probable, suspected and confirmed cases

4.2 Sub-national indicators

There is a list of indicators[2] for the countries Guinea, Sierra Leone and Liberia. The data set consists of 26 indicators. First of all, the International Wealth Index (IWI) is given. The IWI is an index that can be used to measure a household's long-term economic status for all low and middle income countries [28]. Locations with a lower wealth index might intensify the dissemination of the disease. Furthermore, 25 other indicators are given, such as the number of households with a bad quality toilet, and the number of households with a bad quality floor. A description of each of the indicators is denoted in appendix B.

The sub-national indicators are from 2012 (for Guinea) and 2013 (for Sierra Leone and Liberia), and are therefore slightly outdated. However, these are the only indicators that are available on sub-national level for the Ebola affected countries. We do not believe the indicators changed significantly in the past years, since there have not been large changes in the environment of the countries. The data is supplied by the Global Data Lab, an organization that performs research and develops instruments for measuring and analyzing progress of societies [9].

4.3 Weather circumstances

We think that the weather circumstances could influence the physical contact within households. On the one hand, we hypothesize that if the weather is cold people tend to stay inside and therefore physical contact is higher in these houses. On the other hand, in warm weather circumstances people are more likely to sweat, a bodily fluid that could contain Ebola and therefore intensifies the chance of spread. In both hypotheses, the weather circumstances influence the way Ebola disseminates and thus is relevant for the research.

A large number of organizations supply data about the weather history. The World Weather Online API is able to provide a daily average of the weather back to July 2008 on the basis of the latitude and longitude. The daily average values consist of the precipitation, minimum temperature and maximum temperature of the day. By using geocoding, we created a data set with the corresponding latitude and longitude of a location to find the weather circumstances of a specific location.

4.4 Ebola treatment centers and units

A data set with the number of Ebola treatment centers and units with opening- and closing dates denote the accessibility towards care for local residents on a location. If the availability of these centers and units are

[2] An indicator is defined as a thing that indicates the state or level of something.

low at a location, the accessibility towards care for local residents is lower and therefore an outbreak might be more intense at the specific location. Combining the accessibility with for example the prevalence can provide us with new insights about the disease dissemination.

Data sets that are supplied on the open data initiatives are updated until the end of December 2014. Therefore, The data sets are heavily outdated and not reliable. Moreover, in most of the data sets they provide information about the health care facilities, but the information contains missing attributes, such as time and location. For example, a health care facility is stated to be closed without a closing date. The data set would be extremely relevant for the research, but does not fulfil the three conditions that we stated. The lack of quality in these data sets may critically impact the results of the analysis.

4.5 Control efforts

Control efforts are defined as the interventions that have been taken by aid organizations to contain the disease dissemination and gain control of Ebola. There are data sets available that consist of information about the control efforts that are taken on a location. However, the date of the intervention is not given in the data sets. Therefore, we have no availability to combine the control efforts with the prevalence data and observe whether the speed of transmission decreases once interventions have taken place. The lack of this information reduces the impact of the data set and can be an underlying issue in why policymakers and health officials are experiencing difficulties to contain the spread of Ebola.

4.6 Discussion

From the 55 Ebola-related data sets that were available on the HDX, only two data sets can be used to provide policymakers and health officials with adequately insights into the way the disease disseminates over time [14]. Most of the other data sets consist of (1) the same kind of information, (2) data that is not reliable, and/or (3) missing information (e.g. time and location). There are several problems that lead to the lack of quality of the data sets. We mention three of them.

(1) The origin of the disease outbreak is in West Africa. In West Africa, the health systems were already broken before Ebola struck [4]. The software created to track Ebola outbreaks was designed to let one person manually input the data into the database, which made it difficult to suddenly process all the data of different aid organizations that supplied case reports. Therefore, collecting data amid an outbreak of this intensity is a challenge.

(2) The priority of combating the disease was to respond to the outbreak, instead of focusing on the challenge of collecting data [4]. Direct actions and interventions had a higher priority than collecting data, and therefore organizations preferred investing in resources and other materials for direct impact instead of collecting data for analysis in the longer run. Even though data analysis could support and optimize in coordinating the resources and materials in an efficient matter.

(3) There already is a large amount of criticism towards the health organizations that combat the disease [11, 19]. We can imagine that the health organizations prefer to over report in a situation, because they would otherwise even receive more criticism for making wrong estimations.

Collecting data is a challenge in the West African countries. However, we have shown in section 4.1 that the inserted data also lacks of quality, which is the responsibility of the organizations themselves. The organizations underestimated the power of data science in an early stage of the disease outbreak, and now they see the consequences of not collecting data of high quality. An open data initiative could be of real power in fighting a disease such as Ebola, but only if the quality of the data sets will be improved. The power of data science could be of great value in fighting a disease such as Ebola [25].

4.7 Chapter conclusions

In this chapter we investigated research question 1: *What kind of disparate Ebola-related open data sets could provide adequate insights into the dissemination of Ebola?* We analyzed the available data sets provided by aid organizations on the open data initiatives and other sources. We searched for data sets that meet three conditions; (1) the attributes time and location, (2) relevancy, and (3) reliability.

In most of the data sets we found that (1) the attributes time or location were missing, and/or (2) the data sets were outdated and thus unreliable. Therefore, availability of data sets that meet these conditions is one of the key issues in finding data sets that provide adequate insights into the disease dissemination. We found three data sets that met the three conditions; (1) the prevalence data set, (2) the sub-national indicators, and (3) the weather circumstances. But even these data sets consisted of a lack of quality, which can be explained by the health organizations that undervalued the power of data science in an early stage of the disease outbreak. Other data sets did not meet the conditions and therefore limit the research, because the lack of quality in the data sets may critically impact the results of the analysis and therefore obscure underlying correlations.

In this chapter we focused on the collection of data sets that can be relevant for the research. However, gaining knowledge about the disease dissemination by combining the data sources can provide us with insights into fighting diseases such as Ebola. Therefore, the next chapters will investigate how we can combine the data sets to gain new knowledge about the way the disease disseminates.

Chapter 5

Data integration

Data integration is defined as combining data from multiple sources to provide an unified view of the disparate data sets [18]. This chapter discusses the integration of the data sets that are used for the research. We integrate the chosen data sets from chapter 4 on a weekly basis ranging from the 7th of April 2014 to the 22nd of June 2015 for all the possible locations: (1) the prefectures of Guinea, (2) the districts of Sierra Leone, and (3) the counties of Liberia. The first three sections (5.1 to 5.3) discuss the combination of the three data sets, with the corresponding problems that have occurred. In section 5.4 we define the target attributes for forecasting and add them to the data set. In section 5.5 we consider the final data set that is created by integrating the data sets.

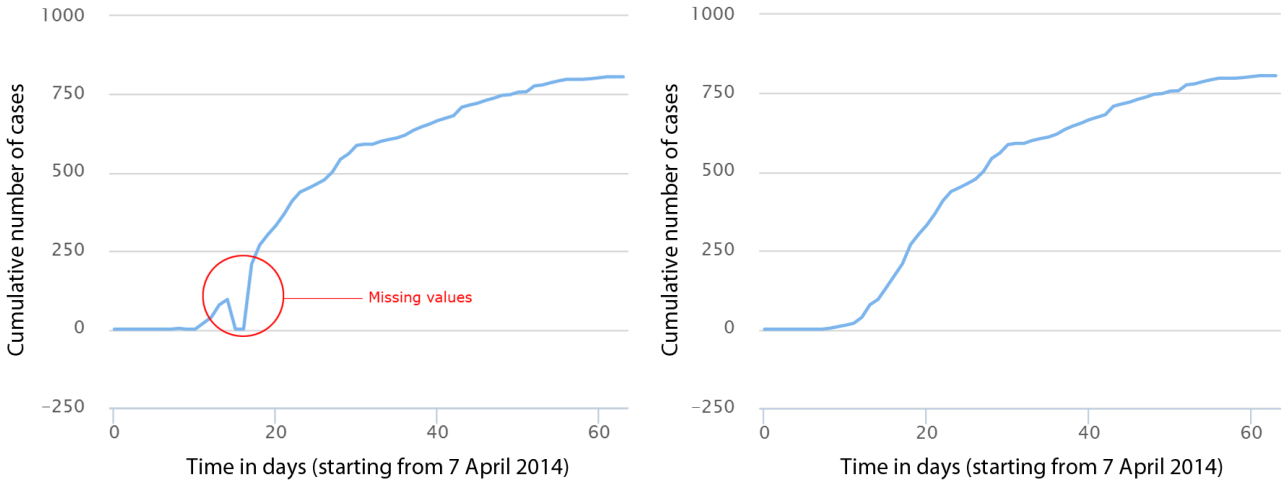
5.1 Case rates

We use the Ebola prevalence data set from Simon J. Johnson from the British Red cross. The data set provides the number of cases on a weekly basis. As we are analyzing the way the disease disseminates, we denote how infectious diseases spread over time. Infectious diseases start off slowly. If each infected person infects two other persons, the number of cases grows exponential. Once organizations start to intervene and gain control of the disease outbreak, the number of new cases slowly diminishes over time [20]. Therefore, infectious diseases spread in terms of a S-shape, also known as a sigmoid curve. Our data set shows the same kind of sigmoid curves. An example of a S-shape of the data set is given in figure 5.1b.

In figure 5.1a we notice some missing values. We interpolated the missing values using a linear equation. Mathematically, the equation is denoted as follows.

$$y = c_0 + \frac{c_1 - c_0}{t_1 - t_0}$$

Where c_0 is the cumulative number of cases that has been seen for the last time and c_1 is the cumulative number of cases that is observed in the future. t_0 is the day that the last number of cases has been seen and t_1 is the new day that the new cases are found. Figure 5.1 provides an example of the results of the linear imputation for the total number of cases. The figure shows that dates with a value of zero where cases have been seen earlier, will be imputed to a representative value. We imputed the missing values for the cumulative number of (1) cases, (2) deaths, and (3) confirmed cases.



(a) Before imputation

(b) After imputation

Figure 5.1: An example that shows the results of the imputation into the data set for the district Kenema

There are also other indicators that we can derive from the prevalence data: (1) the number of new cases of a week, since it provides an indication of the intensity of the outbreak on that specific date. (2) The prevalence of the past three weeks. The transmission time of Ebola is 2 to 21 days, so the virus can still be active in this time span [31]. (3) The number of new cases of neighbouring locations, as there is a high mobility across border in West Africa [24].

Regarding (1) and (2), we computed the incidence and number of new cases in the last three weeks by subtracting the cumulative number of cases of the observed date, from the cumulative number of cases for one or three weeks. Regarding (3), we manually created a data set with all the neighbouring locations for each location. Then, we computed a summation of the incidence data of all the neighbouring locations and inserted it in the data set.

5.2 Sub-national indicators

To integrate the sub-national indicators with the prevalence data, the sub-national indicators data set should be sufficiently location specific. The sub-national indicators data set provides information about the districts of Sierra Leone, the counties of Liberia, and the regions of Guinea. We analyze the dissemination of Ebola

on the prefectures of Guinea, and thus we have to specify the indicators of each prefecture instead of the regions. Therefore, we used reversed geocoding to derive the prefectures that correspond with the regions. The sub-national indicators of each region are then assigned to the corresponding prefectures. Afterwards, we joined the prevalence data set and the sub-national indicators data set on the location attribute.

5.3 Weather circumstances

The data set that consists of the latitude and longitude of each location that we created in section 4.3, provides the possibility to analyze the weather circumstances on a location. By using the World Weather Online API, we requested the daily (1) precipitation, (2) minimum temperature, and (3) maximum temperature. By using the World Weather Online API and the generated latitudes and longitudes data set, we requested the average weather circumstances of a week and added it to the total data set by joining the data sets on the attributes time and location.

5.4 Target attributes

The target attribute of the data set is a categorical variable that describes what kind of risk an observed location has in the upcoming week. The risk values can be (1) none, (2) low, (3) medium, and (4) high. We discretized[1] the amount of risk by observing the incidence in the upcoming week. Table 5.1 provides the assigned risk values based on the incidence for the upcoming week.

Risk	**Incidence(i)**
None	$i < 1$
Low	$1 >= i < 5$
Medium	$5 >= i < 20$
High	$i >= 20$

Table 5.1: Assigned risk values based on incidence of the upcoming week

5.5 Result

The resulting data set consists of 60 locations over 64 weeks, ranging from the 7th of April to the 22th of June. The total number of rows of the data set is 3,840. The data set consists of 38 attributes. There are 37 attributes that consist of historical data, and 1 attribute that describes the risk for the upcoming week. The 37 attributes can be analysed to observe whether they help in forecasting a risky location. A description of each attribute of the resulting data set is provided in table 5.2.

[1] Discretization is defined as converting or partitioning continuous attributes to categorical attributes.

#	Attribute	Description
1	Date	The date of the day
2	Location	The location on sub-national level
3	Cases	The cumulative number of cases
4	Confirmed cases	The cumulative number of confirmed cases
5	Deaths	The cumulative number of deaths
6	New cases	The number of new cases
7	New cases in last two weeks	The number of new cases from this day to two weeks ago
8	New cases of neighbouring countries	The number of new cases of the last two weeks in neighbouring locations
9	Precipitation	The average amount of rain, snow and other water vapours of the upcoming week
10	Maxtemp	The average maximum temperature of the upcoming week
11	Mintemp	The average minimum temperature of the upcoming week
12	IWI	The international wealth index
13	Edyr	The mean years of education of persons aged 20-49
14	Edyr_fem	The mean years of education of women aged 20-49
15	Edyr_male	The mean years of education of men aged 20-49
16	Urban	Percentage of people living in urban area or region
17	Wrk agri	Index of married men aged 20-49 working in agricultural occupation
18	Wrk inagr	Index of married men aged 20-49 working in lower non-agricultural occupation
19	Wrk unagr	Index of married men 20-49 working in an upper non-agricultural occupation
20	Electr	Index of households with electricity in the region
21	Small house	Index of households with none or one sleeping room in region
22	Large house	Index of households with three or more sleeping rooms in region
23	Qual floor	Index of households with high quality floor in region
24	Bad floor	Index of households with bad quality floor in region
25	Tap water	Index of households with piped water in region
26	Bad water	Index of households with bad quality water supply in region
27	Flush toilet	Index of households with flush toilet in region
28	Bad toilet	Index of households with bad quality or no toilet in region
29	Age 0-9	Percentage of population aged 0-9 in region
30	Age 10-19	Percentage of population aged 10-19 in region
31	Age 20-29	Percentage of population aged 20-29 in region
32	Age 30-39	Percentage of population aged 30-39 in region
33	Age 40-49	Percentage of population aged 40-49 in region
34	Age 50-59	Percentage of population aged 50-59 in region
35	Age 60-69	Percentage of population aged 60-69 in region
36	Age 70-79	Percentage of population aged 70-79 in region
37	Age 80-89	Percentage of population aged 80-89 in region
38	Risky	Categorical variable describing the risk in the upcoming week

Table 5.2: The resulting attributes of the data set

5.6 Chapter conclusions

This chapter supports to answer research question 2: *To what extent can we combine the Ebola-related data sets to gain more knowledge about the disease dissemination?* We combined the Ebola-related data sets of chapter 4 on the attributes time and/or location. We created a total data set where the three data sets are combined: (1) the prevalence data set, (2) the sub-national indicators data set, and (3) the weather circumstances. We computed and imputed the missing values in the prevalence data set and the sub-national indicators data

set. We also defined a target attribute for forecasting purposes.

By combining the data sets, we can now investigate whether we can transform the information of the data sets into knowledge. The next chapters will investigate if we can gain new insights into Ebola and the way it disseminates.

Chapter 6

Public health triangulation

Public Health Triangulation is defined as combining and interpreting data from multiple sources to empower and support decision making [27]. This chapter discusses two of the insights into the disease dissemination that we found by combining and analyzing the data sets. In section 6.1 we discuss how we measure the speed of dissemination, which we can compare against the sub-national indicators to observe whether there exists a correlation between the speed of dissemination and the sub-national indicators. In section 6.2 we compute the correlations between the sub-national indicators and (1) the total number of cases of a location, and (2) the speed of dissemination. In section 6.3 we discuss our results. Lastly, in section 6.4 we provide a provisional answer to research question 2.

6.1 Speed of Ebola dissemination

A method to compute the speed of dissemination of Ebola using prevalence data is by computing the reproduction number [31]. However, it is extremely difficult to determine the contact rate and transmission rate without intensive control [17]. Therefore, we thought of our own method. By determining the take-off date and stabilization date, we can measure the growth per day over the period of take-off and stabilization. The take-off date is the date where the incidence is consistently higher than zero. The stabilization date is the date where the incidence consistently decays and slowly tends to become zero. We can describe the speed of dissemination by measuring the growth per day in the worst time of the epidemic period of a specific location. We computed both values using a different method.

We are not able to compute the take-off date by observing on what moment the first case occurs, because sometimes only one case occurs and is immediately dealt with. The data is also inconsistent since we deal with under- and over reporting. Therefore, we averaged the last five weeks of new cases seen from the week

that we observe. Afterwards we look forward three weeks to determine whether the growth is consistently higher than the average of the last five weeks. If this is the case, we decide this is a take-off date for the disease outbreak at a specific location.

For measuring the stabilization date we cannot use the same approach in a reversed matter, because if the disease stabilizes, the incidence can still consistently grow over time. Therefore, we computed the stabilization date by taking the maximum growth of incidence in a week over the entire period and afterwards analyzing whether a lower number than 40 % of the maximum growth occurs four times.

Figure 6.1 shows the moment of take-off and stabilization for the district Bo, and we can see that the estimations are representative for a take-off and stabilization date. The number of cases does keep growing after the stabilization date, but we measure the speed of dissemination on the worst period of the disease outbreak. However, because of inconsistencies in the data set we could not always compute the correct take-off. Six locations did not meet the requirements to observe a take-off date. Therefore, we updated the take-off date manually by observing on what moment the incidence tends to grow consistently at the locations Gueckedou, Dabola, Gbarpolu, Kouroussa, Grand Gedeh and Pita.

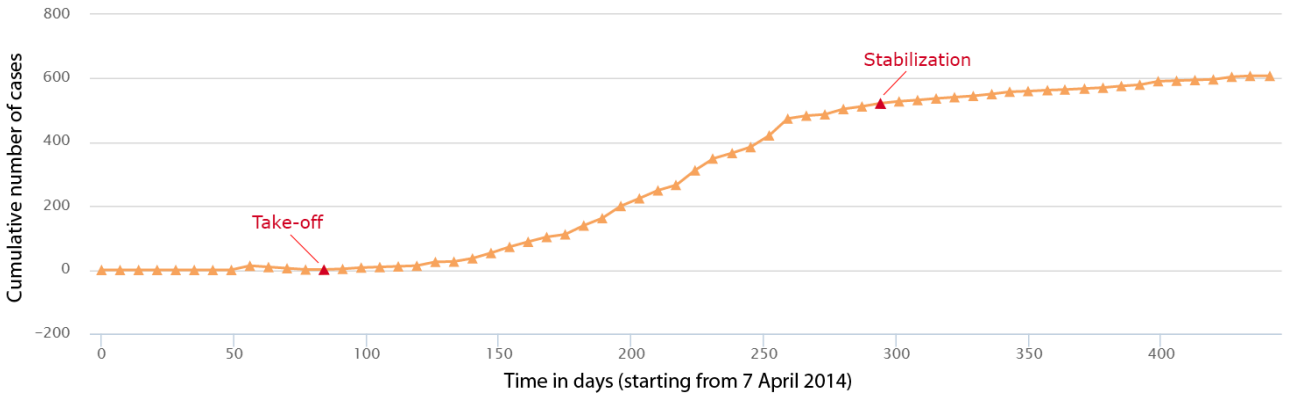


Figure 6.1: Example of the take-off date and stabilization date measured by our approach for the district Bo

Afterwards, we calculate the speed of dissemination(y) as follows:

$$y = \frac{c_1 - c_0}{t_1 - t_1}$$

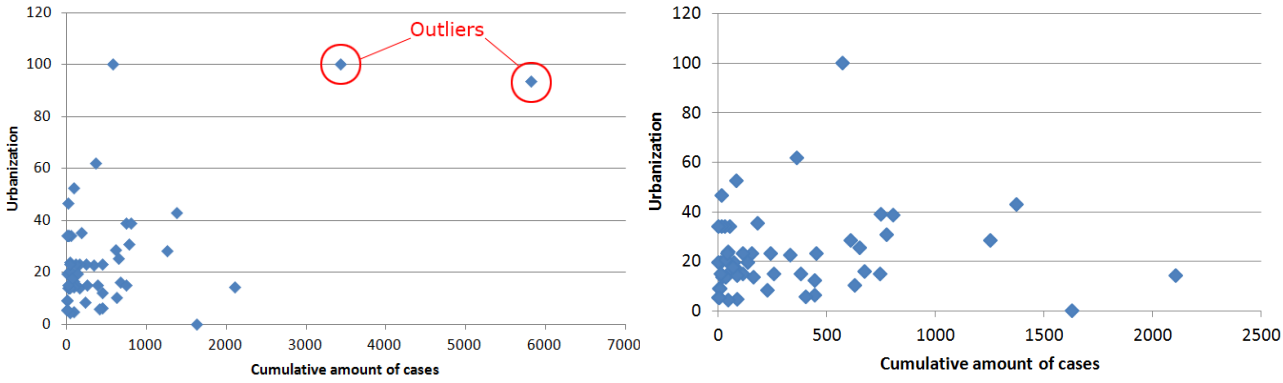
Where c is the number of cases on a certain time stamp (date of take-off (0) and date of stabilization (1)) and t is the date in days for stabilization and take-off. By using this method, it computes the same speed of dissemination for locations where both the number of cases and the duration of the outbreak is twice as low.

6.2 Sub-national indicators

Searching for a correlation between the sub-national indicators and indicators of the dissemination of Ebola provides us with insight into the disease outbreak. We analyzed (1) the Pearson correlation between the

sub-national indicators and the total number of cases, and (2) the Pearson correlation between the speed of dissemination and the sub-national indicators.

We found that four sub-national indicators had a correlation of above 0.5, meaning that they have a strong association with the indicators of the Ebola dissemination. However, Figure 6.2a shows that the indicators of the Ebola dissemination of the locations Montserrado and Western Area Urban are outliers in the observed data. As the Pearson correlation is sensitive to outliers, we removed these points from the data set. The right figure shows the data set after removal of these points, and the correlations decreased to a value lower than 0.21, meaning there is no strong association between the indicators. Table 6.1 demonstrates the indicators that had a correlation of above 0.5, but after removing the outliers had a weak correlation of lower than 0.3. We observe in figure 6.2b that the data points are randomly separated and there is not a linear association to be found. A total list of the correlations can be found in appendix C.



(a) With outliers

(b) Without outliers

Figure 6.2: Scatterplots of the urbanization index with the cumulative number of cases with and without outliers

Indicator	Correlation with total number of cases	Correlation with speed of dissemination
Edyr	0.188	0.204
Edyr_fem	0.216	0.231
Urban	0.109	0.188
Wrk_unagr	0.215	0.242

Table 6.1: Correlation of four of the sub-national indicators with the speed of dissemination and the total number of cases

6.3 Discussion

Measuring the correlation of the sub-national indicators with indicators about the disease outbreak showed not to have a strong association with the disease outbreak. However, we do believe that some of the sub-national indicators can influence the way the disease disseminates. We mention two of them: (1) the urbanization index and (2) the IWI.

Regarding (1), this is the first time the disease broke out in more urbanized countries in comparison to earlier

outbreaks, and it struck harder than ever before [6]. Physical contact is inevitable in cities these days, making the chance of an infected person to transmit the disease higher than in rural areas. We therefore hypothesize that the urbanization index can influence the way the disease disseminates.

Regarding (2), we hypothesize that wealth can influence the Ebola dissemination in two ways. On the one hand, bad wealth circumstances can lead to a faster spread of Ebola because citizens have less accessibility towards health care. On the other hand, good wealth circumstances may lead to highly infrastructured environments that could relate to our hypothesis of (1).

One of the current limitations is that the sub-national indicators and prevalence data is not sufficiently location specific. On locations with a low urbanization rate, the disease could still have been broken out in a city, explaining the high number of cases on that location. These kinds of situations make it hard to conclude whether the sub-national indicators have impact on the disease outbreak. However, we can see that investigating the disease outbreak on sub-national level could not provide us with more insights into the disease outbreak.

6.4 Chapter conclusions

This chapter provides a provisional answer to research question 2: *To what extent can we combine the Ebola-related data sets to gain more knowledge about the disease dissemination?* We measured the speed of dissemination to compare against the sub-national indicators. Then, we observed whether the sub-national indicators correlate with (1) the total number of cases and (2) the speed of dissemination. The sub-national indicators that we compared did not have a strong association. However, we do believe that some of the sub-national indicators influence the way the disease disseminates, if the data is sufficiently location specific.

Therefore, our provisional answer to research question 2 reads as follows. At the moment, we are not able to gain more knowledge about the disease dissemination by triangulating the data sets. This is explained by the fact that the data has to be more location specific to investigate whether indicators such as the urbanization index or IWI influence the way the disease disseminates in the geographical areas.

Chapter 7

Forecasting risk

This chapter discusses our approach to forecast risky locations. In section 7.1 we discuss how we train the machine learning algorithms and test the forecasts. In section 7.2 we discuss how we evaluate and compare the forecasting models. In section 7.3 we provide a baseline algorithm that we will use to compare against three other forecasting algorithms: (1) C4.5, (2) Naive Bayes, and (3) a mathematical model. Section 7.4 discusses the C4.5 and Naive Bayes algorithms of Weka that we used and their results. Section 7.5 considers an alternative approach with the use of a mathematical model to forecast perilous locations. In section 7.6 we investigate if we can improve accuracy by smoothing the data. In section 7.7 we discuss the algorithms and analyse what algorithm scored best by measuring performance of the test sets. Finally, section 7.8 provides an answer to research question 2 and research question 3.

7.1 Data preprocessing

Since Weka does not support time series analysis, we removed the date attribute from the data set and analyzed situations on other attributes. To train the forecasting models and test our forecast, we created twelve training sets and twelve test sets. Each training set consists out of the historical data until the first of each month, and each test set consists of the risk of the locations in the upcoming week from the week after the last month of the training set. For example, if our training set consists of the historical data towards the first of July, our test set consists of the risk of the 60 locations in the week between the first and seventh of July. We created twelve training sets and twelve test sets, ranging from June 2014 to May 2015. The risk distributions can be found in appendix D.

7.2 Data evaluation

In order to evaluate our forecasting methods, we used the confusion matrix to derive the overall forecasting accuracies for each of the algorithms that is used. The confusion matrix maps each of the actual values and forecasted values in a matrix. The confusion matrix defines four types of values; (1) true positives (TP), the values that are correctly identified, (2) true negatives (TN), the values that are correctly rejected, (3) false positives (FP), the values that are incorrectly identified, and (4) false negatives (FN), the values that are incorrectly rejected.

By investigating the confusion matrix and computing three indicators we can investigate the forecasting accuracies of the models. We measure (1) the precision, that indicates how many times we correctly identify the risk if we forecast the risk, (2) the recall, that indicates how many times we correctly identify the risk in comparison to the total amount of times the risk occurs, and (3) the F-measure, that combines the precision and recall. Mathematically, we denote the indicators as follows.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F - measure = \frac{2*Precision*Recall}{Precision+Recall}$$

7.3 Baseline algorithm

To analyse whether the data integration of chapter 5 helps in forecasting the future risk, we need a baseline algorithm to compare to the algorithms that take the additional indicators into account. An usual baseline algorithm for machine learning is the ZeroR algorithm, which predicts the target attribute that occurs most often. The ZeroR algorithm has an accuracy of 54.9%, which forecasted all the test cases to be not risky. For infectious diseases the prevalence data is known for a specific location. Therefore, policymakers and health officials currently analyse the prevalence data to support decision making and to decide what locations are currently in the highest need of resources and other materials. Since policymakers and health officials analyse the prevalence data, we used the OneR algorithm as baseline algorithm. It splits on only one attribute that predicts the risk with the least errors in it [13]. The OneR algorithm splits on the incidence of the last two weeks, making it the most important attribute to forecast situations. With a minimum bucket size of 6, table 7.1 provides the forecasting results of the baseline model.

The OneR algorithm beats the ZeroR algorithm with an accuracy of 69.7%. The precision and recall of the model demonstrate that the biggest mistakes are made in forecasting the low and medium risk locations. This could partially be explained by the inconsistencies of the prevalence data. A second reason is that the

number of no risk training examples is of a larger size than the number of low risk examples, and therefore sometimes low risk cases are classified as no risk cases. We also observe that forecasting high risk cases scores relatively high in comparison to forecasting low and medium risk cases. This can be explained by the fact that if a high risk location is observed there is a high chance of observing a high risk location again.

Risk	None	Low	Medium	High		
None	384	19	10	1		414
Low	88	20	25	1		134
Medium	26	12	42	18		98
High	4	3	11	56		74
	502	54	88	76		720

Risk	**Precision**	**Recall**	**F-measure**
None	0.765	0.928	0.886
Low	0.370	0.149	0.212
Medium	0.477	0.429	0.452
High	0.737	0.757	0.747

Table 7.1: Forecasting results of the baseline algorithm

7.4 Forecasting by Weka

We used Weka's feature selection to investigate whether the sub-national indicators of section 4.2 and weather circumstances of section 4.3 possibly influence the risk on a specific location. Several evaluators in combination with their corresponding search methods have been used to investigate whether the indicators are of any value. The evaluators measure the importance of attributes by using algorithms and by ranking them. The evaluators that we used (InfoGain, GainRatio, CfsSubset, ChisquaredAttribute) showed that the sub-national indicators and weather forecasts are of low value to forecasting the risk. We also verified this by measuring the correlation towards the speed of dissemination in section 6.2. Forecasting without the indicators gave better results in testing the algorithms. Therefore, we removed the indicators from the data set.

Feature extraction

Since the sub-national indicators and weather circumstances did not have a strong association with the target attribute, we extracted other features out of the prevalence data. We are working with time series data, so we extracted new features that could help in forecasting risky locations by using time. We came up with the following features; (1) slope, the weekly growth from the time of take-off (from 6.1) until the day we are looking at. If the slope is high, there might be a higher chance of risk. (2) Reproduction, the growth of the number of cases in comparison to the last week (computed by dividing the current incidence with the incidence of the week before), and (3) the days since the disease took off, also based on section 6.1.

By using the same evaluators to measure the value for forecasting risky locations, we found that these indicators have a better association with the risk compared to the total number of cases, number of new neighbouring cases and previously analyzed indicators.

Algorithms

We tested two algorithms to forecast perilous situations; (1) C4.5, a classifier that builds a decision tree by searching for the attribute with the highest information gain using the information entropy. It iteratively splits on the attribute with the highest information gain until all examples are covered [26]. (2) Naive Bayes, a probabilistic classifier that is based on the Bayes theorem, which describes the probability that a certain event occurs on the basis of related other events [15]. After testing and debugging the data with parameters chosen in Weka, we found that the C4.5 algorithm scored best on the test sets. We optimized the parameters of the C4.5 algorithm (J48 in Weka). By changing the confidence factor to 0.05, and the minimum number of instances per leaf to 10, we forecasted most accurately. The forecasting results of the C4.5 algorithms can be found in table 7.2.

With an accuracy of 71%, the C4.5 algorithm scored slightly better than the OneR algorithm. The decision tree of the C4.5 algorithm varied with each training set. It used the following attributes to forecast risk; the number of new cases in the last week, the number of new cases in the last three weeks, the slope, the neighbouring cases, the days since take-off, and the cumulative number of cases. Like the OneR algorithm, the F-measure of forecasting low and medium risk locations is relatively low compared to the no and high risk forecast. The reason is explained in section 7.3.

Risk	None	Low	Medium	High		
None	370	31	12	1		414
Low	58	36	37	3		134
Medium	16	18	45	19		98
High	2	1	11	60		74
	446	86	105	85		720

Risk	**Precision**	**Recall**	**F-measure**
None	0.830	0.894	0.861
Low	0.419	0.269	0.328
Medium	0.429	0.459	0.443
High	0.706	0.811	0.755

Table 7.2: Forecasting results of the C4.5 algorithm

7.5 Forecasting with the IDEA model

Adding multiple indicators forecasted slightly better than the baseline algorithm. Fisman et. al. accurately fitted data in an early stage of the disease outbreak with the use of the IDEA model to forecast the future incidence [8]. If we use the historical data to fit the IDEA curve until the day that is known, we can extrapolate the data to observe what the risk will be in the upcoming week.

Curve fitting

The IDEA curve is bell-shaped. We expected the incidence data to be normal distributed as well, but as we demonstrated in figure 4.2, the data is inconsistent because of over- and under reporting. The equation to fit the data on the IDEA model is as follows:

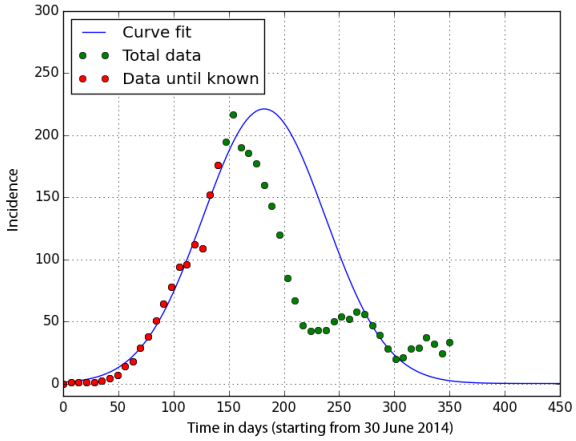
$$I_t = (\frac{R_0}{(1+d)^t})^t$$

Where R_0 is the reproduction number, d is the discount factor that explains the control efforts against the disease or any other dynamic changes that slow the disease dissemination. Furthermore, t is the time in days and I_t is the incidence given on a certain time t.

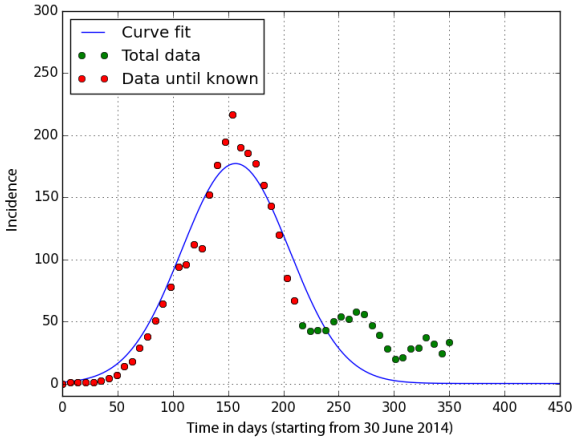
On the basis of earlier results of the IDEA model we made an initial guess of $R_0 = 1.89$ and $d = 0.0009$ [8]. The model will try to fit the data by using this equation and initial guess. The model best fits the data if it looks from the point of take-off, as it will otherwise overestimate in forecasting risky locations in an early and late phase of the disease outbreak, because the curve will try to fit on the entire prevalence data. Therefore, we observe whether there are more than zero cases seen in the data. If the incidence is higher than zero in the data, we use this moment as starting point for curve-fitting. Otherwise, it will fit the data on the zero point and it will forecast no risk. This also occurs if the disease stabilized to a zero point and afterwards new cases occur. At these moments, we re-fit the curve from that moment to better forecast the upcoming risk. Some examples of the curve fitting problems and solutions are given in figure 7.1 and 7.2. Lastly, sometimes it is hard for the program to find the optimal amount of parameters to fit the curve. Increasing the amount of calls to the curve fit function solves the problem and fits almost all data points that are available.

Forecast

We defined the risk of the upcoming week by using the same discretization values of table 5.1. Then, we checked whether the forecasted risk corresponds with the observed risk of that week. Table 7.3 shows the forecasting results by using the IDEA model. With an accuracy of 65.6% the model scores worse than the C4.5 and OneR algorithm. One of the limitations of this method is that it cannot deal with weird inconsistencies

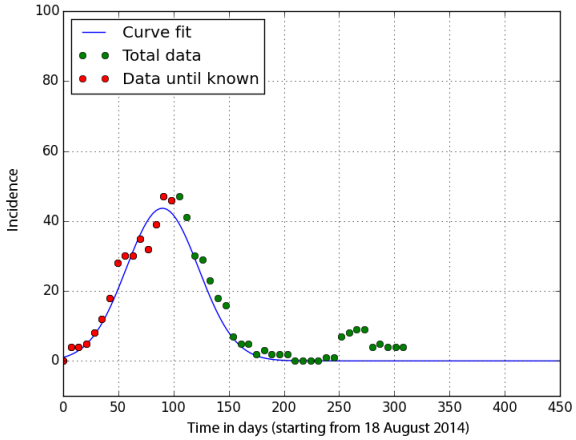


(a) Example where curve fits forecast

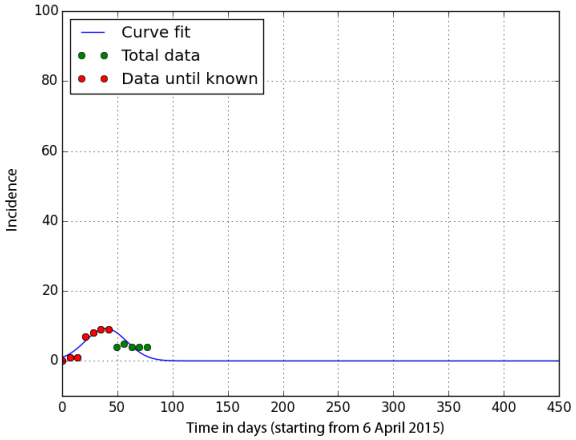


(b) Example with overestimation

Figure 7.1: Example of curve fitting of the mathematical model with overestimation



(a) Example where curve fit happens on total incidence



(b) Example where model re-fits the data from day 230

Figure 7.2: Example of re-fitting the forecasting curve because the disease stabilized in an earlier stage

in the data (e.g. a sudden unexplained drop of 15 cases for 4 weeks), while the OneR and Weka algorithm adapt by only analyzing the current situation instead of the entire historical data. Moreover, the model tends to forecast no risk when there is a low risk. We can explain this by the fact that the curve stabilizes earlier than the data does, and that the model has to deal with inconsistencies in the data.

7.6 Improvements

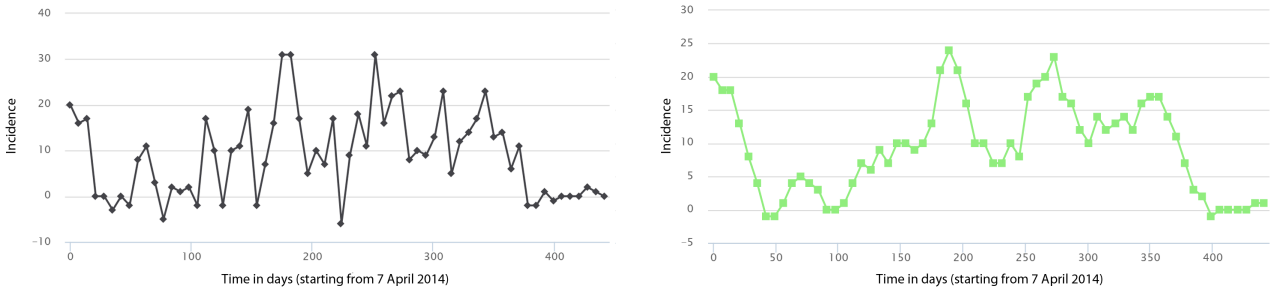
The prevalence data has a large number of inconsistencies in it. If we smooth the data, we may improve accuracy of the forecasting models. Therefore, we smoothed the data by taking the average of the incidence of the last four weeks. By smoothing the data, we can deal better with inconsistencies in the data set. Figure 7.3 demonstrates the result of smoothing the data, and shows that a decrease in the amount of inconsistencies.

Risk	None	Low	Medium	High		
None	320	55	17	22		414
Low	38	53	23	18		132
Medium	16	18	34	30		98
High	2	1	7	64		74
	376	127	81	134		718

Risk	Precision	Recall	F-measure
None	0.851	0.773	0.809
Low	0.417	0.402	0.409
Medium	0.420	0.347	0.380
High	0.478	0.865	0.616

Table 7.3: Forecasting results of the mathematical model

Next, we defined the risk by the same discretization factor as we did in section 7.5, but by using the average number cases of the last four weeks for the upcoming week.



(a) Before smoothing

(b) After smoothing

Figure 7.3: Incidence of Ebola for Conakry before and after smoothing

By smoothing the data, the OneR algorithm and C4.5 algorithm both had a slightly higher accuracy. The C4.5 algorithm forecasted with an accuracy of 71.1%, and the OneR algorithm forecasted with an accuracy of 69%.

However, the accuracy of the mathematical model grew significantly. The accuracy of the model using the smoothed data is 75.5%. The model scores higher than the other algorithms, because at first it was limited towards inconsistent data. Smoothing the data decreases inconsistency, therefore increasing accuracy. The model could even be more accurate, if the model also re-fits data if we signalise a new outbreak in an earlier stage of stabilization on the same location. The current model only re-fits the IDEA model after the disease is completely stabilized, while in some cases the disease comes back in an earlier stage of stabilization. The results of the improved model are shown in table 7.4

Risk	None	Low	Medium	High		
None	361	8	2	0		371
Low	71	77	9	4		161
Medium	24	20	42	16		102
High	4	3	13	55		75
	460	108	66	75		709

Risk	Precision	Recall	F-measure
None	0.785	0.973	0.869
Low	0.713	0.478	0.572
Medium	0.637	0.412	0.500
High	0.733	0.733	0.733

Table 7.4: Forecasting results of the improved mathematical model

7.7 Discussion

The OneR, C4.5 and IDEA algorithm scored well in comparison to the ZeroR algorithm. By using the prevalence data supplied by the Simon J. Johnson, the C4.5 algorithm scored best. All algorithms had a bad F-measure on the locations with a low and medium risk, because of inconsistencies in the data set. Moreover, low and medium risk locations are harder to forecast because they describe a change towards the take-off or stabilization of the outbreak.

The algorithms used in Weka and the mathematical model both have limitations. The OneR and C4.5 algorithm both forecast given a certain situation. They both investigate the indicator that is of most importance, and based on the importance the algorithms will make the decision and forecast. In addition, the C4.5 algorithm also uses other indicators that could be of any value and splits the data in larger decision tree. Deciding by analyzing a certain situation could be powerful, but the data could be in an entirely other phase of the disease outbreak(e.g. splitting on the new cases when the disease takes off could give an other forecast than when the disease almost stabilizes). Therefore, analyzing time series data could be of better value. The IDEA model has problems in forecasting because the quality of data is low. Inconsistencies occur in the prevalence data, because under- and over reporting occurs in the incidence data. In these situations, the decision trees adapt easier than the mathematical model, since they analyze on a specific situation instead of the entire historical data.

If we smooth the incidence data, the accuracy of the forecast by using the mathematical model is higher than all the other methods. In this way, fitting a curve in the historical data to forecast the risk can be done better because the data is less inconsistent. However, by smoothing the line we can only forecast the future average cases of the last four weeks. The improvement of accuracy proves that data of better quality provides the possibility to improve forecasting perilous locations. We cannot use this model to forecast risky locations in the current situation, because the average incidence could be dependent on the incidence that was seen four weeks ago. It does provide us with the knowledge that cleaner data will improve in forecasting risky locations. Therefore, it can provide insights into a way to act more preventive towards an outbreak that is as vigorous as the Ebola outbreak of 2014.

7.8 Chapter conclusions

This chapter partially provides an answer to research question 2: *To what extent can we combine the Ebola-related data sets to gain more knowledge about the disease dissemination?*. This chapter also provides an answer to research question 3: *To what extent can we forecast perilous locations with the use of the Ebola-related datasets?*

Answer to research question 2

We used multiple Weka evaluators to check whether (1) the weather forecasts help in forecasting the disease outbreak, and (2) whether the sub-national indicators help in forecasting risky locations. None of the indicators helped in forecasting future risk.

Therefore, our answer to research question 2 reads as follows. Based on the results of section 6.2 and section 7.4 we can conclude that the combination of the Ebola-related data sets cannot provide us with more information about the way the disease disseminates. However, if the data would be more location specific, we could analyze the disease outbreak more specifically and decide whether the indicators could or could not provide us with new knowledge about the way the disease disseminates.

Answer to research question 3

We defined a baseline algorithm to forecast perilous locations by using the OneR algorithm. Then, we used three forecasting models to forecast risky Ebola locations and compare them to the OneR algorithm: (1) C4.5, (2) Naive Bayes, and (3) a mathematical model. Both the sub-national indicators and weather circumstances did not help in forecasting a risky location, and therefore we extracted features out of the time series data. In all models, the most difficult part was forecasting whether there is a low risk, since all forecasting models tend to forecast no risk when there actually is a low risk. This is explained by (1) the small number of low risk examples in comparison to high number of no risk examples to train the model and (2) the inconsistencies in the data.

Hence, our answer to research question 3 reads as follows. We are able to forecast perilous situations by using machine learning algorithms with a certainty of 71%. The Ebola-related data sets that we combined did not add any value in forecasting risky locations. Extracting features out of the time series data provided us with better results. Moreover, mathematical modeling did not forecast better than the machine learning algorithms, but the mathematical model adapts less easy towards inconsistencies in the data. Smoothing the data showed that the accuracy of the mathematical model can improve significantly if the data is of higher

quality and thus can help in providing insight into a way to act more preventive towards an Ebola outbreak on a specific location.

Chapter 8

Conclusions

In this chapter we provide an answer to the three research questions and the problem statement formulated in chapter 1. In section 8.1 we provide answers to the research questions. In section 8.2 we answer the problem statement. Finally, we discuss future work in section 8.3.

8.1 Answers to the research questions

Research question 1: *What kind of disparate Ebola-related open data sets could provide adequate insights into the dissemination of Ebola?*

The answer of the first research question is derived from chapter 4. We searched for data sets provided by aid organizations on the open data initiatives that meet three conditions; (1) the data set has to consist of the attributes time and location, (2) the data set has to be relevant for the research, and (3) the data set has to be reliable. Although most of the data sets met the relevancy condition, they were lacking the attributes time and location, and/or were not up to date and thus unreliable. We were able to find three data sets that provide information about the dissemination of Ebola; a prevalence data set, (2) the sub-national indicators data set, and (3) the weather circumstances data set. A data set about the Ebola treatment centers and units, and a data set about the interventions taken by organizations were not available. These limit the research because they may critically impact the results of the analysis and obscure underlying correlations.

Research question 2: *To what extent can we combine the Ebola-related data sets to gain more knowledge about the disease dissemination?*

The answer of the second research question is derived from chapter 6 and chapter 7. The 25 sub-national indicators that we combined on the attribute location did not have a strong association with (1) the speed

of dissemination of the disease and (2) the total number of cases. Both the forecasting models (see section 7.4) and the measurement of correlation (see section 6.2) verified that we could not use the sub-national indicators to gain more insight into the disease outbreak. Moreover, the weather circumstances did not have a strong association with forecasting perilous locations. However, we do believe that some of the indicators can influence the way the disease disseminates, such as the urbanization index and the International Wealth Index. To investigate whether these indicators influence the disease dissemination, the location has to be sufficiently location specific. Therefore, we may conclude that the data has to be more location specific to be of any value in investigating the disease outbreak and to gain insights into the disease dissemination.

Research question 3: *To what extent can we forecast perilous Ebola locations with the use of the Ebola-related data sets?*

The answer of research question 3 is derived from chapter 7. We are able to forecast perilous locations by using machine learning algorithms with a certainty of 71%, meaning that we forecasted better than forecasting based on only the prevalence data with the OneR algorithm. The key aspect in improving forecasting accuracy was to extract features out of the prevalence data based on time. However, by smoothing the data and using a mathematical model we showed that we could forecast risky locations with a higher accuracy, if the data would be of higher quality. The lack of quality is explained by under- and over reporting done by the aid organizations that supply the data. Therefore, we may conclude that we are able to forecast risky situations, but our accuracy can be improved significantly if the data is of higher quality.

8.2 Answers to the problem statement

Problem statement: *To what extent is it possible to help policymakers and health officials in reacting effectively to an outbreak of Ebola that is as vigorous as the Ebola outbreak of 2014?*

In this thesis we investigated whether we can provide insights into the Ebola dissemination by analyzing and combining data sets. Our aim was to improve the combat against diseases such as Ebola in the field of data science. We investigated (1) what kind of data sets were available and useful for the research, (2) how we could use these data sets to gain new insights, and (3) if we are able to forecast perilous locations by using these data sets.

From our investigations, we may conclude that the collection of data sets is a big issue in the field of Ebola. There is a lack of quality in each of the data sets, limiting the research. We may also conclude that the data that is provided should be on a more specific location level than on sub-national level to gain more insights into the way the disease disseminates. Moreover, we may conclude that we could provide insight into a way to act more preventive towards the outbreak by forecasting upcoming perilous situations by analyzing the

prevalence data. However, the forecast can only be used as an indication since it is not perfectly correct.

Our answer to the problem statement reads as follows. If we would like to help policymakers and health officials in reacting effectively to an Ebola outbreak in the field of data science, we should start with recommending them to not underestimate the power of data science. It is time for the public health organizations to make a larger step towards the data century and improve the way to collect and supply data. Then, more specific investigations could lead to new insights into the disease dissemination and data science could become of real power. The power of data science could maybe not prevent an outbreak, but it could definitely help in analyzing the outbreak and even prevent it from getting so intense as the Ebola outbreak of 2014.

8.3 Future work

This section provides recommendations to further analyze the disease dissemination and thus support health officials and policymakers in decision making. We have five recommendations for future work.

(1) Data quality is of critical importance to investigate the disease outbreak. In our approach we imputed the missing values in the data set to improve the quality of the data sets. However, analyzing methods to signalize under- and over reporting could help in improving the quality of the data sets. We have shown that the improvement of data quality could increase the accuracy of forecasting risky situations. Therefore, searching for methods to clean the data can significantly help in investigating the disease outbreak and forecasting perilous situations.

(2) The IDEA model has shown to fit the prevalence data and was able to forecast perilous situations. However, the current model is limited to only re-fit the model if we observe that the disease is completely stabilized and another outbreak occurs on the same location. However, signalizing another outbreak if the disease is close to stabilizing could therefore improve the accuracy of the model.

(3) Forecasting risk by using our mathematical model only uses the prevalence data set. Accuracy may improve by adding additional indicators to the model and by analyzing what kind of indicators could be of importance to improve the model.

(4) We analyzed the forecasting possibilities by discretizing the incidence for the upcoming week. Regression models can also be used to analyze the forecasting accuracy and give an indication of the upcoming risk. Evaluating and optimizing by using regression models can provide more specific information about the risk for the upcoming week.

(5) We combined the data sets and analyzed whether the sub-national indicators were associated with the way the disease disseminates. Triangulating on more indicators could help in investigating the disease outbreak.

For instance, investigating whether the total number of cases at a location correlates with the way the disease decays over time on the specific location, may provide us with new knowledge about the disease outbreak.

Besides these recommendations, one of the most important aspects to improve future work in the field of Ebola and data science is by improving the data collection methods and the way it is supplied. The organizations should supply data of higher quality, and the data should be more location specific. Then, at least two data sets can provide us with interesting insights: (1) a data set about the interventions taken by organizations to control the outbreak, and (2) a data set about the Ebola treatment centers and units. These data sets can explain underlying correlations and therefore be of great importance to improve the combat against diseases such as Ebola.

References

[1] Althaus, C. L. (2014). Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. *PLOS Current Outbreaks*, 1.

[2] Batty, A., Fishkind, A., and Moyela, C. (2014). *IBM launches Humanitarian intitiatives to Help Contain Ebola Outbreak in Africa*. IBM Press Room.

[3] Centers for Disease Control and Prevention (2015) (2015 (accessed July 8, 2015)). *Ebola Hemmorrhagic Fever - CDC*. `http://www.cdc.gov/vhf/ebola/`.

[4] Cheng, M. and Dilorenzo, S. (2014). In ebola outbreak, bad data adds another problem. *The Big Story - Associated Press*.

[5] Denzin, N. (2006). *Sociological Methods:* A Sourcebook. Aldine Transaction, 5 edition.

[6] Dissel, J. T., Wychgel, W., and Timen, A. (2014). Ebola - Hoe is Nederland voorbereid? (in dutch). *Magazine nationale veiligheid en crisisbeheersing*, 5:29–31.

[7] Drake, J. M., Kaul, R. B., Alexander, L. W., O'Regan, S. M., Kramer, A. M., Pulliam, J. T., Ferarri, M. J., and Park, A. W. (2015 Jan. 13). Ebola Cases and Health System Demand in Liberia. *PLOS Biology*, 1(1).

[8] Fisman D., Khoo E., T. A. (2014 Sep. 8). Early epidemic dynamics of the west african 2014 ebola outbreak: Estimates derived with a simple two-parameter model. *PLOS Current Outbreaks*, 1(1).

[9] Global Data Lab (2015 (accessed July 15, 2015)). *Global Data Lab - GDL*. `http://www.globaldatalab.nl`.

[10] Google (2015 (accessed August 8, 2015)). *The Google Geocoding API*. `https://developers.google.com/maps/documentation/geocoding/intro`.

[11] Gostin, L. O. and Friedman, E. A. (2014). Ebola: a crisis in global leadership. *The Lancet*, 384:1323–1325.

[12] HealthMap (2015 (accessed July 10, 2015)). *Ebola Map Virus & Contagious Disease Surveillance*. `http://www.healthmap.org/ebola/#timeline`.

[13] Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90.

[14] Humanitarian Data Exchange (2015) ((accessed July 8, 2015)). *Ebola Crisis Page.* `https://data.hdx.rwlabs.org/ebola`.

[15] John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[16] Johnson, S. J., British Red Cross(2015) (2015 (accessed July 8, 2015)). *Ebola dashboards*. `http://simonbjohnson.github.io/`.

[17] Kretzschmar, M. and Wallinga, J. (2010). Mathematical models in infectious disease epidemiology. In Krmer, A., Kretzschmar, M., and Krickeberg, K., editors, *Modern Infectious Disease Epidemiology*, Statistics for Biology and Health, pages 209–221. Springer New York.

[18] Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA. ACM.

[19] Liu, J. (2015). Disease outbreak: Finish the fight against ebola. *Nature*, 524:27–29.

[20] Macdonald, T. H. (2007). *Basic concepts in statistics & epidemiology*. Radcliffe Publishing Ltd, Abingdon, United Kingdom.

[21] Minakshi, Vohra, R., and Gimpy (2014). Missing value imputation in multi attribute data set. *International Journal of Computer Science and Information Technologies*, 5(4):5315–5321.

[22] NuCivic (2015) (2015 (accessed July 8, 2015)). *Ebola Open Data Jam.* `https://www.eboladata.org`.

[23] Philips, M. and Markham, A. (2014). Ebola: a failure of international collective action. *The Lancet*, 384.

[24] Piot, P., Muyembe, J., and Edmunds, W. (2014). Ebola in west africa: from disease outbreak to humanitarian crisis. *The Lancet*, 14:1034–1035.

[25] Plaat, A. (2015). Inaugural lecture: Data science and ebola. *Leiden University*.

[26] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[27] Rutherford, G., McFarland, W., Spindler, H., White, K., Patel, S., Aberle-Grasse, J., Sabin, K., Smith, N., Tache, S., Calleja-Garcia, J., and Stoneburner, R. (2010). Public health triangulation: approach and application to synthesizing data to understand national and local hiv epidemics. *BMC Public Health*, 10(1):447.

[28] Smits, J. and Steendijk, R. (2015). The international wealth index (iwi). *Social Indicators Research*, 122(1):65–85.

[29] The Republic of Sierra Leone (2015) ((accessed July 8, 2015)). *Open Government Initiative.* http://www.ogi.gov.sl/.

[30] Weka (2015) (2015 (accessed July 8, 2015)). *Weka 3 - Data mining with open source machine learning software in Java.* http://www.cs.waikato.ac.nz/ml/weka/.

[31] WHO Ebola Response Team (2014). Ebola virus disease in west africa the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, 371(16):1481–1495. PMID: 25244186.

[32] World Health Organization (2014a). *Case definition recommendations for Ebola or Marburg Virus Diseases*. Technical report.

[33] World Health Organization (2014b). *HR Procedures concerning public health emergencies of international concern (PHEIC)*. Technical report.

[34] World Health Organization (2014c). *Statement on the 1st meeting of the IHR Emergency Committee on the 2014 Ebola outbreak in West Africa*. Technical report.

[35] World Health Organization (2015). *Ebola Situation Report*. Technical report.

Appendix A

Map of Ebola affected countries

The investigations of the research are in the geographical areas. We analyze (1) The prefectures of Guinea, (2) the districts of Sierra Leone, and (3) the counties of Liberia. The figure below provides a map of the locations on sub-national level.

Figure A.1: The geographical areas on sub-national level of Guinea, Liberia, and Sierra Leone

Appendix B

The sub-national indicators

The sub-national indicators data set that is supplied by the Global Data Lab consists of indicators of the Ebola affected countries. Below a table is provided with the indicators and their corresponding description.

#	**Indicator**	**Description**
1	IWI	The international wealth index
2	Edyr	The mean years of education of persons aged 20-49
3	Edyr_fem	The mean years of education of women aged 20-49
4	Edyr_male	The mean years of education of men aged 20-49
5	Urban	Percentage of people living in urban area or region
6	Wrk agri	Index of married men aged 20-49 working in agricultural occupation
7	Wrk inagr	Index of married men aged 20-49 working in lower non-agricultural occupation
8	Wrk unagr	Index of married men 20-49 working in an upper non-agricultural occupation
9	Electr	Index of households with electricity in the region
10	Small house	Index of households with none or one sleeping room in region
11	Large house	Index of households with three or more sleeping rooms in region
12	Qual floor	Index of households with high quality floor in region
13	Bad floor	Index of households with bad quality floor in region
14	Tap water	Index of households with piped water in region
15	Bad water	Index of households with bad quality water supply in region
16	Flush toilet	Index of households with flush toilet in region
17	Bad toilet	Index of households with bad quality or no toilet in region
18	Age 0-9	Percentage of population aged 0-9 in region
19	Age 10-19	Percentage of population aged 10-19 in region
20	Age 20-29	Percentage of population aged 20-29 in region
21	Age 30-39	Percentage of population aged 30-39 in region
22	Age 40-49	Percentage of population aged 40-49 in region
23	Age 50-59	Percentage of population aged 50-59 in region
24	Age 60-69	Percentage of population aged 60-69 in region
25	Age 70-79	Percentage of population aged 70-79 in region
26	Age 80-89	Percentage of population aged 80-89 in region

Table B.1: Description of the sub-national indicators of the Ebola affected countries

Appendix C

Correlations with the sub-national indicators

The tables below show the correlation of the sub-national indicators with (1) the number of total cases and (2) the speed of dissemination before and after removal of the outliers.

Indicator	Correlation with total amount of cases	Correlation with speed of dissemination
IWI	0.282	0.263
Edyr	0.558	0.553
Edyr_fem	0.645	0.635
Edyr_male	0.494	0.492
Urban	0.579	0.569
Wrk agri	-0.423	-0.410
Wrk inagr	0.363	0.344
Wrk unagr	0.502	0.512
Electr	0.220	0.212
Small house	0.370	0.352
Large house	-0.330	-0.331
Qual floor	0.472	0.455
Bad floor	-0.389	-0.366
Tap water	0.040	0.305
Bad water	-0.035	-0.037
Flush toilet	-0.136	-0.128
Bad toilet	-0.120	-0.169
Age 0-9	-0.110	-0.131
Age 10-19	-0.032	0.005
Age 20-29	0.061	0.086
Age 30-39	0.143	0.122
Age 40-49	0.001	-0.026
Age 50-59	-0.041	-0.061
Age 60-69	0.016	0.014
Age 70-79	0.049	0.046
Age 80-89	-0.025	-0.388

Table C.1: Correlation of the sub-national indicators with the speed of dissemination and the total number of cases before removal of outliers

Indicator	Correlation with total amount of cases	Correlation with speed of dissemination
IWI	-0.097	-0.094
Edyr	0.188	0.204
Edyr_fem	0.216	0.231
Edyr_male	0.165	0.180
Urban	0.109	0.188
Wrk agri	0.079	0.058
Wrk inagr	-0.148	-0.132
Wrk unagr	0.215	0.242
Electr	-0.075	-0.071
Small house	-0.041	-0.051
Large house	-0.056	-0.050
Qual floor	0.126	0.015
Bad floor	-0.133	-0.148
Tap water	0.054	0.054
Bad water	0.086	0.085
Flush toilet	-0.077	-0.076
Bad toilet	-0.253	-0.258
Age 0-9	-0.098	-0.13
Age 10-19	0.083	0.122
Age 20-29	0.129	0.154
Age 30-39	0.096	0.066
Age 40-49	-0.136	-0.155
Age 50-59	-0.175	-0.179
Age 60-69	-0.017	-0.016
Age 70-79	0.014	0.015
Age 80-89	-0.078	0.015

Table C.2: Correlation of the sub-national indicators with the speed of dissemination and the total number of cases after removal of outliers

Appendix D

Training and test sets

The distribution of the training sets and test sets for analyzing and building the machine learning algorithms in chapter 7 are given below.

Risk	Training set 1	Training set 2	Training set 3	Training set 4	Training set 5	Training set 6
None	438	591	871	1011	1163	1265
Low	24	52	85	122	164	219
Medium	17	27	42	72	121	156
High	1	10	22	55	112	160

Training set 7	Training set 8	Training set 9	Training set 10	Training set 11	Training set 12
1357	1480	1593	1717	1889	2049
274	351	414	467	529	577
208	263	300	337	386	412
201	246	273	299	316	322

(a) Risk distribution of the training sets without smoothing

Risk	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Test set 6
None	50	48	37	34	27	22
Low	7	8	11	7	13	12
Medium	2	2	6	8	8	14
High	1	2	6	11	12	12

Test set 7	Test set 8	Test set 9	Test set 10	Test set 11	Test set 12
24	26	29	36	34	47
15	15	15	8	15	8
13	12	9	12	8	4
8	7	7	4	3	1

(b) Risk distribution of the test sets without smoothing

Table D.1: Risk distribution of the training sets and test sets without smoothing

Risk	Training set 1	Training set 2	Training set 3	Training set 4	Training set 5	Training set 6
None	426	679	868	1004	1142	1244
Low	36	64	93	142	204	255
Medium	15	30	41	75	119	156
High	2	7	18	39	95	145

Training set 7	Training set 8	Training set 9	Training set 10	Training set 11	Training set 12
1320	1419	1524	1636	1794	1919
321	409	472	542	613	673
206	273	320	354	404	452
193	239	264	288	309	316

(a) Risk distribution of the training sets with smoothing

Risk	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Test set 6
None	49	51	38	29	25	23
Low	7	5	13	14	13	14
Medium	4	2	5	9	9	10
High	0	2	4	9	13	13

Test set 7	Test set 8	Test set 9	Test set 10	Test set 11	Test set 12
16	22	27	33	31	36
22	17	18	11	14	15
12	13	8	10	13	7
10	8	7	6	2	2

(b) Risk distribution of the test sets with smoothing

Table D.2: Risk distribution of the training sets and test sets with smoothing

Appendix E

Results of the forecasting models

We have shown the results of the most important forecasting models in chapter 7. This appendix shows the results of the forecasting models that we also used to analyze the forecasting possibilities, but did not score considerably better than the other algorithms.