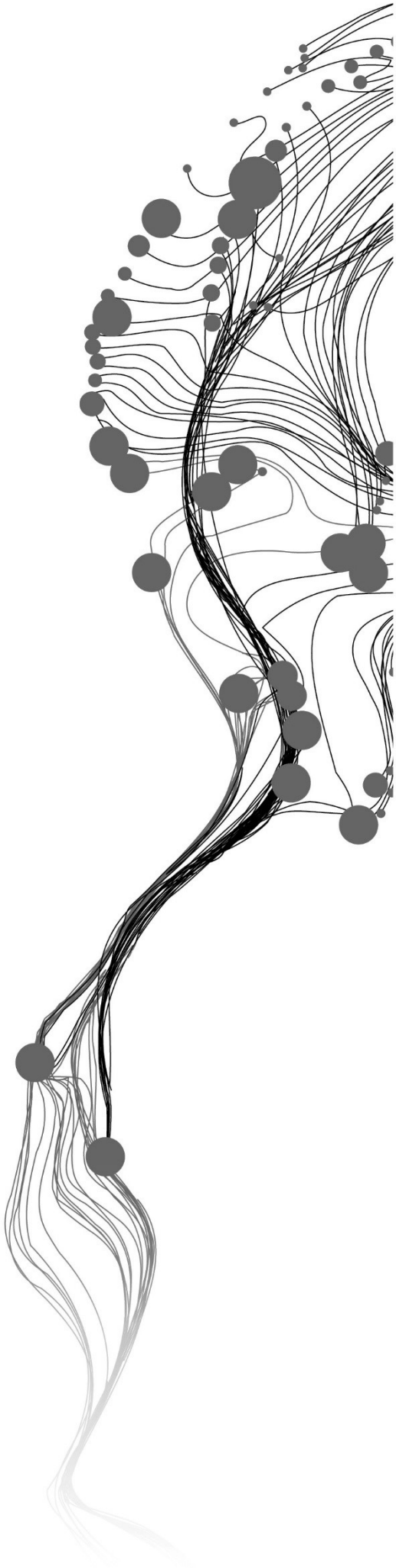
HUMAN MOBILITY AND PERTUSSIS DISEASE DIFFUSION PATTERNS

BIRTUKAN FIKADIE MESFIN
February, 2016

SUPERVISORS:
Ir. P.W.M. Ellen-Wien Augustijn
MSc. P. Pasha Zadeh Monajjemi



HUMAN MOBILITY AND PERTUSSIS DISEASE DIFFUSION PATTERNS

BIRTUKAN FIKADIE MESFIN
Enschede, The Netherlands, February, 2016

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geo-informatics

SUPERVISORS:
Ir. P.W.M. Ellen-Wien Augustijn
MSc. P. Pasha Zadeh Monajjemi

THESIS ASSESSMENT BOARD:
Prof. Dr. M.J. Kraak (Chair)
Dr. S. Amer (External Examiner, University of Twente, ITC-PGM)

ABSTRACT

Emerging and re-emerging of infectious diseases have had a significant impact on the well-being of humans. Studying the spread of these infectious diseases is important for better interventions. It is known that mobility plays an important role in the spread of diseases because when people move they can introduce an infection in new, previously uninfected locations. Mmobility scale varies from long distance air travels to short distance commuting and it is unknown at what level of mobility scale is relevant to explain epidemic behaviour. Besides, modelling human mobility in relation to disease spread is challenging because of two reasons: conceptual complexity and data limitation.

A solution for the data limitation problem is to use different mobility proxies such as road network data. Studying the relationship between mobility and disease diffusion at different spatial-temporal scales can help to determine the relevance of commuting data as indicators for spreading of infectious diseases.

In this study, we assessed the diffusion pattern of the infectious disease pertussis in the Netherlands for three epidemic years (1996, 1999 and 2001). Human mobility was modelled by identifying high commuting areas using different mobility proxies such as the road and railway network. Based on these mobility proxies, municipalities were grouped into urban zones and series of highly connected cities. The diffusion pattern of pertussis for the thee epidemic years was analysed using these high commuting zones and comparing the disease patterns of municipalities in these zones and between these zones with patterns of municipalities outside the zones.

The disease diffusion pattern was measured in two different diffusion patterns: hierarchical diffusion and synchrony. Hierarchical diffusion was assessed using the number and duration of fadeout (extinction of a disease) in relation to the population size. Hierarchical diffusion between urban zones was shown for both the number of fadeouts and duration of fadeouts. But for within urban zones, only duration of fadeouts shows a hierarchy of diffusion and there was no significant difference in hierarchical diffusion between the three epidemic years.

Synchrony was measured at two different scales using the frequency (number of disease cases per 100000 inhabitants) of infection. We found that measuring synchrony between urban zone scale is more robust than within urban zones scale. This is not surprising as urban zones consist of a combination of municipalities with lower and higher population sizes and all urban zones have a population size above the critical community size CCS above which the disease can maintain itself. In addition, synchrony was measured using municipalities connected by railway network and intercity tracks (lines) finding a higher level of synchrony during the year 1996. During the years 1999 and 2001 no synchrony of diffusion was found and further studies are required to confirm why these two epidemic years were not showing synchrony of diffusion. Overall, this study indicates that a multi-scale analysis of disease diffusion using different types of mobility data is important for gain a better understanding of the diffusion of an infectious disease for different epidemic years.

Eventually we assessed the general diffusion pattern of the disease in urban (high commuting) versus non-urban (low commuting) areas using the timing of the peak of the epidemic year. We found that the disease reached the peak in urban areas one month sooner than in the non-urban areas. Nevertheless, this difference in the peak timing of urban versus non-urban is not significant.

Key words: *Infectious diseases, Mobility, Spatial-Temporal diffusion patterns, Pertussis*

ACKNOWLEDGEMENTS

First, I would like to thank GOD. He has given me his divine grace, protect and hold me with his merciful hands. All these would not have been possible without his support and blessings. Lord, you are the leader of my life thank you!

I am deeply indebted to my first supervisor Ir. P.W.M. Ellen-Wien Augustijn for he unlimited support and guidance throughout the progress of this thesis. Specially, her immediate and detailed feedbacks on every of my work, her dedication to assist me, patience and knowledge she had kindly share with me. I do not have a special word to show my thankfulness to her. THANK VERY MUCH. Also I would like to extend my special thanks to my second supervisor Ms MSc. P. Pasha Zadeh Monajjemi for her contractive comments and suggestions.

I would like to thank the Netherlands Fellowship Program, NFP for giving me this greater opportunity to build up my career and financial support. I also would like to thank University of Twente, ITC for selecting me to attain this quality education.

I also would like to thank all my friends here at ITC and at back home for your accompany in study and life. My special thanks go to Mehreteab Yohannes who always support me during this difficult time at ITC. I also would like to extend my special thanks to Saron Araya Wubie. Sari you are hilarious, I was surprized, and we had the same feeling and the same thinking all the time konjiye.

Most importantly, I would like to express my deepest love and special thanks to my beloved parents, brothers and sisters. All my successes are based on support and, encouragement and prayer of my family. Specially, my elder sister Meazi, thank you for being a role model for my life. Meazye thank you so much you are the reason I am here. I have also special thanks to my brothers Woretaw and Muluken for their support and guidance in my life.

Lastly, I dedicated this thesis to my hubby Melaku Wondie. Babye thank you my darling for being always there for me when I am in need for you, for your patience and sacrifices. You deserve all the words for what you did. We had different hardships which we faced during my stay here and thank you for understanding me while I was away. I ALWAYS LOVE YOU WUDIE.

Birtukan Fikadie
Enschede
The Netherlands

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation and problem statement

Emerging and re-emerging of infectious diseases have had a significant impact on the well-being of humans. Due to the rampant growth of populations and the continuous movement of humans, the danger of infectious diseases diffusion is rising (Mei et al., 2015) and keeps on being a major public health challenge in both developed and developing countries (Bell et al., 2013). Infectious diseases may emerge locally in a certain place and can diffuse to neighbouring regions through different factors. The spatial distribution of hosts, their mobility behaviour and interactions are among the factors that facilitate the spread of infectious diseases (Poletto et al, 2012).

Modelling of spatial-temporal diffusion patterns of infectious diseases is important to gain insight into their diffusion dynamics and to apply better treatment of infected individuals and control of disease outbreak. However, as discussed by Merler & Ajelli (2010), in countries with high variability in socio-demographic structure and mobility, spatial-temporal diffusion patterns of infectious diseases are not easily detected. In such type of countries there is a need to consider human mobility patterns, and it is clearly necessary to model the location of hosts and their mobility to analyse patterns of diffusion (Riley et al.,2014).

Modelling human mobility in relation to disease spread is challenging because of conceptual complexity and data limitation (Balcan et al., 2009). The commuting network structure of links connecting settlements display a complex structure and are difficult to model due to different orders of magnitude in number of locations and number of travellers between each origin and destination(Chowel et al., 2003; Barrat, 2004; Brockmann et al., 2006). The varying multi-scale nature of mobility and especially the interrelations among the multiple scales of human mobility is challenging. Currently it is unknown at what level of resolution of epidemic behaviour a given mobility scale is relevant (Balcan & Vespignani, 2011). Another problem is the difference in mobility patterns between age groups (adults show different mobility patterns compared to children). Charaudeau et al. (2014) discussed the impacts of changes among age groups in epidemic spread and they suggest considering the age differences during analysis of mobility patterns.

Availability of mobility data varies per country. Data may be available at national or regional aggregation level, lacking the necessary level of detail. Mobility data used for disease spread analysis can be split into real data, proxies of real data and modelled data. Real data are data based on measured mobility between spatial locations (settlements). Proxies of real data are data collected from official census surveys, statistical organizations or by using online data collection tools that are used as a measure of real mobility data where real data are missing. Modelled mobility data are simulated data which are generated by using models to produce mobility networks. Using real data on human mobility has uncovered the effects of various heterogeneities that characterize human movement patterns (Poletto et al., 2013). For instance, the use of real data from commuting networks add extra challenge of complexity related to high level predictability and recurrence (Balcan & Vespignani, 2011). In addition, due to the high degree of fluctuation in mobility, theoretical analysis of epidemic spread in a heterogeneous network is a difficult task (Gong et al., 2014). Because fully up-to-date data are hard to get and some research requires less complex data, researchers have used simulated data and human mobility proxies. One of the examples of

the use of proxies is Gonzalez et al.(2008) who used mobile phone data, measuring the distance between user's position at consecutive calls as a measure for human mobility. In another example, Vazquez-Prokopec et al.(2013) used GPS data points of mobility tracks of individuals to measure age-specific mobility parameter to design the dynamics of influenza infectious disease. For the simulation of mobility data, either radiation (Tizzoni et al., 2014) or gravity models (Ortúzar & Willumsen, 2011) can be used. Simulating mobility data has some problems including the required input data for calibration and parameter fitting and the assumption of "universal" commuting behaviour at a given scale (Tizzoni et al., 2014).

Spatial-temporal spread of communicable disease differs per disease and requires a good understanding of the type of disease and its diffusion process. Most studies distinguish three different types of spatial-temporal diffusion patterns at three different scales: travelling waves, synchrony and hierarchical diffusion (Viboud et al., 2006). These three types of diffusion patterns can be observed at different spatial scales. Travelling waves are observable at a world or country scale, synchrony can be observed at a country to regional scale, and hierarchical spread at a regional to local scale. Pertussis is a world-wide infectious disease that can be studied at all of the mentioned scales. Travelling waves of Pertussis were observed by Choisy & Rohani (2012). The patterns of synchrony of pertussis were analysed by Rohani (1999), hierarchical patterns of pertussis were also detected by Broutin et al. (2004). Researchers also concentrate on different aspects of the diffusion. For travelling waves some studies focus on the speed of (spatial) propagation and the direction of propagation. Synchronized infection of dispersed locations deals with time of infection (frequency at the same time). Hierarchical spreads most of the time capture number of fadeouts or duration of fadeouts. These different aspects of the disease diffusion are visible at different scales and can be related to different types of human mobility in different scales.

To build a good model of the spatial-temporal spread of infectious disease, there is a need to understand the relationship between the potential driving variable and the disease diffusion pattern. One of the potential disease diffusion driver tools is mobility and there is a need to understand the link between mobility and disease diffusion. Analysing the relationship between these will improve the quality of existing disease models and will help to overcome some of the limitations encountered in previous models (Balcan et al., 2009).

In this context, considering the complexity of human mobility patterns and findings from previous studies, this study is motivated to experiment with different human mobility variables related to disease spread in a country by investigating disease frequency, number of fadeouts, duration of fadeouts and timing of peak cases at multiple spatial-temporal scales.

1.2. Research Identification

Studying the disease spread is important for better interventions. Mobility plays an important role in disease diffusion, but this role is not completely understood yet. Therefore, further research is needed. The following sub-sections discuss the proposed objectives and the innovation of the current study in relation to the existing literature in this domain.

1.2.1. Research Objectives

The main objective of this research is to compare different mobility variables at different spatial-temporal scales to determine their relevance as indicators for spreading of infectious diseases, using a case study of Pertussis in the Netherlands.

To achieve this main objective, the following sub-objectives will be addressed.

- To identify and compare different mobility indicators as measures (proxies) for aggregating areas and identifying high commuting zones/lines.
- To detect and quantify (analyse) the correlation between pertussis disease data and these identified zones/lines.
- To check if the observed relationships between can be identified in different epidemic years.

1.2.2. Research Questions

1. How to aggregate areas into highly commuting zones/lines using different mobility indicators?
   1.1. What are the mobility indicator variables that can be used as proxies for identifying zones/lines with high commuting?
   1.2. Which mobility indicator variable is more appropriate for identifying zones/lines in the context of disease diffusion at a certain spatial-temporal scale?
2. How to determine the relationship between disease diffusion and highly commuting zone/lines?
   2.1. What type of disease variable can be used to study the disease in different zones/commuting lines?
   2.2. Is there a hierarchy of disease diffusion and if there is, how does it vary between more and less connected zones?
   2.3. Is there a synchrony of disease diffusion and if there is, how does it vary between more connected and less connected zones?
3. Do different epidemic years show similar diffusion pattern on the identified zones/lines or different?
4. What is the relationship between the disease diffusion pattern of urban and non-urban areas?

1.2.3. Innovation

Although the impact of human mobility on disease diffusion has been analysed previously, this research is unique because it evaluates (compares) different mobility indicator variables at multiple spatial-temporal scales in a country with a very high level of spatial complexity. Most existing models select one mobility variable without a critical evaluation of multiple scales.

Innovative in this work is also that it includes three aspects of disease diffusion, timing of the peak, number (and duration) of fadeouts and frequency of incidence; whereas most existing studies only use frequency. Contrary to many of the existing models, it does not use a simulation model, but applies a number of readily available (statistical) analytical approaches.

1.2.4. Related Work

Several studies have investigated the association between infectious disease diffusion and human mobility. For instance, Merler and Ajelli (2010) developed an individual-based epidemic simulation model that considers human mobility and population heterogeneity and quantified their effects on impacts and timing of a highly infectious influenza pandemic. Balcan et al. (2010) present global epidemic and mobility (GLEaM) model based on meta-population approach that integrates the disease dynamics, population and airline transportation system and simulates the diffusion of epidemics at the worldwide scale. Similarly, Frias-Martinez et al. (2011) designed an agent-based system to model mobility patterns and an epidemic spread of N1H1 using individuals' mobility and their phone call records.

Moreover, researchers have used analytical (statistical) models to analyse human mobility and disease spatial-temporal patterns. Ge et al. (2015) analysed the association between Tuberculosis (TB) incidence

and regional transport networks by using Geographical Information System (GIS) and K-function statistics. They also study environmental variables such as population density and elevation in combination with transportation networks to predict the TB occurrence in space. They found that TB occurrence is not only associated with the transportation network of the regions, but also with differences in elevation. Their study provides an indication for considering transport network connection for effective treatment and control of TB diffusion. Charaudeau et al. (2014) performed spatial auto-correlation analysis (Moran's I) to extract typical paths of spread and evaluate spatial auto-correlation of influenza-like illness data in France to analyse the correlation with commuting network. Wu et al. (2014) used time-series regression analysis to analyse the impacts of mass vaccination campaign against pandemic H1N1 2009 influenza in Taiwan. They assess the effectiveness of the mass influenza vaccination and the impact of the prioritization program among people at different levels of risk using a multiple linear regression model. The daily number of patients is used as a proxy variable for H1N1 activity at hospital level and the impacts of age groups, establishment of flu clinics, and average daily immunity against influenza were analysed. They conclude that, mass vaccination of influenza was an effective control measure, and priority should be given to school-aged children than older adults during an influenza pandemic.

Although the relation of human mobility patterns with infectious disease diffusion is modelled by researchers, in most studies the pattern of diffusion of the epidemic behaviour of a given mobility scale is selected randomly without any comparison.

1.3. Project set-up

1.3.1. Method adopted

The proposed research was carried out by following the steps listed below to achieve the specified objectives and answer research questions:

- Literature review: The research work was started with reviewing relevant literatures to select and justify the methods, to identify mobility indicator variables.in order to choose statistical (analytical) analysis methods.
- Initial data exploration of disease data: An initial exploration of the disease data was performed to examine important properties of the data. The exploration was performed based on different perspectives: using number of cases per epidemic year, the relationship between disease cases and population, and spatial distribution of disease cases.
- Data preparation and aggregation: Both disease and mobility data were prepared for consistency, aggregation and scale.
- Comparison of mobility variables: Comparison of mobility data were performed on three different methods: Here we compared percolation method, urban agglomerations and commuting distance methods to select mobility data to define urban zone.
- Correlation and pattern analysis: Correlation and pattern analysis were used to understand the relationship between the selected mobility data and disease frequency, total number of fadeouts, duration of fadeouts and the timing of the peak case. These analyses were performed at different scales to study the patterns of diffusion at different scales.
- Comparison of diffusion patterns of disease with the effects of different spatial scales were applied for three epidemic years.

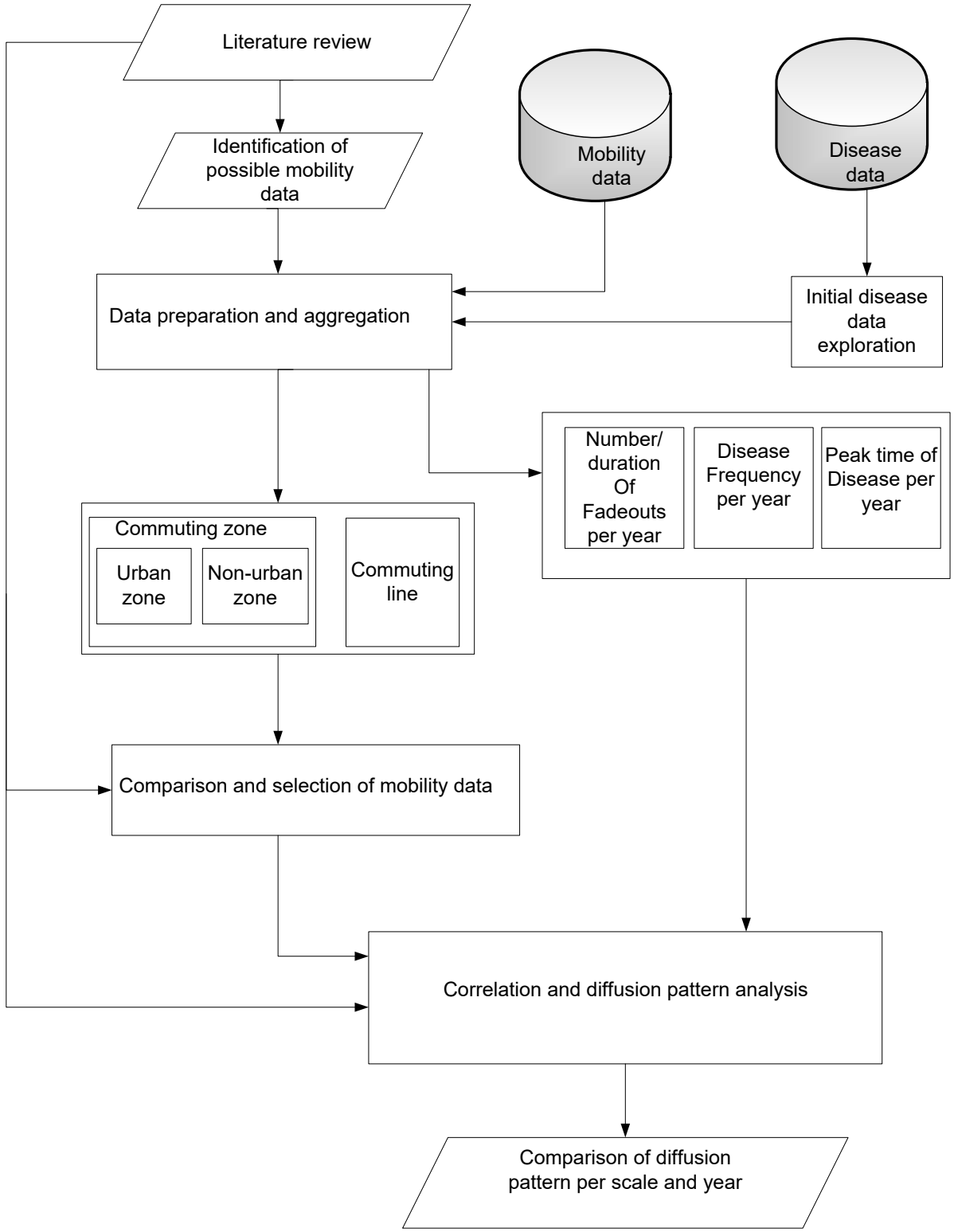


Figure 1.1: Methods adopted for addressing the research objective

2. LITERATURE REVIEW

This section describes the preliminary concepts needed to be understood about characteristics of disease time series, human mobility measuring methods and infectious disease diffusion patterns.

2.1. Characteristics of disease time series

2.1.1. Epidemic curve

An epidemic curve is a plot of time trends in the occurrence of a disease for a defined population (Wallinga & Teunis, 2004). Epidemic curve can be used to: determine the duration of an epidemic, help to determine the source of infection, used to propose hypothesis on the nature of disease and its mode of transmission. We can learn a lot about an outbreak from an epidemic curve such as: identify the earliest case, the peak of the epidemic, the outbreak's time trend, the distribution of the case over time, and determine the likely period of exposure that led to the outbreak. An epidemic curve offers a great deal of information and better illustrated changes in the frequency of outbreaks over time. For instance, for pertussis, an epidemic curve showed an infectious period of 14 to 21 days (Rohani, 1999). As investigated by (de Melker et al., 1997) epidemic curve of pertussis also characterized by double peaks.

2.1.2. Fadeouts

Fadeout is defined as the average time that a patient takes to recover from infection counted to the moment in which the spread counted (Broutin et al., 2007). Meaning that a period with no new disease cases are reports that there was an infection previously observed. Lloyd-Smith and Cross (2005) differentiate two types of fadeouts based on their starting conditions: epidemic fadeout refers to the extinction of disease after a major outbreak reduces the available number of susceptible and it occurs when a small number of infectives are introduced in to a totally susceptible population (Cullen, 2003). Whereas, endemic fadeout describes the extinction of disease from a relatively stable state and it occurs for populations of less than a critical community size and irrespective of the number of susceptible present (Cullen, 2003). For pertussis a fadeout was defined as a period of at least three consecutive weeks without any cases (Grenfell, 1997). The number of consecutive weeks without disease cases corresponds to the duration of fadeouts.

2.1.3. Peak of disease cases

A peak of a disease is defined as a point which is the maximum number of disease case was observed and timing of the peak is the time at which these maximum cases were reported. Timing of the peak case can be used as an indicator for the analysis of disease diffusion patterns. For instance, Broutin et al. (2004) used dates of the peak number of cases observed to analyse urban versus rural pattern of pertussis transmission in a small region of Senegal.

2.1.4. Critical community size

Critical community size (CCS) is the minimum population size below which any given infection cannot persist itself without external inputs (Broutin et al., 2004). Critical community size is traditionally set by subjective assessment or arbitrarily chosen criteria and is often used as a general term for all population thresholds for disease persistence (Lloyd-Smith & Cross, 2005). CCS can be determined by using population size in relation to disease fadeout. For Pertussis, 250,000-400,000 people are set as a CCS to allow the persistence of the disease (Grenfell, 1997). CCS is used to analyse the persistence of disease diffusion in a certain region. Broutin et al. (2004) suggested that the population size has a profound

implication in the persistence of pertussis disease diffusion. The main point here is that there must be large enough number of population size to maintain a chain of transmissions.

2.2. Spatial-Temporal disease diffusion patterns

The spatial-temporal diffusion of communicable diseases differs per disease and requires a good understanding of the type of disease and its diffusion process. Thus, it is crucial to study disease diffusion at different spatial-temporal scales. Most studies distinguish two different types of spatial-temporal diffusion patterns at two different scales: synchrony can be observed in a country to regional scale, and hierarchical spread at a regional to local scale. The following subsections explain each of the two disease diffusion patterns.

2.2.1. Hierarchical diffusion

Hierarchical diffusion is the spread of disease from its sources to new locations following a decreasing sequence of classes or places based on their size (Viboud et al., 2006). For example, in socially structured populations the source of the disease may originate in the upper level of the social hierarchy and then trickle down to the lower levels (Cliff et al., 1981). In such type of diffusion, disease lasts long within originated area but spreads out to a new location at a later period.

For instance, Broutin et al. (2004) showed that the hierarchically diffusion of pertussis is from larger population size to smaller population size sequentially. Figure 2.1 shows the hierarchical diffusion of an infectious disease. The diffusion may originate from a larger city or urbanized region (origin) and spread to smaller cities in different parts of the country (spatially detached). Hierarchical diffusion is characterised by: long duration of infection in a large city, short duration of infection in small settlements, frequent fadeouts in medium settlements, few fadeouts in large or small settlements and order of first infection in order of city size.

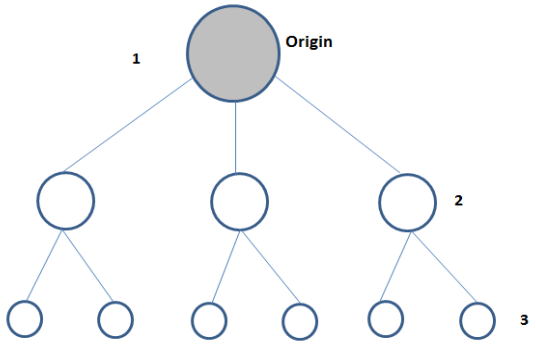


Figure 2.1: Hierarchical diffusion adopted from (Cliff et al., 1981)

Hierarchical diffusion can be influenced by movement of people carrying a disease and geographical distance they travelled (Cliff & Haggett, 2004). Because of this, hierarchical diffusion is characterized by the display of cascading diffusion and drop down at longer distance. Cascading diffusion means the diffusion of process assumed to be downwards from larger cities to smaller cities to villages (in the order of population size) (Cliff et al., 1981). Therefore, population size is the key factor and places closer to the source of the disease source are most likely to have higher disease cases.

Although human activities can be similar in all regions, there are regions with a large population that function as a centre of activity for a certain region (Sattenspiel, 2009). These centre regions cause the diffusion of disease to diffuse hierarchically from the centre region to its surrounding and it fadeouts when the number of the population is too small to maintain the disease.

Several researchers have determined the hierarchical diffusion of disease and identify the diffusion of disease is from these centre cities to smaller regions. Broutin et al. (2004) determines the hierarchical diffusion of pertussis in a small region of Senegal using critical community size method. They showed the diffusion of pertussis is from centre city to village. Moreover, Rohani et al. (2000) studied pertussis time series in 60 cities in England and Wales using the mean duration of fadeout in relation to population size. They identified the bigger the locality, the shorter the mean duration of the fadeout. Wallace and Wallace (1999) also studied the diffusion of AIDS in US and it diffuse hierarchically among the metropolitan areas, from larger to smaller regions, beside the national travel path. Similarly, Bartlett (1960) used critical community size 250,000 to 300,000 and she pointed out the diffusion of measles in United states is from large cities ignited sub-sequent epidemics in smaller towns. Lloyd (2001) studied the persistence of childhood viral diseases such as measles relating fadeout percentage with population size using a mathematical model. He showed hierarchical diffusion of disease and highlighted, larger population size has high disease persistence and in smaller populations the chain of transmission is likely to be interrupted. Broutin et al. (2004) calculated Pearson correlation coefficients between the total rural time series and the proportion of pertussis disease cases of each city to determine urban-rural hierarchical diffusion. They plotted these correlations against the population size and their plot's indicate that a strong negative urban-rural correlation. From the plots, they suggest that disease cases arise in towns and diffuse hierarchically to rural areas.

2.2.2. Synchrony diffusion

Synchrony or synchronicity in the diffusion of diseases occurs when parallel development in the number disease cases or timing of infection between geographically separate locations that have similar size or timing (Cliff et al., 1981). Synchrony exists also as part of a hierarchical diffusion pattern. Cities at the same level of the hierarchical diagram shown in Figure (2.1) are synchronised. However, it can also occur in a non-hierarchical setting between spatially segregated populations which have similar timing of diffusion.

Synchrony has been measured in different ways and varieties of mechanisms have been applied to explain this synchrony. A number of studies have defined synchrony of disease diffusion between different spatially separated areas (time series) in different ways such as, size synchrony, timing of synchrony, coincidence of peak synchrony and change synchrony. For example, Schanzer et al. (2011) determine synchrony for the geographical units of Canadian provinces and US surveillance regions using the peak of epidemics. They defined as the epidemic peak timing for each city and assessed synchronization by calculating the difference in timing of epidemic peak between the largest and all other communities within the province.

The most common explanation provided to assess synchrony is using a Pearson correlation coefficient between two spatially separated areas or two time series. Pearson correlation works by measuring the similarity of a linear association between two time series (variables) as a function of one relative to the other. Several researchers have used Pearson correlations to analyse the association between two series. For example, Jacobs et al. (2014) used Pearson correlation coefficients to analyse the association of timing of the peaks of meningococcal disease with influenza in the united states. Broutin et al. (2004) determine synchrony between urban and rural areas using auto-and cross-correlations. Similarly, Bjørnstad et al. (1999) used Pearson correlation coefficients to quantify synchrony based on change in the proportion of disease cases. They showed that change synchrony often exists when the patterns of diffusion rise and fall together.

2.3. Spatial grouping methods

Interaction between people is the main factor to facilitate disease diffusion (Bharti et al., 2010). This highlights the need to understand the diffusion pattern of a given infectious disease based on connectivity and distance travelled by the host. It is important to group areas based on connectivity for a better understanding of diffusion. Especially in cases like in the Netherlands where cities are very close together and so much commuting takes place and no real large cities are available to maintain the CCS (250000 – 400000 inhabitants). There are only three cities that meet these numbers and these large cities are still small compared to large cities in some other countries like London, Paris and New York. Previously, different urban grouping techniques were applied in different disciplines to group areas based on connectivity and distance. For example, the national government of the Netherlands (Ministerie van VROM, 2004) groups the country using urban networks. Arcaute et al. (2015) used the distance between road intersection points from the road networks to split Britain through percolation method. Commuting distance is also used to group areas. Commuting means regular movement of people between their residence and place of work or school and it is the main factor to facilitate disease diffusion. So grouping an area based on the commuting distance is also a promising approach for analysing the diffusion pattern of diseases. In this research three different approaches were followed for grouping area and defining different urban zones in the Netherlands. Based on connectivity and commuting distances, percolation method, urban networks and commuting distance have been used for this purpose.

2.3.1. Percolation method

Percolation is the movement of material through paths (Meyers, 2012), for example, water can percolate through the soil (finding space to move through the soil). Percolation theory describes the behaviour of connected clusters in a graph. We can replace the water by commuters and the graph by the road network. When we do this, percolation theory can help us to determine network clusters corresponding to areas which define commuting units.

Percolation can be associated with critical phenomena that are related to different environmental, economical, and social aspect. For instance, percolation can be related with network of nodes. A network percolation model is a collection of nodes distributed in space and the network of those nodes looks like a collection of isolated urban zones thus causing breaks in the percolation process (Essam, 1980). Saberi (2015) defined a percolation model as the simplest fundamental model in statistical mechanics that exhibits phase transitions expressed by the occurrence of a giant connected component. A giant component is a connected sub cluster that contains a majority of the entire cluster nodes. The percolating network responds to transitions by changing the fraction of this giant component.

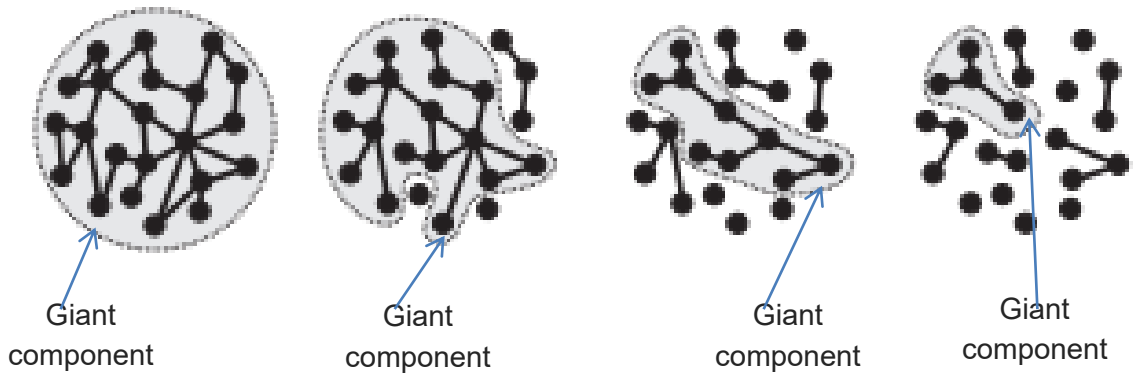


Figure 2.2: Shows a giant component of connected nodes (Soriano et al., 2010)

Percolation method very recently has been applied to a large variety of critical phenomena. Hennemann et al. (2015) defined percolation in a different way as a geometrical–mathematical model that can be easily implemented in different disciplines. The percolation model uses very simple rules and it has been applied to characterize a large variety of critical phenomena.

Several researchers have used percolation models for many applications to cluster and understand important features of physical, biological, chemical, natural, technological or social systems. Grimaldi and Balberg (2006) applied a percolation model in physics to characterize the conducting phase in a conducting material. They used a distance threshold and the distance between the centres of the two nearest neighbour particles. In their study, the percolation model was successfully applied to partition the conductivity phase of the particle. Li et al. (2015) used a percolation model to define the reliability of a network. They showed, network failure can be regarded as a percolation process by setting the critical threshold of percolation as a network failure criteria. Another recent application of percolation theory is for clustering areas based on their connectivity. For instance, the percolation method was applied to partition Britain into a hierarchy of urban zones at different scales that are independent of administrative arrangements using intersections of road network by successively reducing a distance threshold (Arcaute et al., 2015). The urban zones they obtained by using this method are in a very good correspondence to the morphological definition of cities given by satellite images and other clustering methods. Similarly, Molinero et al. (2015) used the percolation method to cluster Britain into distinct regions and nations using the intersection of a road network at a very fine spatial scale. They used these percolation urban zones to allocate the voting patterns to a new hierarchy of constituencies and applied this method to generate voting predictions. The results show a picture of how Bratian might vote on geographical lines that are important to think about the hierarchy of cities and regions.

3. MATERIALS AND METHODS

This chapter describes the datasets used during this research (section 3.1), the data preparation performed to make these datasets suitable for analysis (3.2), and the corresponding methods that were applied during the analysis to answer the research questions and achieve the main objective (section 3.3).

Section 3.3.1 describes identification of urban areas and how the identified urban areas were compared and section 3.3.2 explains how the disease data were analysed on the defined urban zones and diffusion patterns of the disease were measured.

3.1. Data

There were four datasets used in this analysis: a dataset of disease, a dataset of mobility, a dataset of population and a dataset of municipalities.

3.1.1. Disease data

For the purpose of this research three epidemic years of pertussis disease data (1996, 1999 and 2001),as a monthly count of the number of pertussis cases per municipality, were received from the National Institute for Public Health and the Environment (RIVM). The dataset covers the whole country of the Netherlands at a monthly temporal aggregation level and a municipal spatial aggregation level. The disease dataset consists of a polygon shape file for each month making 12 shape files for each year. These shape files for each year were joined using a spatial analysis tool "join field" in ArcGIS 10.3. The total number of municipalities used in the analysis was 538.

3.1.2. Population data

One of the dataset used was the population data of the Netherlands. Population data of the Netherlands at municipality spatial aggregation level were downloaded from statistics the Netherlands[1] for the year 1998 and used to analyse the diffusion pattern of pertussis disease in relation to population size. The reason to choose the population data of 1998 is that it contains 538 municipalities which is the same number of municipalities in disease dataset.

3.1.3. Mobility data

The Netherlands roads and railways shape files were downloaded from free Netherlands ArcGIS shape file map layers[2] and used as mobility data. For these dataset, data projection and transformation were performed on this dataset to change from GC WGS-1984 to GCS Amersfoort. The layout of the road layers was visually assessed to determine which road layers are important for the analysis. However, the road layers dataset suffers from quality problem. For example, there are road layers that are not connected to any other roads and road types that do not contribute to connectivity of cities. In this research, such problematic types of road layers were removed and the primary, secondary, trunk, motorways, residential and track road layers were kept.

Similarly in the railway dataset there were different types of railways such as: platform preserved, construction, disused, industrial, light-rail, station, subway, tram and rail. Among those types of railways the only rail types were selected and used for the analysis of synchrony of diffusion. In another step for

[1] http://www.cbs.nl/nl-NL/menu/home/default.htm
[2] http://www.mapcruzin.com/free-netherlands-arcgis-maps-shapefiles.htm

preparing the road dataset, the municipality data were obtained from Public service card (PDOK)[3] for masking the selected road datasets in order to remove water bodies is a central facility which provides the geographic accessibility of geo-datasets of national importance.

3.2. Data preparation

Data preparation is the step to make the datasets fit for use in the research work. It is required to structure datasets both in space and time.

3.2.1. Data preparation of road data

Here, the first step was data preparation for mobility data. For this data, road layer were used and from this road layers a road network was generated. Then, the road intersection points were extracted from the road network. There were lots of intersection points that are not nodes of road intersection points. To clean those intersection points only the intersection of the roads which are connected to at least three road segments were selected and the rest were removed. In addition, those selected intersection points were masked by municipalities to limit the boundaries of the country and to exclude water bodies.

To apply this method, first the road layer was changed to a road network. From the road network the intersection points were found. Also the number of counts of each line connected to a single node was calculated. The number of counts of intersection points was computed by adding a count field to the road networks that count number of lines intersected with a single node. These numbers of counts were used to select intersection of the roads which are connected to at least three road segments. And about 40% of the intersection nodes were selected for the analysis of percolation method.

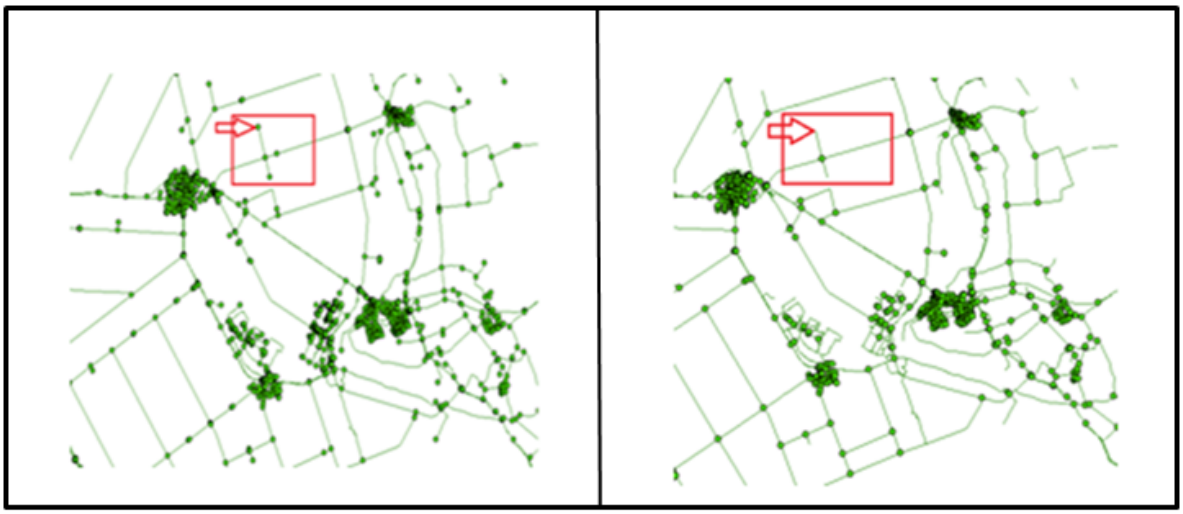


Figure 3.1: (left) before removing endpoint nodes, (right) after reducing endpoint nodes of road intersection points.

3.2.2. Data preparation disease data

The next step during this stage was data preparation of the disease datasets and also at this stage the following assumptions and definitions of fadeouts were made prior to the preparation of the disease data. It was assumed that the absence of the disease during one month is a fadeout and where fadeout is

[3] www.pdok.nl

defined as the first month without infection given that there was an infection in the previous month. Duration of fadeout is defined as the length of consecutive months without infection.

Next, municipalities with total sum of zero disease cases were discarded from the dataset to avoid their effects during the counting of fadeouts and frequencies of the disease. After removing municipalities with zero disease cases, the number and duration of fadeouts for each municipality per year were calculated. Then, the generated dataset was used to analyse the total number and duration of fadeouts of the disease in each municipality.

In addition the frequency (number of disease cases per 100000 inhabitants) of the disease was calculated at the municipality spatial aggregation level. The following methods explain in detail how those prepared datasets were applied.

3.3. Methods

This section discussed the methods that were applied to identify the urban zones, to analyse the disease data in order to achieve the objectives of the current research. The data analyses were performed using R version 3.1, which is statistical computing and graphics software (Grunsky, 2002). In addition, Python version 3.4, an interpreted, object-oriented scripting language (Dobesova, 2011) and ArcMap 10.3 were used for performing different analysis.

3.3.1. Identification of urban areas

The dense social-contact networks of urban areas characterize them to form a perfect match for fast, uncontrolled disease diffusion (Eubank et al., 2004) and this urban areas appear at different scales and have different spatial distributions with a numerous type of transition zones. In addition, the movement of most individuals is limited within particular urban areas (Kang et al., 2012). As a result, this research intended to identify urban areas and analyse how the disease diffuses in these areas. In order to do this, three different methods: percolation method, urban agglomerations and commuting distance method were applied in this research work. The next sections explain how these urban zones were identified using each method.

3.3.1.1. Percolation Method

Percolation method can be used to split areas based on distance of road intersection points (Arcaute et al., 2015). In this study percolation on the intersection points of road networks was applied to group the Netherlands' municipalities by treating the whole country as a connected massive region based on road network connectivity.

After removing endpoint nodes, splitting areas based on road intersection points (percolation method) were conducted using different cell sizes and distance thresholds. Details of the technique to generate the percolation urban zones are explained by Arcaute et al. (2015) and implemented as indicated by the offers in ArcMap 10.3. The percolation procedure applied using road intersection distance consists of the following steps:

1. Convert selected intersection points to raster
2. Calculate Euclidian distance
3. Apply distance threshold
4. Region grouping
5. Averaging urban zone
6. Selection of urban zones

The general overview of the algorithm follows the following steps.

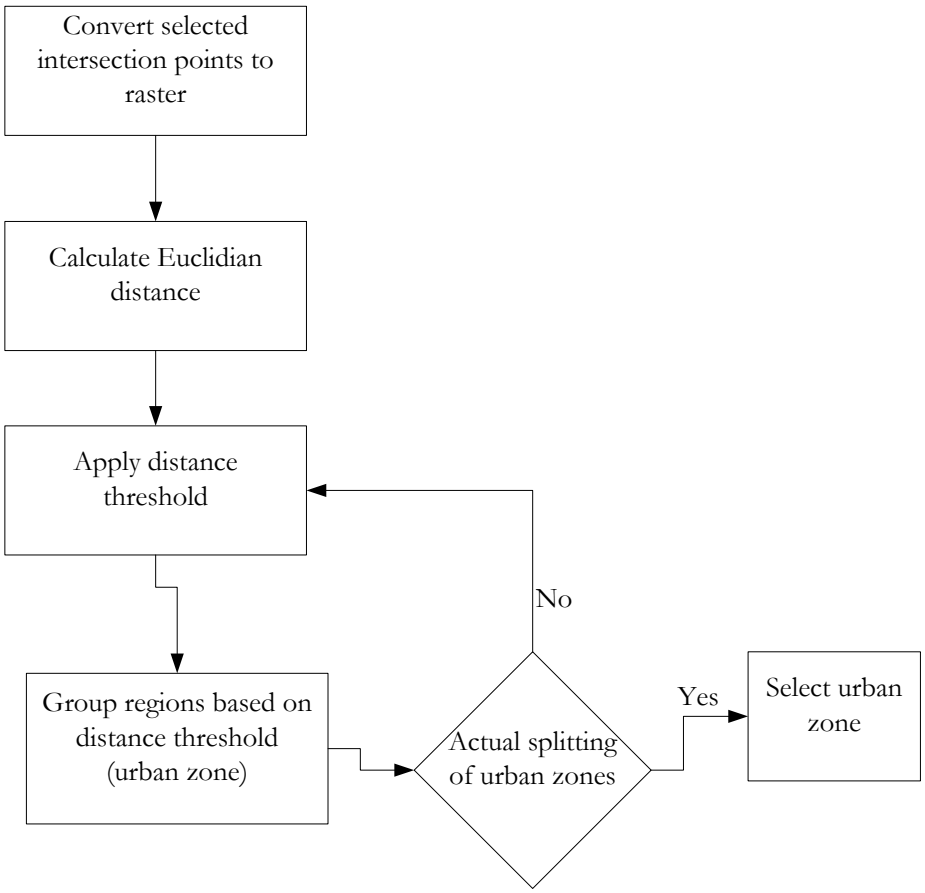


Figure 3.2: General overview of percolation method

First, the selected nodes were converted to a of 10x10 meter raster. And then, these selected intersection points are masked by municipalities to limit the boundaries and to calculate the Euclidian distance between the road intersection points. From the raster the Euclidian distance between intersection points was calculated and given below in Figure (3.3).

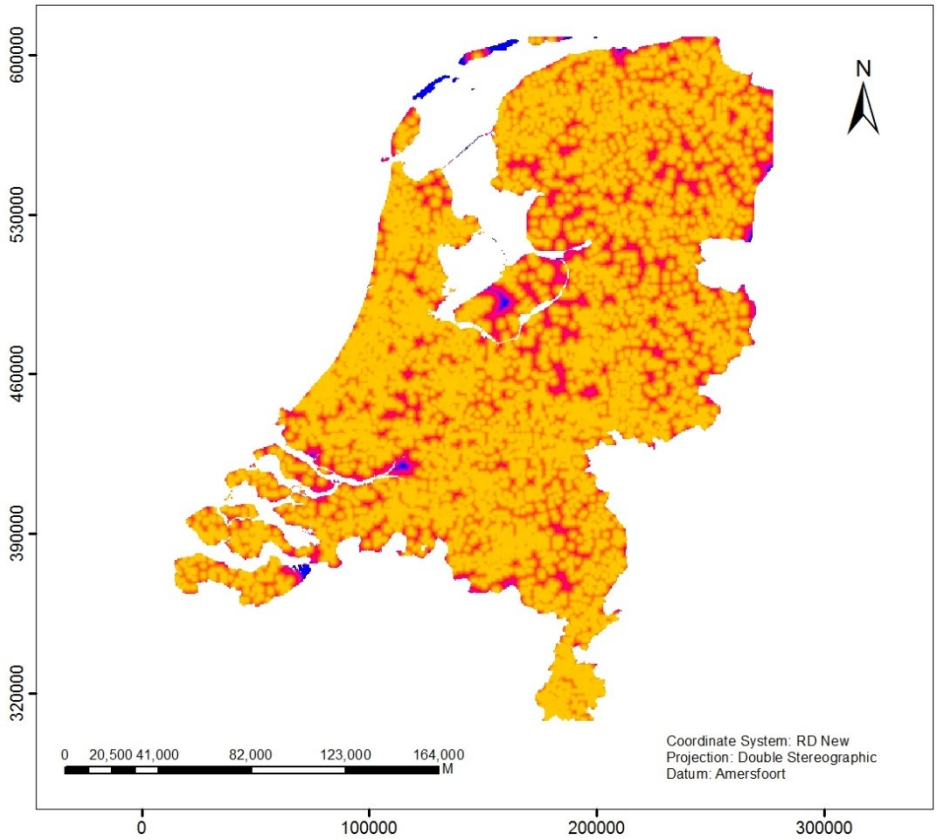


Figure 3.3: Euclidean distance based on road intersection points.

After calculating the Euclidian distance the algorithm starts from the entire country by specifying the maximum distance threshold. The algorithms began with the distance thresholds which contain all the points connected each other and it take these connected points as a giant component. As explained in the literature review section, the giant component responds to changes on the connected nodes when the distance threshold changes. Different distance thresholds were applied and different sizes of urban zones appeared by changing this giant component. Next, region grouping using map algebra was applied by creating same unique identifier for each area which is applied within the distance threshold. Finally, the distance threshold which gives the actual splitting of urban zones were selected and used for next analysis steps.

The size and shape of the urban zones differ based on the cell size and distance between the road intersection points. The urban zones used for this method were grouped by setting the distance threshold to 590 meters because this distance threshold value gives the actual splitting of urban zones (the country was grouped in terms of urban network connectivity). The splitting at this step was not because of other breakings such as rivers, forests, polders, or islands; it is a break between urban and non-urban areas. Figure (3.4) showed the selected urban zones that show the actual splitting of urban zones and 30 largest urban zones extracted from the raster map of 590 meters.

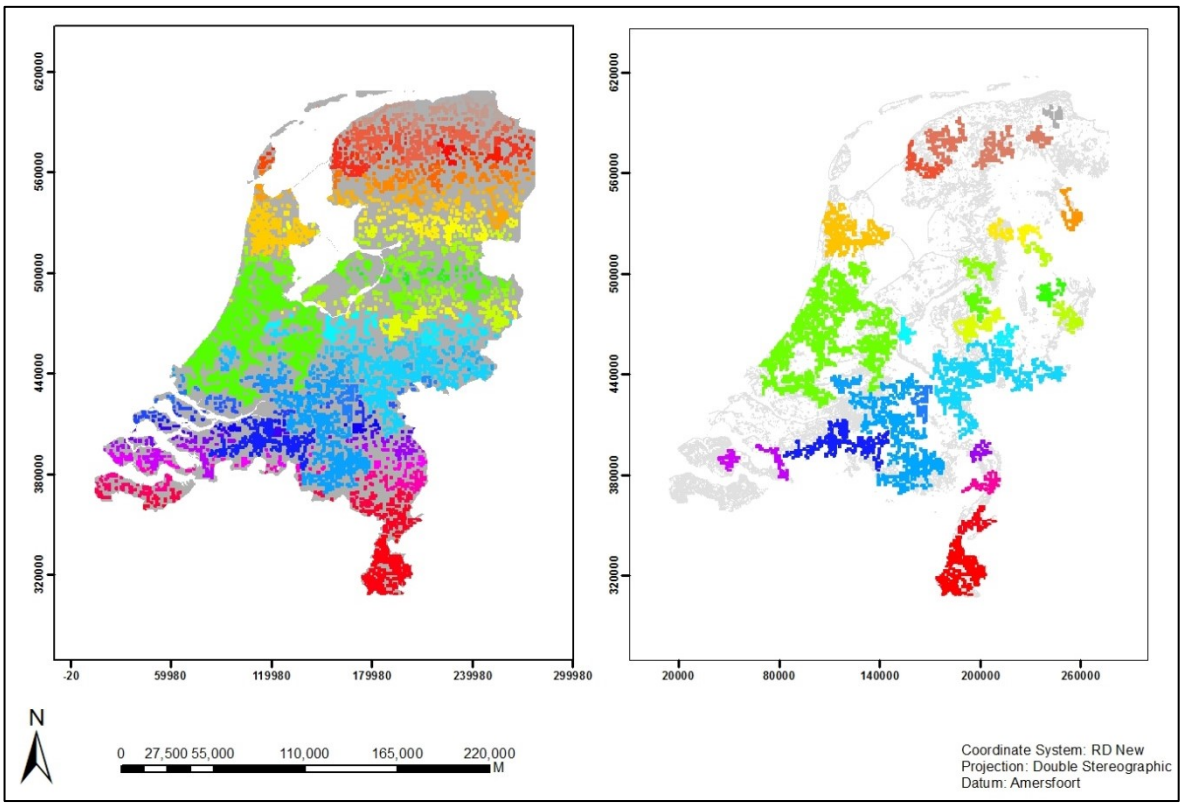


Figure 3.4: Map of urban zones at 590 distance threshold and extracted 30 largest urban zones

These extracted urban zones of raster were converted into polygons. However, there were gaps inside the converted polygons. Since we need the areas inside the urban zones, we need to fill these polygon gaps. In order to achieve this, the following steps were performed in ArcMap 10.3.: Minimum bounding geometry => Erase by the polygon => Multipart to single part => Edit holes => Union => Dissolve. The generalization was performed on nine largest urban zones. These generalized selected nine polygons were used for the next analysis steps.

3.3.1.2. Splitting using urban agglomerations

The Netherlands has been divided into regions, provinces and municipalities. Among the whole country two third of the land area are still under agricultural use (Koomen & Groen, 2004) and there are some areas more urbanized than the others. Urban regions show high levels of incoming and outgoing commuting (van der Laan, 1998). The national government in the Netherlands has selected six urban agglomerations (Ministerie van VROM, 2004). These are: Randstad Holland, Brabantstad, Zuid-Limburg, Twente, Arnhem-Nijmegen and Groningen-Assen. In this study, these identified urban agglomerations of the Netherlands were used to split the country into different urban zones. Because these urban agglomerations show high level of interaction between people and can be used to split areas based on connectivity and used as a mobility data to analyse the diffusion of the disease in relation to human mobility. The image in Figure (3.5) given below show the urban agglomerations of the Netherlands and this image was used to cut out those urban agglomerations and have been used to split the Netherland into different urban zones.

Figure 3.5: Image of urban agglomerations in the Netherlands, according to Ministerie van VROM (2004)

After geo-referencing this image, those urban networks were digitized. These digitized polygons were used for the comparison of the urban network mobility data.

3.3.1.3. Splitting using commuting distance

Commuting is one of the basic elements of diffusion of disease from one place to another because of creating interaction between peoples. Regular or random commuting may occur in space and time at various scales. Within literature we found several different average commuting distance values in the Netherlands. For instance, according to Susilo and Maat (2007) average commuting distance in the Netherlands is 35kms. Based on Statistics Netherlands (2015) Dutch people move 30kms a day. For this study, the 12 largest municipalities: Amsterdam, Rotterdam, The Hague, Utrecht, Eindhoven, Tilburg, Groningen, Almere, Breda, Enschede, Apeldoorn and Nijmegen were selected based on their population size from Wikipedia[4] Hence, the spatial diffusion of communicable disease is mainly tied to mobility; the movement of peoples around these selected municipalities were chosen for the analysis. A commuting distance of 30kms from each centre of selected 12 municipalities was used to split the country based on commuting distance and the Figure (3.6) shows those selected 12 largest municipalities. The analysis was done by drawing a buffer of 30kms from the centre of the selected municipalities.

[4] https://en.wikipedia.org/wiki/Template:Largest_cities_of_the_Netherlands

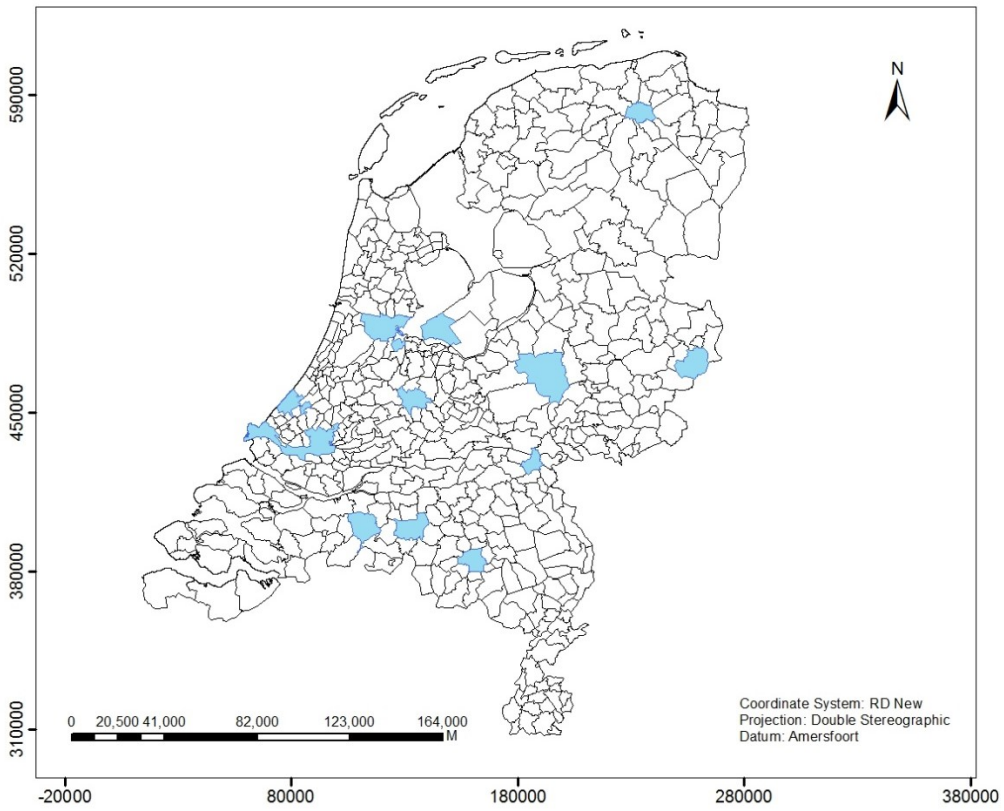
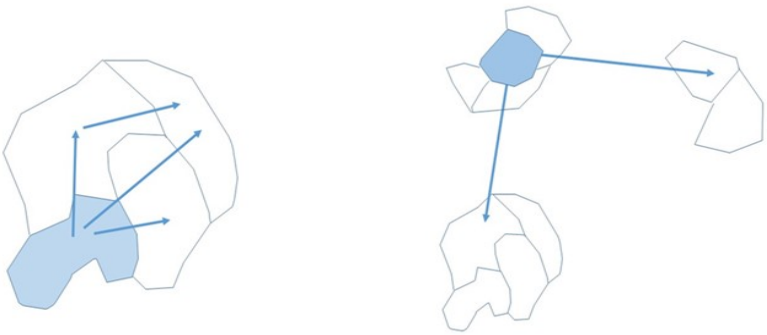


Figure 3.6: 12 largest municipalities in the Netherlands selected for the identification of urban zones based on commuting distance

3.3.2. Method for disease analysis

After comparison of mobility data and selecting one method, the next step was performing the disease analysis on the identified areas. These analyses can be grouped into analysis on the hierarchy and analysis on synchrony. Each of these sets of analysis is applied at different aggregation levels, within an urban zone and between urban zones (3.7).



Hierarchy within an urban zone

Hierarchy between urban zones

Figure 3.7: Within and between urban zone hierarchy spatial aggregation levels

The next sections describe how these different aggregation levels were and how difference in aggregation level affects the diffusion patterns.

3.3.2.1. Method for measuring hierarchy of diffusion

Disease diffusion follows size hierarchies: cases arrive initially in biggest Population sizes and spread to the surrounding smaller population sizes (Broutin et al., 2007). This size hierarchy can be determined by population size and disease fadeout. In this research this size hierarchy method was applied to measure hierarchical diffusion of the disease and the measure of hierarchy was performed in two ways. The first one was the measure of hierarchy using the total number of fadeout and the other one was a measure of hierarchy using mean duration of fadeouts. In order to apply this method, we first calculated the total number of fadeouts per year for each municipality. Second, we determine the mean duration of fadeouts. Third, the municipalities cases are aggregated into urban zones and the total number of fadeouts and mean duration of fadeouts were calculated per municipality and per urban zone. Next, total number of disease fadeout (mean duration of fadeouts) in relation to population size was plotted. Then, the Pearson correlation coefficient analysis method was used to see the relation between population size and disease fadeout. Pearson correlation coefficient is used to determine the dependence between two variables (Plata, 2006). It gives values between -1 (negative correlation) and 1 (positive correlation) inclusive. Pearson correlation coefficient represented by a Greek letter ϱ (rho) and can be calculated as follows:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma x \sigma y} \quad (1)$$

Where, $\rho_{x,y}$ is the Pearson correlation coefficient, cov(x, y) is the covariance between x and y, σx is the standard deviation of x and σy is the standard deviation of y. These correlation coefficients were used to decide the type of regression model to fit. The model was fitted using the summery statistics, i.e. mean of x ($\bar{x}$), standard deviation of x (SD_x), mean of y($\bar{y}$), standard deviation of y(SD_y) and the correlation coefficient $\rho_{x,y}$. In addition, the trends of the plot also used to determine the type of the fitted model.

After the type of regression model was determined, the analysis was done on the residuals from the predicted relationship of population size and fadeouts. Larger population size with fewest fadeouts has negative residuals and are potentially epidemically important for spread of disease to neighbouring regions (Bharti et al., 2010). So from the fitted model the predicted residuals between population size and fadeouts, the sequence of diffusion can be determined by the size and sign of residuals. Population sizes with fewest fadeouts have negative residuals and smaller size. Medium population size with larger fadeouts has positive residuals and larger in sizes. The flow chart given below shows the steps followed to analyse the hierarchical diffusion.

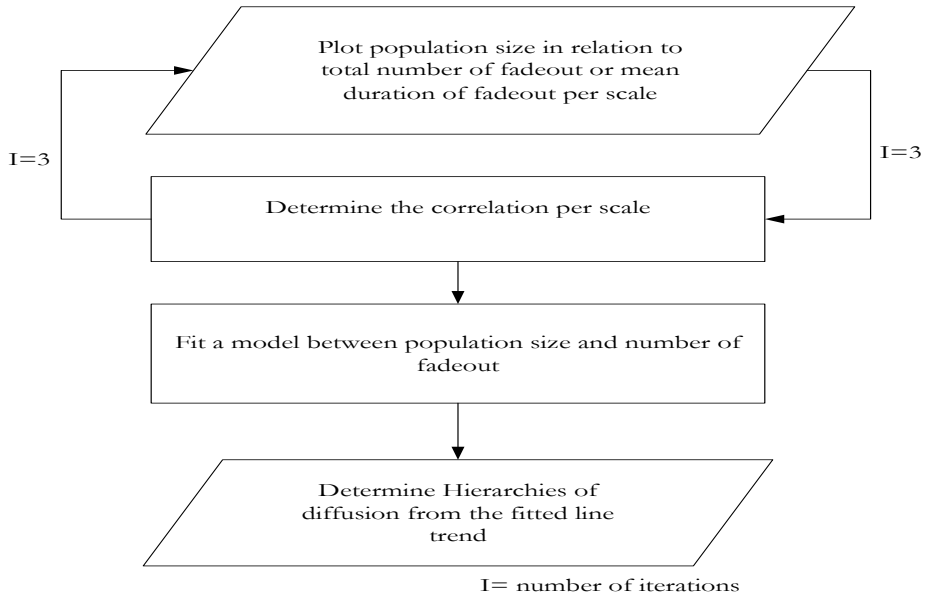


Figure 3.8: Flow chart showing analysis of hierarchical diffusion

3.3.2.2. Method for measuring synchrony diffusion

Synchrony between urban zones and within urban zones were investigated by calculating the frequency (number of disease cases per 100000 inhabitants) of the disease for each municipality. Two sets of analysis were performed to measure the synchrony. The first assessed the synchrony of disease frequency by calculating the Pearson correlation coefficient between urban zones and within urban zone and again, Pearson correlation coefficients between cities connected by railways were calculated for three years. In addition, Pearson correlations between the largest urban zone (Randstad Holland) with the other urban zones were assessed to compare based on their distance from the largest urban zone (Randstad Holland) and to see the impact of separation distance on the diffusion pattern of disease between urban zones.

Moreover, line based synchrony was examined to check if the connectivity between cities has an impact on the diffusion pattern using the intercity line (track) between cities. The intercity lines were identified based on the map given below in Figure (3.9).

Figure 3.9: Intercity network of cities in the Netherlands

Page url= https://commons.wikimedia.org/wiki/File%3AIntercitynet_NL_2015.png

From this figure four intercity lines were selected. The four intercity Lines were listed below from 1-4 using name and municipality code.

1. Rotterdam(599)=>Gouda(513)=>Utrecht(344)=>Amersfoort(307)=>Zwolle(193) =>Groningen(14))
2. Den Haag(518) => Leiden(546) => Amsterdam(363) =>Amersfoort(307) =>Zwolle(193 => Leeuwarden(80)
3. Den Haag(518) => Gouda(513) => Utrecht(344) =>Amersfoort(307)=>Deventer(150) =>Enschede(153)
4. Roosendaal(1674) => Breda(758) =>Den Bosch(796) => Nijmegen(268) => Arnhem(202) => Deventer(150) =>Zwolle(193)

After selecting these intercity lines synchrony between these connected cities was analysed to check if there is synchrony in more connected cities than in less connected cities. Then, synchrony was assessed

based on Pearson correlation coefficients between cities connected by an intercity line in a similar way as the methods of urban zones and railway networks.

The correlation coefficients for all the synchrony measures were calculated using Equation (3.9). The calculated Pearson correlation coefficients lead to a two dimensional matrix. This matrix was visualized using correlogram by installing the corrplot package in R to explore synchrony. A pairwise averaging of Pearson correlation coefficients was performed to assess the synchrony within and between urban zones and also between cities connected by the railway network and intercity lines. This average measure gives the overall proportion of pairs that agree with their change in values (moving up and down together). Higher correlation value means similarity in patterns of synchrony of diffusion. The flow chart given in Figure (3.10) below presents the methods applied during the analysis of synchrony of diffusion.

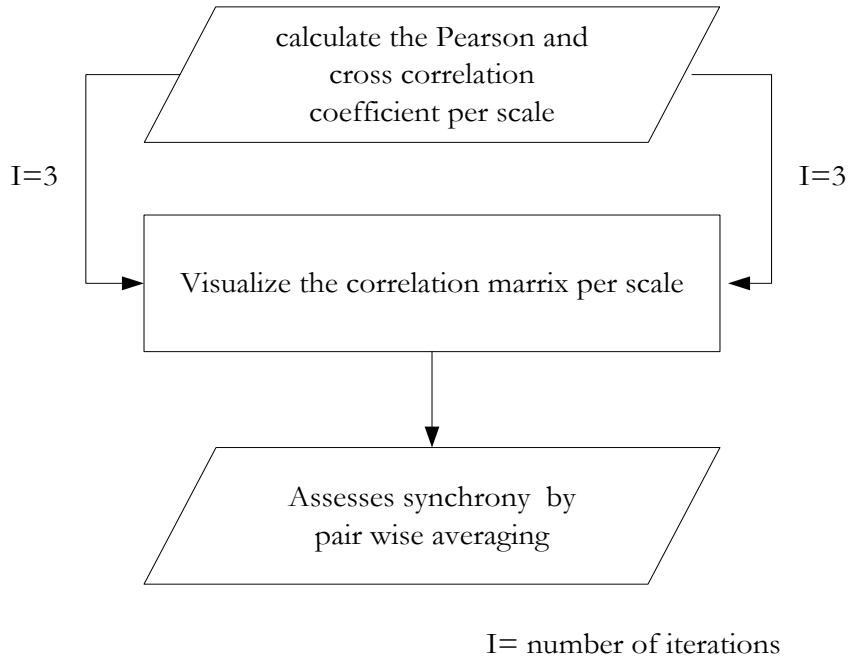


Figure 3.10: Flow chart showing the analysis of synchrony diffusion

3.3.2.3. Method to measure urban versus non-urban areas disease diffusion patterns

During the analysis of non-urban areas, 127 municipalities were selected. These municipalities are those which are not identified as an urban zone during the grouping of urban zones. Then municipalities with zero total disease cases were removed. The numbers of municipalities without zero disease cases are different for the three epidemic years. The selected non-urban municipalities for the three years are shown below in Figure (3.11).

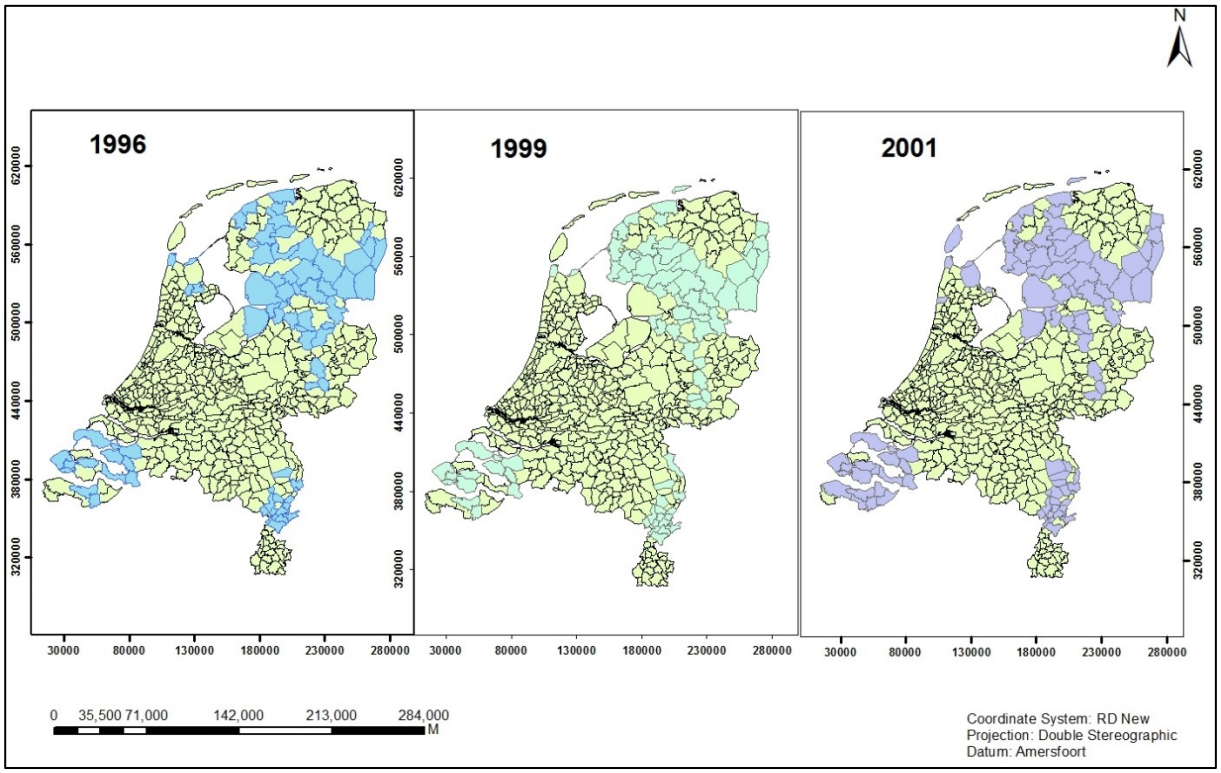


Figure 3.11: Selected non-urban area municipalities for the three epidemic years to analyse urban versus non-urban area diffusion pattern

After removing municipalities with zero disease cases, a month contain the peak value of cases during the year for each municipality were identified using a python code. Municipalities with peak value of one were removed, because one disease case is not a peak, it means the disease is just diagnosed in that area. Next, we rank month with a peak disease case. Months are ranked according to their known sequence, i.e. January= 1, February = 2, March= 3,… and December= 12. Then to see an urban versus non-urban patterns of transmission, a month which is identified with peak value of disease cases during the epidemic year for each municipality in the urban area were also ranked similarly as applied to the non-urban areas. After ranking each month, which contain a peak value of disease case, the means of the ranks were calculated for both urban and non-urban areas. Then, the results were compared by graphical representations.

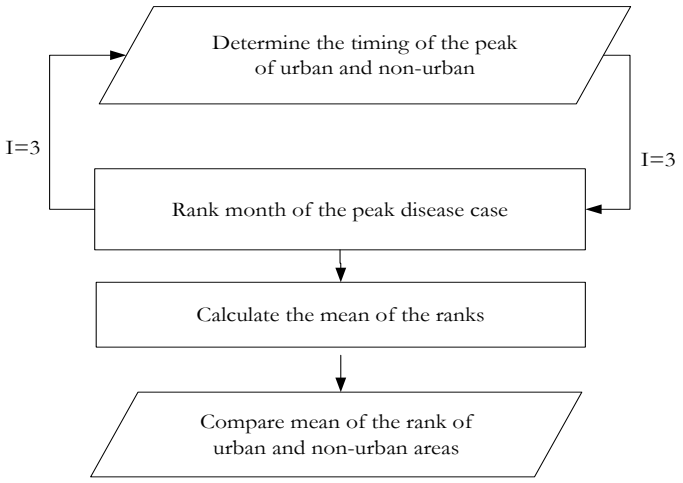


Figure 3.12: Flow chart showing the analysis of urban and non-urban zones

4. RESULTS

4.1. Introduction

This section describes the results of the applied methods and the data analysis involved to achieve the research objectives and answer the research questions. The results are organized as follows: Section (4.2) explores important properties of the disease data based on different perspectives which includes number of cases per epidemic year, the relationship between disease cases and population, and spatial distribution of disease cases. Section (4.3) discussed the identification of urban zones based on three different methods. Section (4.3.1) explains the percolation method which was used to identify urban zones based on distance of road intersection points and the comparison and selection of proper urban zones were present in section (4.3.2).

After selection of urban zones the analysis of disease diffusion was presented in section (4.4). The analysis of measuring hierarchical diffusion using two different datasets duration of fadeouts and number of fadeouts in relation to population size at different scales was discussed in section (4.4.1). Additionally, analysis of measuring synchrony of disease diffusion between urban zones and within urban zones using the frequencies of the disease was discussed in section (4.4.2). In this section, a selection of large cities using railway network and intercity train tracks were performed to measure synchrony in highly connected cities. Finally, the results of urban versus non-urban zone diffusion pattern were discussed in section (4.4.3).

4.2. Initial exploration of the disease data

In order to gain a better understanding of the disease data, an initial exploration has been performed in order to identify:

- Differences in number of disease cases between the epidemic years
- Relationship between the municipal population size and the number of disease cases
- The spatial distribution of the disease cases at municipality level
- Temporal analysis of the three epidemic years

The outputs of these analyses were used in the next stages to analyse the diffusion pattern of pertussis in the Netherlands at different scales.

4.2.1. Number of cases per epidemic year

Data for three epidemic years are being used in this study, 1996, 1999 and 2001. Box plots of the disease data were plotted for the three years and it is given below in Figure (4.1). The number of disease cases per municipality for three years are plotted next to each other and compared based on the size of their box plots.

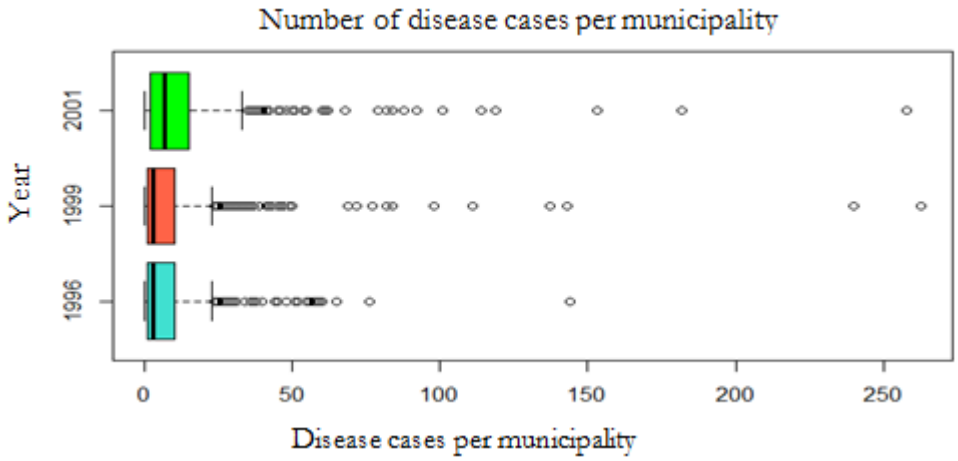


Figure 4.1: Box plots showing the yearly pertussis cases for the three epidemic years

The box plots show the number of disease cases per municipality for the three epidemic years. From the box plots we can see the difference and similarities between the epidemic years. All the three epidemic years show similarity in their box plots. Most of the municipalities have fewer than 50 disease cases for all epidemics in all three epidemic years. Very few municipalities have a higher number of disease cases. There seems to be an increase in the dispersion (more municipalities with a higher number of disease cases) in 2001 compared to 1996 and 1999. This indicates that there was an increase in the number of cases in the last epidemic.

4.2.2. Relationship disease cases and population

The following graphs show the number of inhabitants in relation to municipality disease cases. When we plot the disease cases in relation to the number of inhabitants, we can observe the occurrence of disease cases based on community size of municipalities.

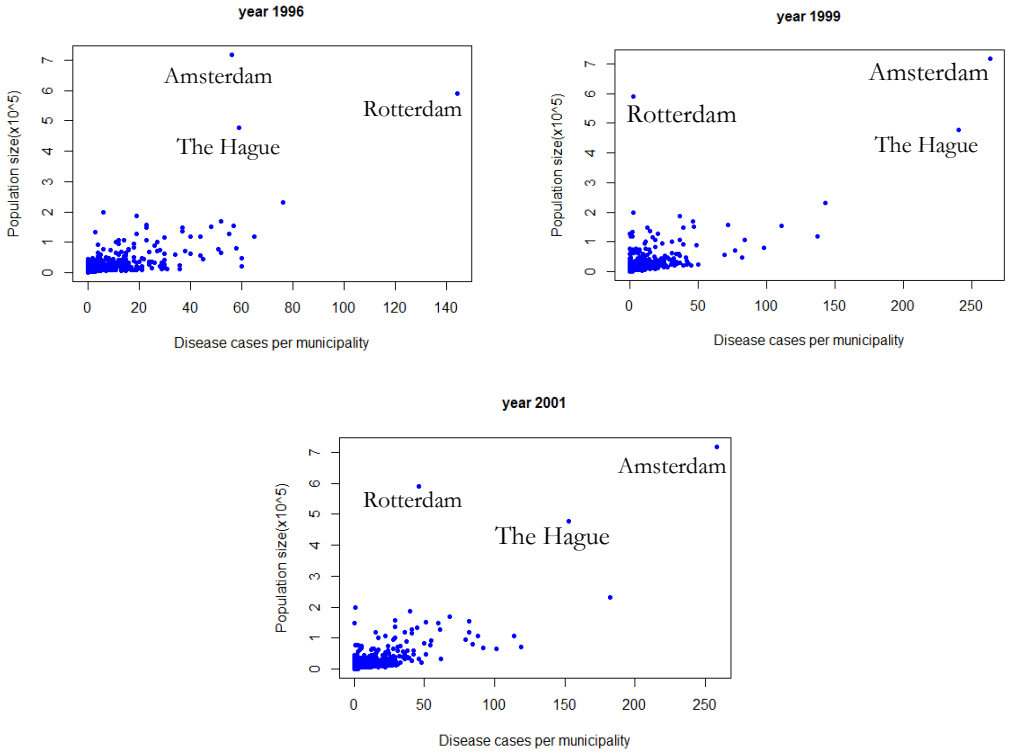


Figure 4.2: The number of inhabitants in relation to municipality disease cases for three years

The plots show that we have only three cities that have a large number of inhabitants with large number of disease cases. These three cities: Amsterdam, Rotterdam and The Hague do not have a higher number of disease cases in all epidemics. In 1996 only Rotterdam has up to 140 disease cases. However, this city has 3 disease cases in the 1999 outbreak and very few (46 disease cases) in the 2001 outbreak. Amsterdam has few cases (less than 60) in the 1996 outbreak, but many in both the 1999 and 2001 outbreak. The last city The Hague has few cases in 1996, many in 1999 and a considerable number in 2001. This indicates diversity in behaviour of cities between the three epidemics.

4.2.3. Spatial distribution of disease cases

This step aims to visualize and understand the spatial distribution of pertussis in the Netherlands. The maps given in Figure (4.3) show where the disease cases are located. The cases are somewhat more dense in the western and central parts of the country, but they are distributed evenly in all parts of the country and do not show a specific location. So, due to this reason, we need to pin down the source and diffusion pattern of the disease. We can see that numbers of disease cases are higher in the year 2001. In this year disease cases are higher in all parts of the country. Disease cases getting higher even in non-urban areas of the country during this year. There are a number of municipalities with disease case between 77 and 300. In the year 1999, there are some municipalities with disease cases between 77 and 300 and these are mostly located in the western part of the country and very few disease cases in the non-urban areas. In year 1996 there is only one city which is Rotterdam, which has a disease case in between 77 and 300. During this year, most of the disease cases located in the western part of the country and very few disease cases for the non-urban areas. In general there is also diversity in the spatial location of disease cases.

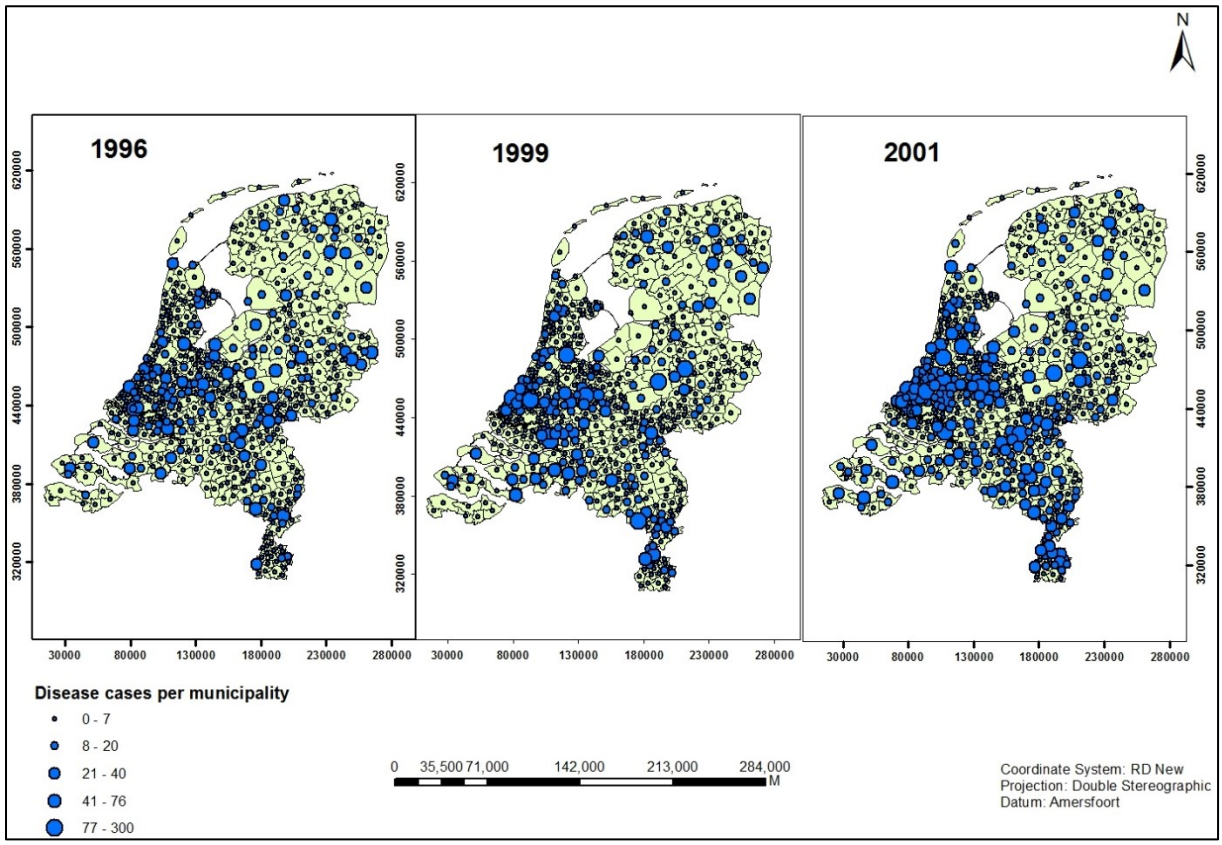


Figure 4.3: Spatial distribution of disease cases for the three epidemic years

4.3. Splitting areas using different mobility data

4.3.1. Percolation method

This step of the research aimed at splitting the country using a road intersection points. To achieve this goal percolation method was selected. Figure (4.5) shows the evolution of splitting of areas at different distance thresholds applied to the percolation method. The graph was plotted by calculating the average number of intersection points (number of intersection points in each cluster divided by the total number of clusters) for each distance threshold. At 4500 meters thresholds, only islands were split. Next, polders were split at 3500 meters. Then, areas separated by natural breaks such as rivers and forests were split at 1300 meters. The splitting of actual urban zones starts at a distance between 600m and 500m. Urban zones which were split at distances less than 500m are very small and do not contribute to the emergence of regions.

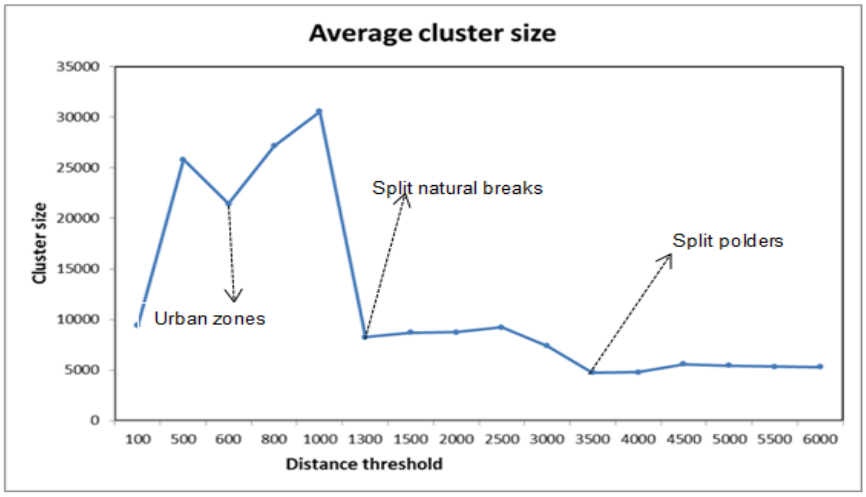


Figure 4.4: The evolutions of urban zones using the percolation method

The threshold distance 590 meters is selected after applying different distance threshold. Because, at this distance the splitting of urban zones is appropriate. Meaning that, areas are grouped based on their urban network and it shows the actual splitting of areas. Above this distance threshold urban areas are not split. Below this distance threshold urban areas are very small and have a size less than a municipality. Different distance thresholds show appearance of different clusters and the results of these different threshold values are shown in Appendix (A). Figure (4.6) shows the nine largest urban zones identified at 590 meters distance threshold after extracting raster and converting to polygons. These polygons were used for the comparison of percolation method with the other urban identification methods.

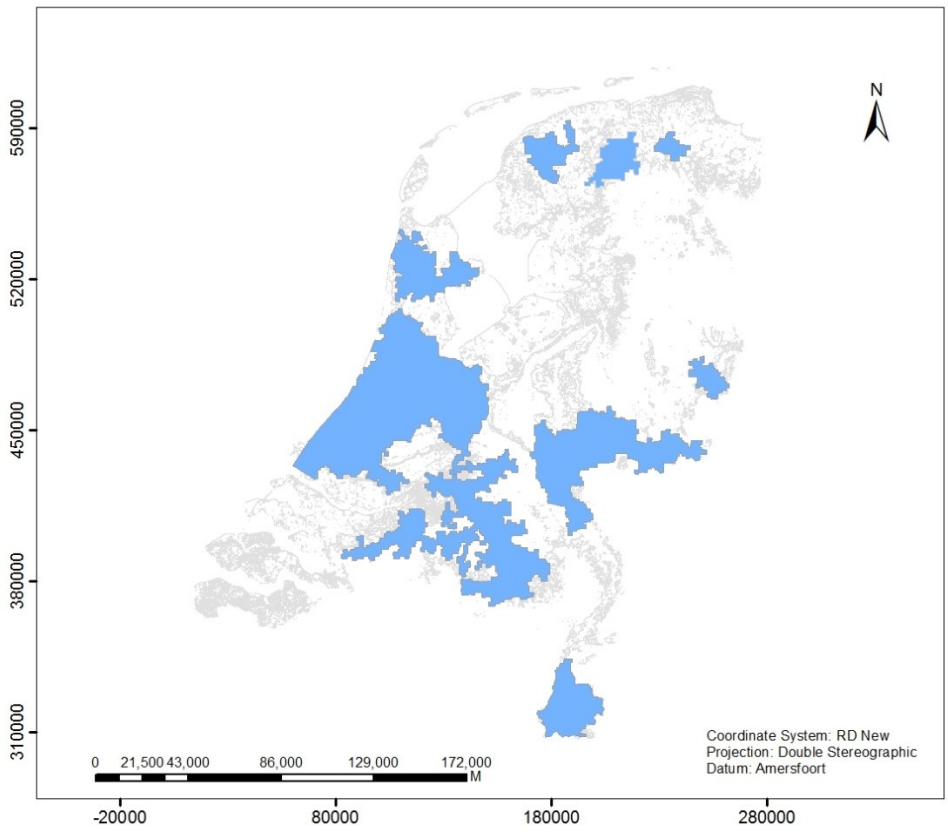


Figure 4.5: Nine largest polygons converted from extracted raster identified by percolation method

4.3.2. Temporal analysis

When plotting the data in time, we create an epidemic curve. This curve can be used to identify the number of occurrence of cases in an epidemic, and illustrate the change of frequency of a specific outbreak for a given period of time. The graphs in Figure (4.4) are epidemic curves of pertussis plotted over time for three years. To construct these epidemic curves, the number of disease cases occurring during a given month in each municipality was counted and plotted over monthly time intervals. The key information from these plots is the volume of disease cases in the epidemics and provides us a visual sense of how the disease changed over time. These plots also show the peak of the epidemic and we can say year 1996 is an epidemic year, which has the highest peak disease cases during October.

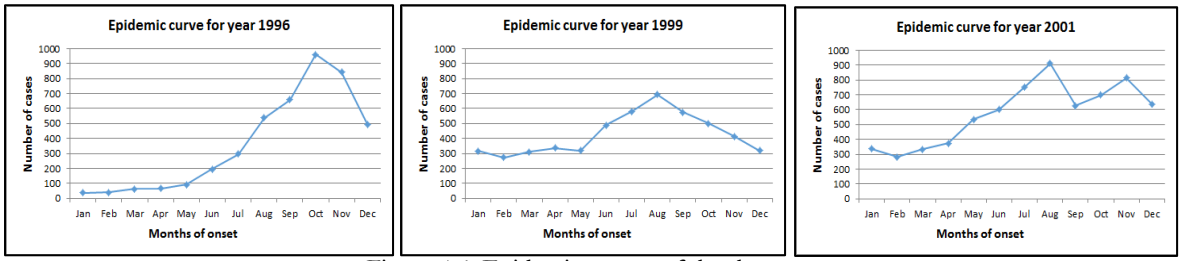


Figure 4.6: Epidemic curves of the three years

When we look at these epidemic curves, in 1999 and 2001 there is far more endemic pertussis (in non-epidemic months like January-April) compared to 1996. The volume of the area under the curve is higher for year 1999 and 2001. And we can see that none of the curves ends at the end of the years, so the outbreak of 1996 continues in 1997, the outbreak of 1999 seems to continue in 2000 and the outbreak of 2001 continues in 2002. We can also see that in 2001 there is a double peak (typical for pertussis) but this

peak cannot be seen in the other years. The total number of disease cases is larger in both 1999 and 2001. But the absolute high peak 962 is in 1996 during October.

4.3.3. Comparison of different mobility data results

This section compares the results of the three methods used as a mobility data in this research. Figure (4.7) shows the results of the three methods used as a mobility data. The left one is the percolation method which displayed the first nine largest urban zones identified by using a road network intersection points. The second (middle) one is for the urban agglomerations which are polygons found by digitizing connected urban systems. The last (right) one is for the commuting distance which shows the grouping of urban zones using a commuting distance of 30kms identified by drawing a buffer of 30kms from the centre of the twelve largest municipalities.

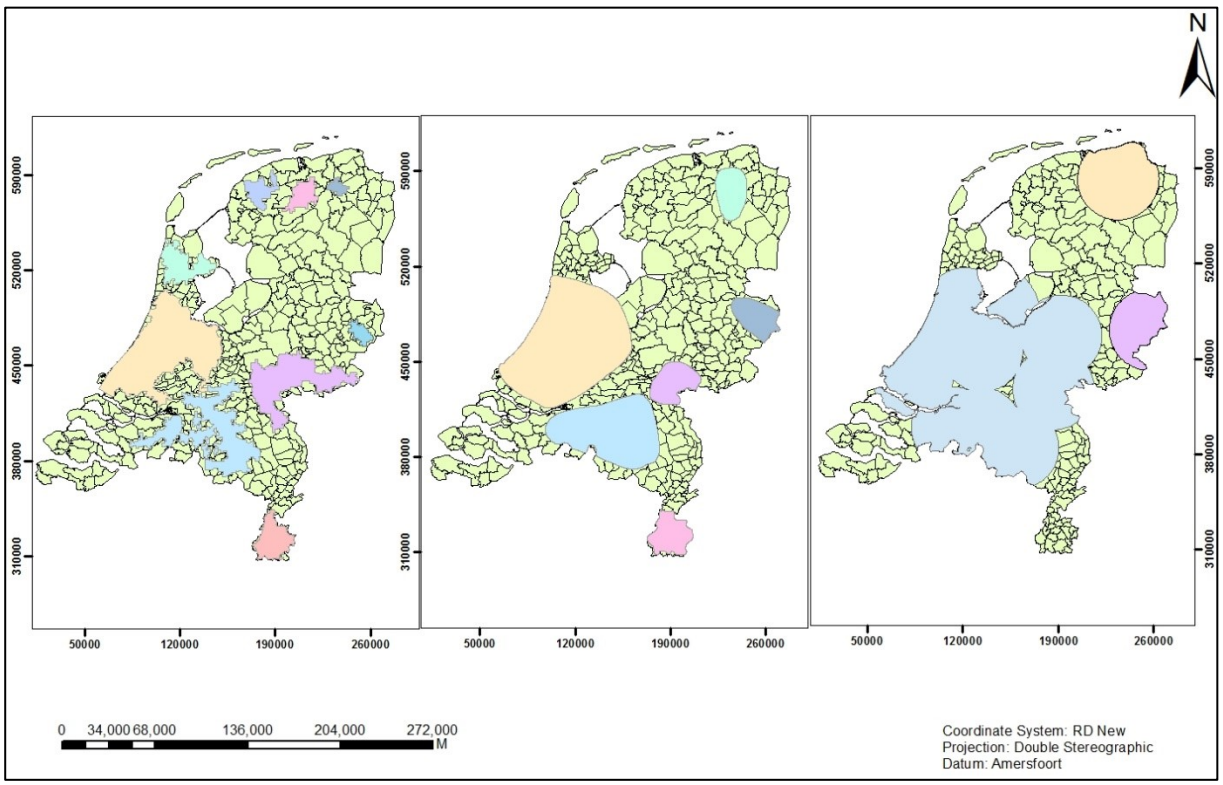


Figure 4.7: Shows identified urban zones based on percolation, urban agglomerations and commuting distance methods from left to right

All the three methods identified similar urban zones and most of the identified urban zones are overlapped for all the three methods. Figure (4.8) presents the union of urban zones identified by the three methods (Percolation method, urban agglomerations and commuting distance).

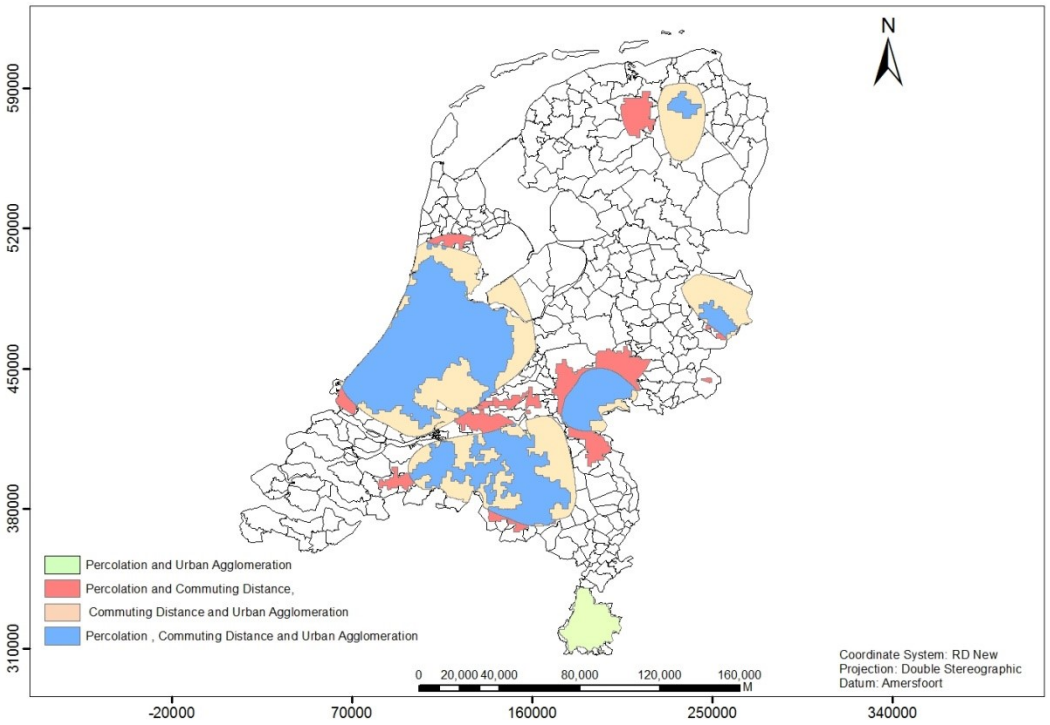


Figure 4.8: Areas identified by all the three urban zone identification methods

To select the appropriate method the characteristics of each method was investigated and the tables below show the characteristics of the areas identified in each selected urban zone for the three methods. Results for the percolation method are shown in Table 1. This method identifies six urban zones with a total population above the CCS and three zones with a total population under the CCS. Some of the zones that exceed the CCS do not have a large city (North_Randstad Holland and Zuid-Limburg). This means that these areas would not have been identified when using the municipalities as the spatial aggregation level of the analysis. There are also urban zones with multiple cities (South_Brabantstad and Randstad Holland). These characteristics of the urban zone were used for comparison of this method with the other two methods.

Table 1: Urban zones identified via the percolation method

Name	Population size	Area (km2)	CCS	Number of large cities
Near_Groningen-Assen	49250	351	0	0
East_Groningen-Assen	123060	350	0	0
Groningen-Assen	168355	116	0	1
Twente	255970	190	1	1
North_Randstad Holland	402015	691	1	0
Zuid-Limburg	638490	609	1	0
Arnhem-Nijmegen	896295	1749	1	1
Brabantstad	1661450	1552	1	1
Randstad Holland	5363450	3607	1	4

In table 2 we see the results for the urban agglomeration method. This method identifies six urban zones and all of the urban zones have a total population above the CCS. There is only one urban zone (Zuid-Limburg) which has no large city. Three urban zones (Groningen-Assen, Twente and Arnhem-Nijmegen) have only one large city. The other two largest urban zones (Brabantstad and Randstad Holland) have more number of large cities. Most of the large cities are included in these two urban zones. For this method all of the urban zones are meeting the CCS and almost all except one urban zone (Zuid-Limburg) a have large city.

Table 2: Urban agglomerations according to Ministerie van VROM (2004)

Name	Population size	Area (km2)	CCS	Number of large cities
Groningen-Assen	356455	727	1	1
Twente	428075	728	1	1
Arnhem-Nijmegen	566640	785	1	1
Zuid-Limburg	636295	721	1	0
Brabantstad	1776185	3067	1	3
Randstad Holland	6320005	5884	1	5

In Table 3 we see the results of the commuting distance method. In this method we identify three urban zones and all of them have a population size above the CCS. In addition, all zones have a large city. Using the commuting distance method most of the urban agglomerations (Randstad_Holland, Arnhem-Nijmegen and Brabantstad) are overlapping and grouped into one urban zone. This urban zone contains the ten largest cities which is the highest number of large cities when we compare them with the other urban zones.

Table 3: Urban zones identified via the commuting distance method

Name	Population size	Area (km^2)	CCS	Number of large cities
Enschede	587710	2642	1	1
Groningen	645160	1546	1	1
Randstad_Holland, Arnhem-Nijmegen, and Brabantstad	10643470	15848	1	10

The methods identified similar urban zones and the amount of spatial overlaps was calculated. The amount of overlap of areas between percolation method and urban systems method is 72%, overlap in areas between percolation method and commuting distance method is 62%, between the commuting distance method and the urban systems method the overlap is 58% and the overlap between all three methods is 41%. The larger overlap is identified between the percolation method and the urban systems method. Figure (4.8) shows the area of overlaps between all the three methods. Percolation method leads to a larger number of urban zones (9 urban zones) compared to the other two methods, the commuting distance method leads to the largest urban zones in area. This largest urban zone in the commuting

distance method leads to the largest number of large cities (contain 10 large cities). For the percolation method, 6 urban zones were identified which meet the CCS, for the urban agglomeration method also 6 urban zones are meeting the CCS and for commuting distance 3 urban zone are meeting the CCS. Even though the percolation method and urban agglomerations have the same number of urban zones which meet the CCS, for the urban agglomerations the polygons may not be exact since no official boundary of urban systems exists, and the identified zones include open spaces in between the zones that are potentially non-urban. For the commuting distance, most of the large city zones are overlapping, because cities in the Netherlands are very close together. Therefore, based on the justifications given above the percolation method is that the chosen method was used for the analysis of disease diffusion in relation to mobility. The six urban zones which meet the CCS should be used for the analysis.

4.4. Analysis of disease diffusion

4.4.1. Hierarchical diffusion

Diffusion of disease originates from one place and spread out to a new location and then at a later period fades away at a certain place. One way to identify the source and extinction of disease diffusion is by measuring hierarchy of diffusion based on the number and duration of fadeouts.

In this study, the disease total fadeouts and mean duration of fadeouts were calculated from disease cases by writing a python script. These data were anticipated for the analysis of hierarchical diffusion at two different scales, i.e. at municipality level and at urban zone level (within urban zone and between urban zones) on the chosen mobility method.

To investigate the hierarchical diffusion between and within urban zones, population size in relation to the total number of fadeouts was displayed using a table and graphical representations. At urban zone level (between urban zones) the disease persists throughout the epidemic years in almost all the selected urban zones and there are no fadeouts in the urban zones. At this spatial scale, there are disease cases at least in one municipality during the epidemic years in each urban zone. Therefore, no fadeouts are found at the urban zone spatial aggregation level. The table (4) has given below shows the total number of fadeouts in relation to the population size between urban zones for the three epidemic years. From the table, one can see that there is no fadeout at urban zones spatial aggregation level; Only Twente has two fadeouts in the year 1996. All the urban zones meet the CCS and the disease persists in all the urban zones.

Table 4: Population size relation to total number of fadeouts for year 1996, 1999 and 2001

Name of urban zone	Population size	Total number of fadeouts		
		1996	1999	2001
Twente	255970	2	0	0
North_Randstad Holland	402015	0	0	0
South of Brabantstad	628065	0	0	0
Zuid-Limburg	638490	0	0	0
Arnhem-Nijmegen	896295	0	0	0
Brabantstad	1033385	0	0	0
Randstad Holland	5363450	0	0	0

The results given in Figures (4.9-4.13) show the hierarchical diffusion at the municipality scale (within urban zones) spatial aggregation level using the total number of fadeouts in relation to population size. In this step urban zones with a small number of municipalities were not used for further analysis. The results of the analysis were displayed using graphical representations, scatter plots and by fitting a non-linear regression model. The model was fitted to draw a trend line between the data and to show the existence of decelerating curvilinear trend between disease fadeouts and population size. The initial values of the fitted model were determined using statistical parameters (Mean, standard deviation and correlation between the total number of fadeouts and population size) of the data. The population size was changed as a power of 105 to match the scale between x and y (population size and total number of fadeouts) during plotting. The non-linear model (curve) was a simple logarithmic function fitted between population size and total number of fadeouts. This model was chosen after considering the fitting trend of different functions (linear functions, power function, and least square exponential functions) by looking at the trend of the plotted data. The logarithmic function gives a better fitting curve for the regression of the total number of fadeouts on population size. The following Figures illustrate the results of hierarchical diffusion and the fitted curve in the selected urban zones for three years (1996, 1999 and 2001).

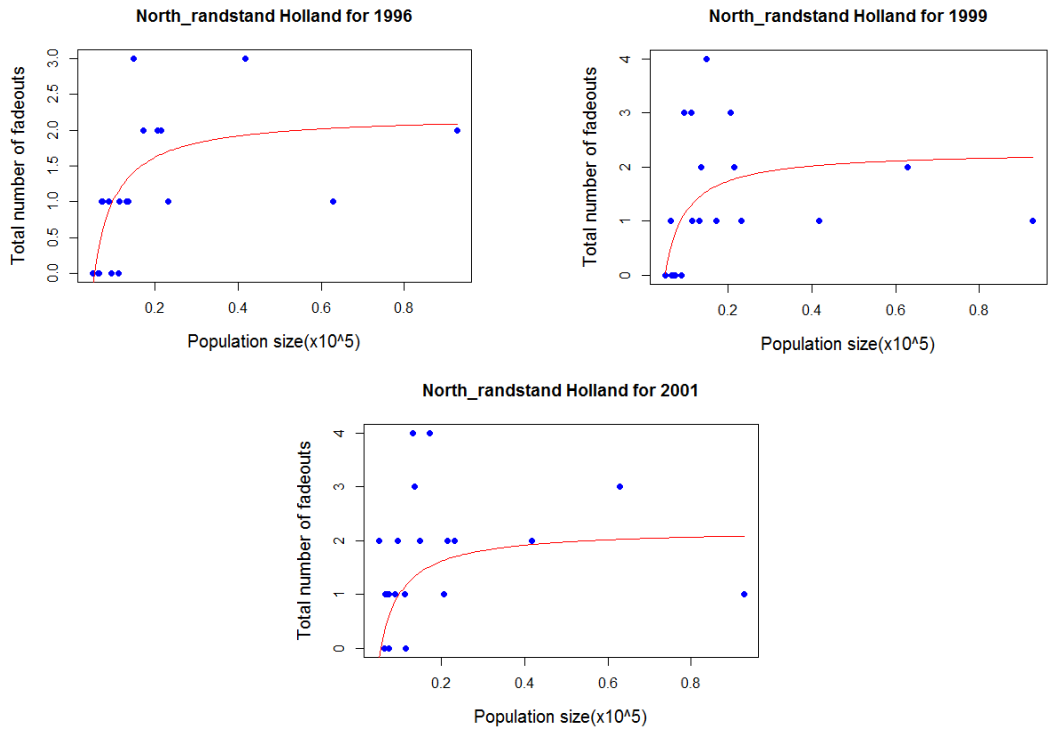


Figure 4.9: Hierarchical diffusion in North of Randstad Holland urban zone using the total number of fadeouts

The result in Figure (4.9) show that there is one large municipality in the year 1996 which has a lower total number of fadeouts and two large cities with a higher total number of fadeouts. In year 1999 two larger cities have lower total of fadeouts and one large city with a higher number of fadeouts. There is also one larger city which has a lower total number of fadeouts in the year 2001 and this year has one large city with a higher total number of fadeouts. The trend line (fitted logarithmic curve) shows an increasing function in this urban zone for all years. This indicates that the total number of fadeouts is higher in cities from very small to small cities (there was a frequent extinction of the disease in small cities). In conclusion, there was no hierarchical diffusion in this urban zone.

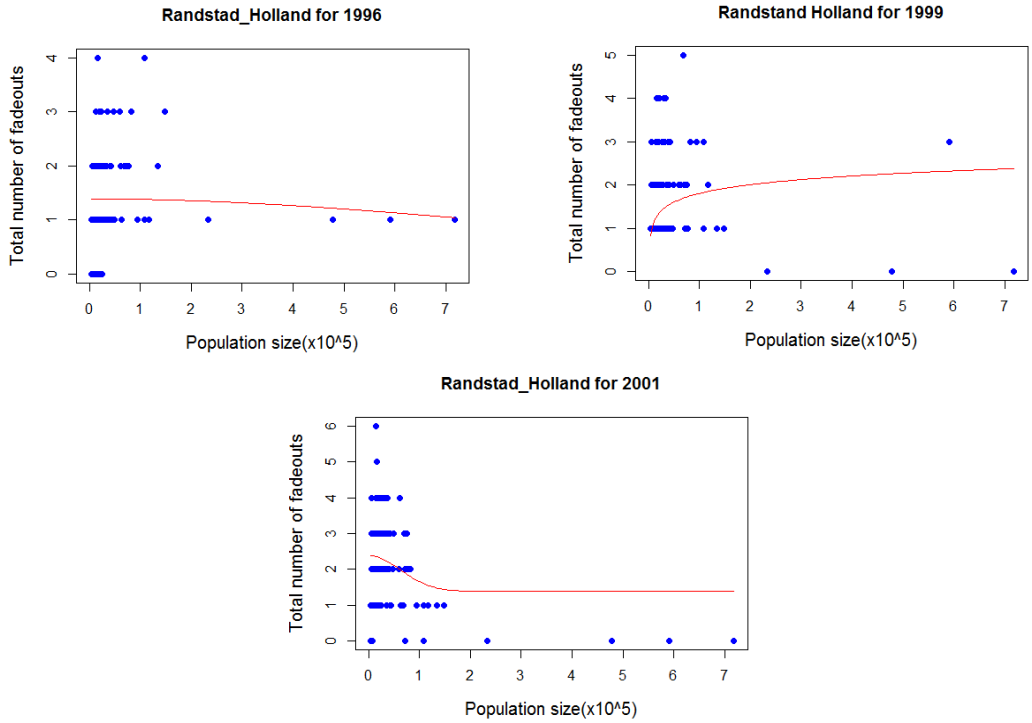


Figure 4.10: Hierarchical diffusion in a Randstad Holland urban zone using the total number of fadeouts

Figure (4.10) illustrates Randstad Holland in the three epidemic years. In all years, this urban zone shows a trend which indicates it has an increasing number of fadeouts when population size increases for population less than 250,000. For population sizes above 250,000, trends of hierarchies of diffusion with lower number of fadeouts were observed. In year 1999, there is an exceptional large city with a higher number of fadeouts. In year 2001, there were no fadeouts in large cities for population sizes above 250,000. The year 1999 also showed similar to year the 2001 except one large city. This is an indication that the disease persist longer in larger cities and this large cities are the source of the disease diffusion. In this urban zone, large cities are the source of the diffusion of the disease because the disease persists longer in large cities (larger cities have no or a lower total number of fadeout). Overall, this urban zone shows the hierarchy of diffusion, indicates that disease diffuse from larger to smaller cities.

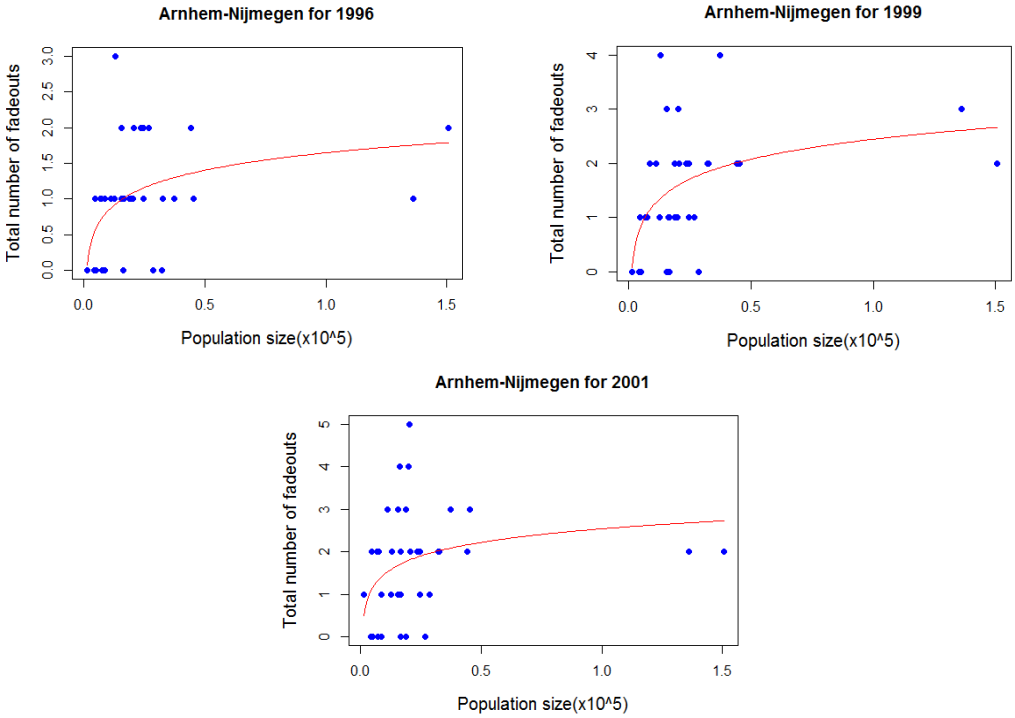


Figure 4.11: Hierarchical diffusion in an Arnhem-Nijmegen urban zone using the total number of fadeouts

In Figure (4.11) the hierarchical diffusion of Arnhem-Nijmegen is displayed for the three years. This figure indicates that all the years in this urban zone show an increasing function, meaning that number of fadeouts increases when the population size increases from very small to small population sizes (for population sizes below 50,000). But, when we look at the general plot, no clear pattern of hierarchy was observed in this urban zone.

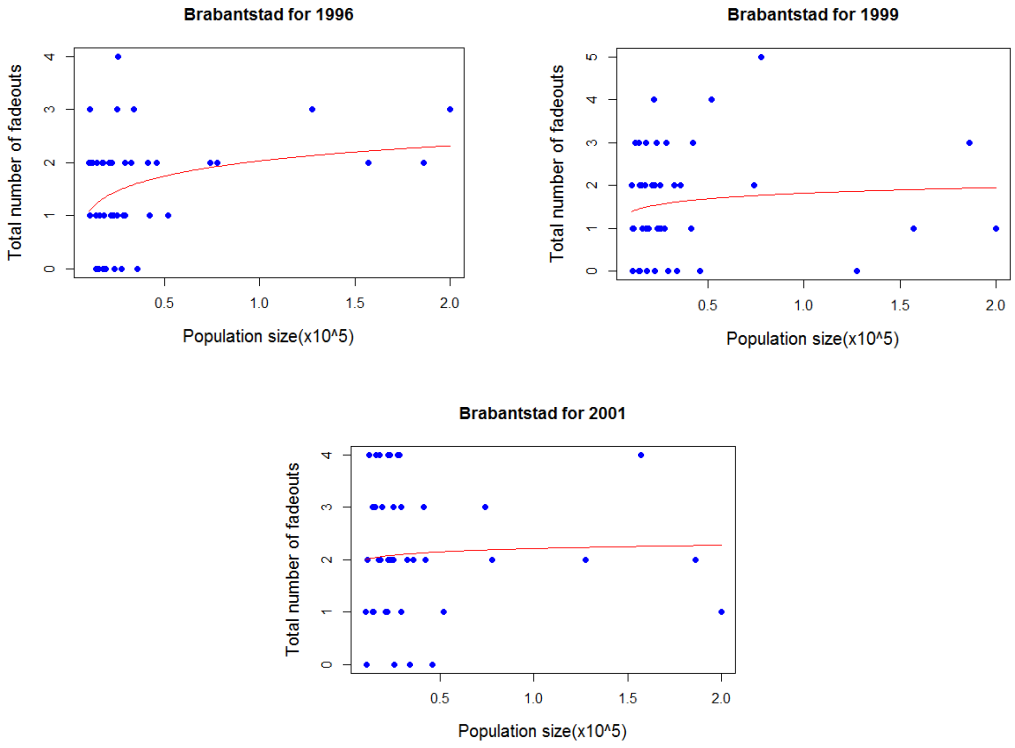


Figure 4.12: Hierarchical diffusion in a Brabantstad urban zone using the total number of fadeouts

In Figure (4.12) total number of fadeouts in relation to population size was investigated in the Brabantstad urban zone. This figure indicates that most large cities have a higher total number of fadeouts. There are also large cities with lower number of fadeouts. This urban zone did not show a clear trend in the hierarchy of diffusion during all the three epidemic years. They show both increasing and decreasing patterns of total number of fadeouts.

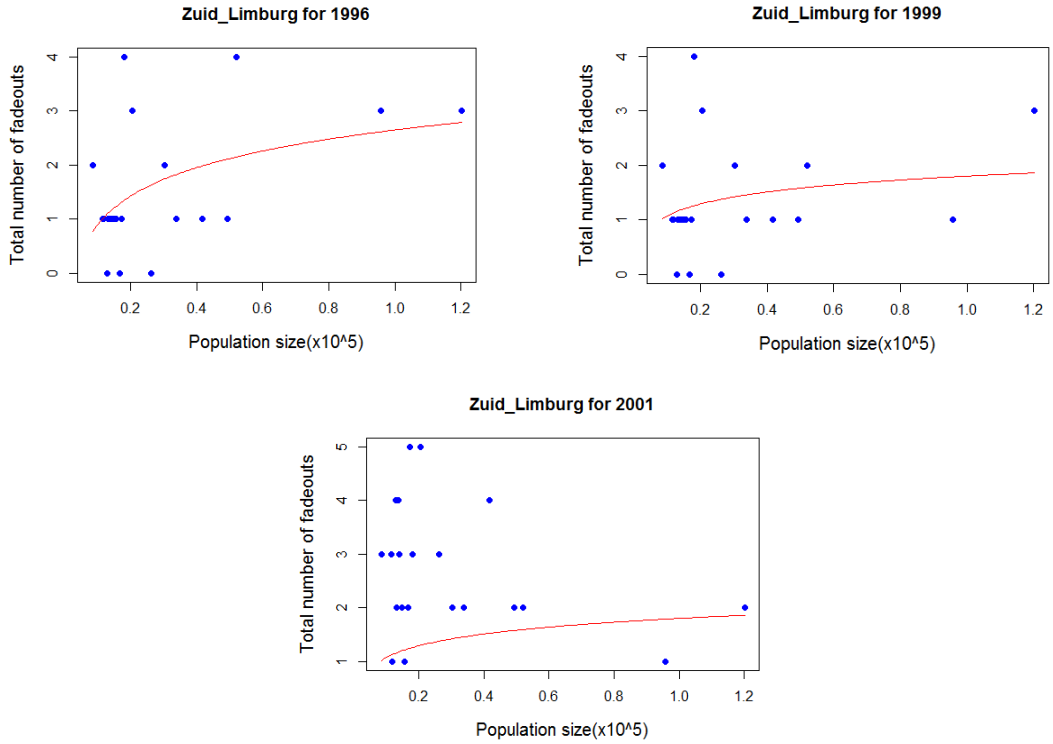


Figure 4.13: Hierarchical diffusion in a Zuid_Limburg urban zone using the total number of fadeouts

In Figure (4.13) the total number of fadeouts in relation to the population size for the Zuid_Limburg urban zone was illustrated for three years. Results show that in this urban zone, the year 1996 and 1999 show similarities in their diffusion patterns. They show an increasing function for smaller sizes and after population size reaches around 60,000 they show a constant number of fadeouts but no clear patterns are shown in both years. For instance, in the year 1996 two larger cities have a higher total number of fadeouts and there is one larger population size with higher number of fadeouts in the year 1999. In year the 2001, there are two large cities which have lower number of fadeouts. This trend indicates, in the year 2001 there was a frequent fadeout in cities with a smaller number of population sizes and less total number of fadeouts in cities with larger population sizes. This urban zone shows the hierarchy of diffusion in this year (year 2001).

Although, using the total number of fadeouts in relation to population size to analyse hierarchical diffusion patterns of pertussis is not showing clear hierarchies of diffusion in all urban zones, most of the fitted curve shows an increasing function, which shows that cities with smaller population sizes were subject to frequent number of fadeouts. And also we can observe that, when the population size increases, there were increasing in frequency of fadeouts for very small to small population sizes (for a population size of below 250,000). The Randstad Holland urban zone shows a different pattern, there were no fadeouts in larger cities and more frequent fadeouts in smaller cities. Overall, only Randstad Holland is large enough to show hierarchical patterns based on the number of fadeouts.

In conclusion, from all the plots we can also observe that most of the smaller cities have a rare number of fadeouts and large cities experienced a number of short fadeouts.

The next step during the analysis of measuring hierarchies of diffusion was using mean duration of fadeouts in relation to population size. In this step, all the methods and parameters of the model were applied in a similar way as the analysis of hierarchies using the total number of fadeouts at two aggregation levels (within urban zones and between urban zones) for three epidemic years. The only two differences are: the fitted model for this data was as a power function and the disease data were analysed as a variable of mean duration of the fadeouts.

The results of measuring hierarchy of diffusion using mean duration of fadeouts in relation to population size at between urban zones spatial aggregations are shown in Table (5). The results show similar results as analysis using number of fadeouts between urban zone spatial aggregation levels. All the urban zones meet the CCS and the disease persists throughout the epidemic years. No duration of fadeouts at urban zone spatial aggregation level was observed. The only Twente urban zone has three months mean duration of fadeouts.

Table 5: Population size relation to mean duration of fadeouts for year 1996, 1999 and 2001

Name of urban zone	Population size	Mean duration of fadeouts		
		1996	1999	2001
Twente	255970	3	0	0
North_Randstad Holland	402015	0	0	0
South of Brabantstad	628065	0	0	0
Zuid-Limburg	638490	0	0	0
Arnhem-Nijmegen	896295	0	0	0
Brabantstad	1033385	0	0	0
Randstad Holland	5363450	0	0	0

Next the results of measuring hierarchy of diffusion using mean duration of fadeouts in relation to population size at the municipality scale (within urban zone) spatial aggregation level was presented in Figures (4.14-4.18).

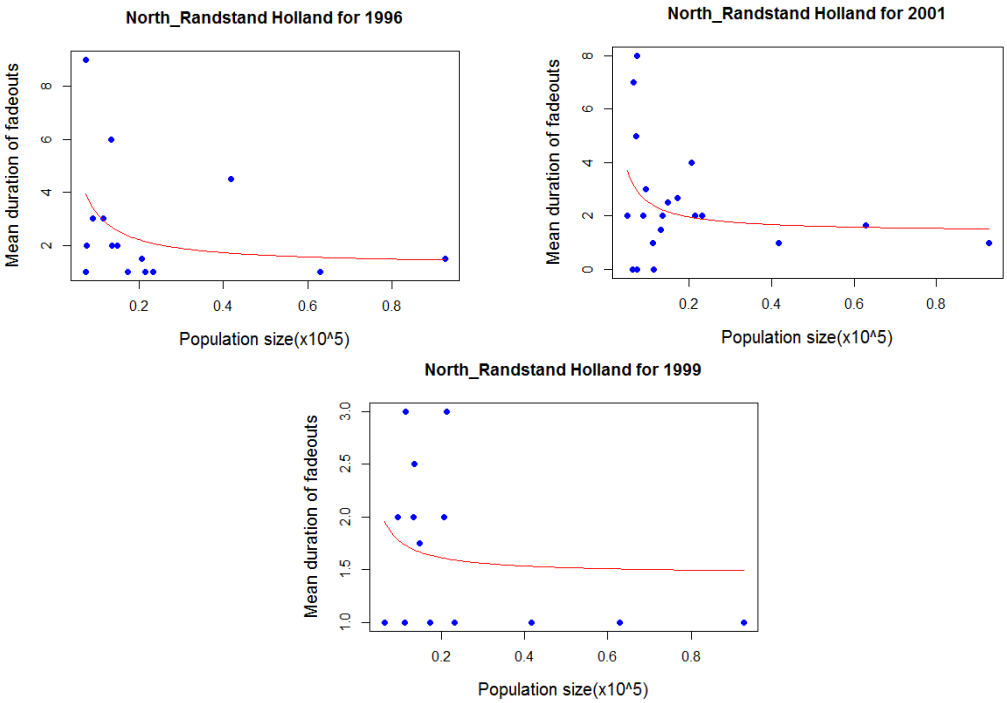


Figure 4.14: Hierarchical diffusion in North of Randstad Holland urban zone using mean duration of fadeouts

In Figure (4.14) hierarchical diffusion in North of Randstad Holland urban zone was presented using mean duration of fadeouts in relation to population size for the three years. In this urban zone all the years shows that long-lasting fadeouts (longer duration of fadeouts) of the disease happened in the smaller cities and larger cities have a shorter fadeout, or more likely to maintain the disease. In this urban zone hierarchy of diffusion is shown clearly, meaning that larger cities were the source of the disease and it has a longer duration of fadeouts in smaller cities.

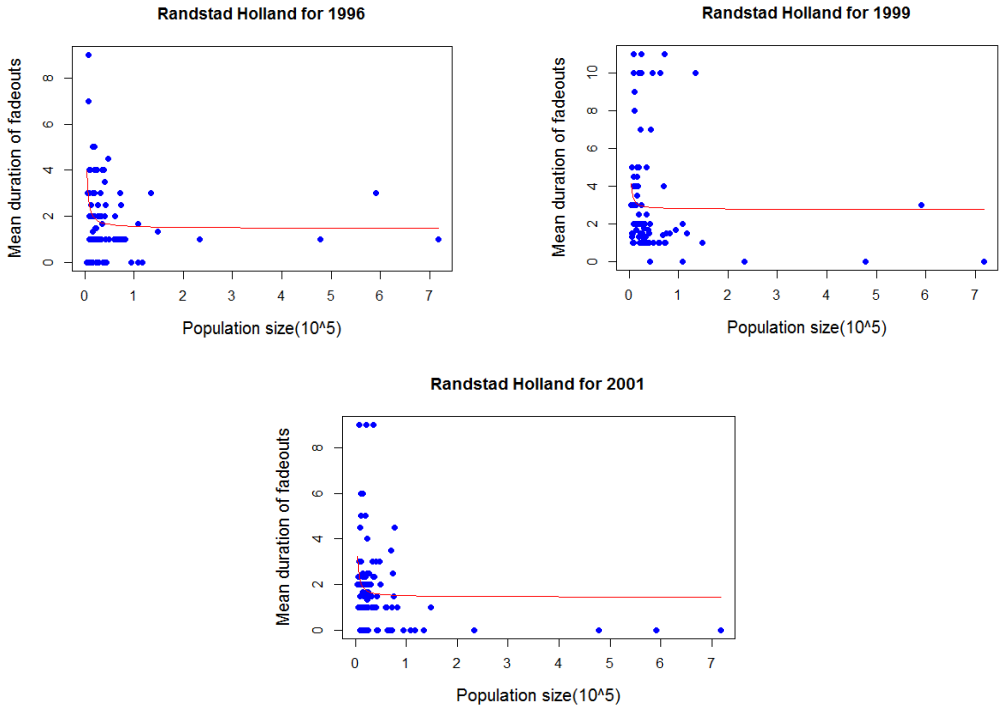


Figure 4.15: Hierarchical diffusion in a Randstad Holland urban zone using mean duration of fadeouts

In this Figure (4.15) hierarchical diffusion in a Randstad Holland urban zone using mean duration of fadeouts in relation to population size was illustrated for the three years. The results show that this urban zone also displayed longer fadeouts in small population sizes and shorter in larger population sizes. The only different trend shown in this urban zone was in the year 1996 and 1999; one large city has longer duration of fadeouts. Except this large city this urban zone has the hierarchy of diffusion patterns from larger cities to smaller cities. In the year 2001, the disease persists throughout the year in the larger cities (no fadeout in large cities). The disease persists longer in large cities and smaller cities are subject to longer duration of fadeouts. This is an indication that these large cities are the source of the disease.

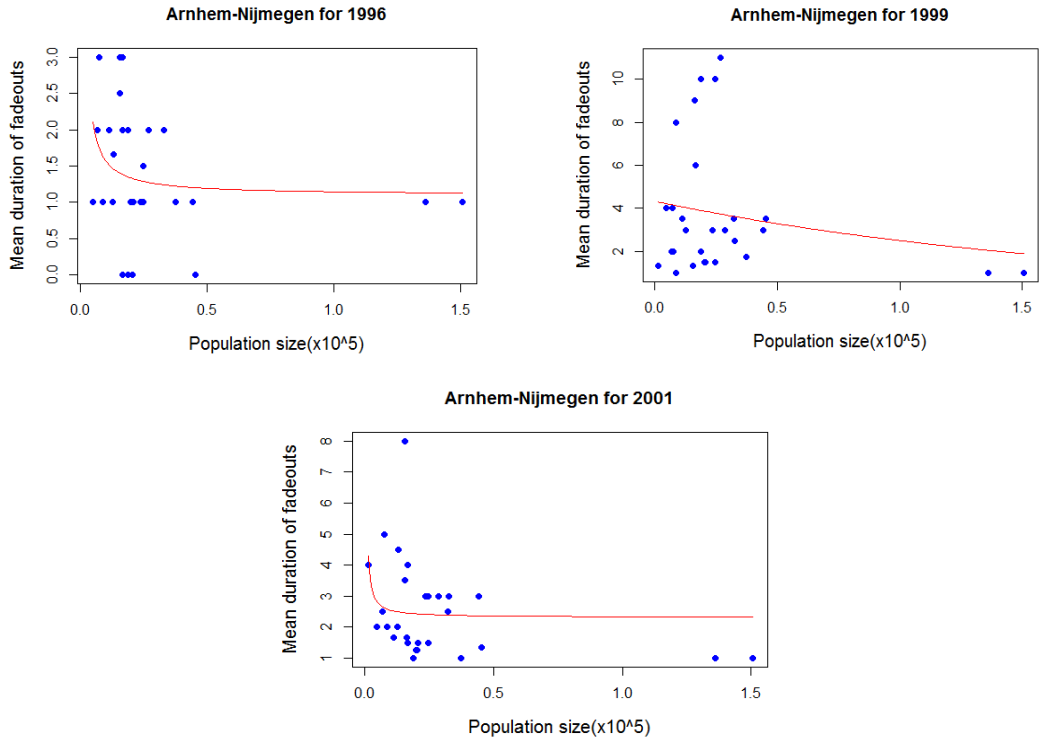


Figure 4.16: Hierarchical diffusion in an Arnhem-Nijmegen urban zone using mean duration of fadeouts

The result of Arnhem-Nijmegen urban zone using mean duration of fadeouts was presented in Figure (4.16) for the three years. This urban zone also shows a trend of the hierarchical diffusion pattern which displays longer fadeouts in small cities and shorter in larger cities, but, the year 1999 shows a slight difference in the diffusion pattern from the other two years (1996 and 2001). There was also a number of shorter duration of fadeouts in smaller cities in the year 1999. In general, this urban zone also shows hierarchies of diffusions which shows long lasting fadeouts in small cities and shorter duration of fadeouts in large cities.

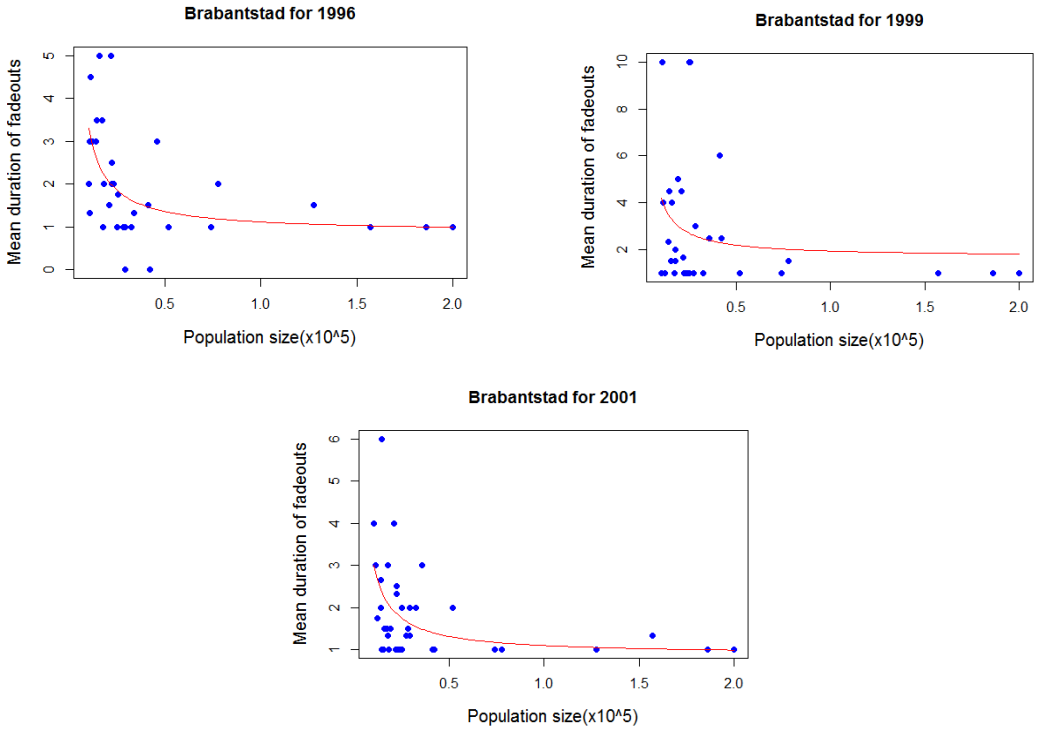


Figure 4.17: Hierarchical diffusion in a Brabantstad urban zone using mean duration of fadeouts

Figure (4.17) illustrates the result of Brabantstad urban zone using mean duration of fadeouts in relation to population size for the three years. This urban zone shows clear hierarchy of diffusion patterns which displays longer duration of fadeouts in small cities sizes and shorter duration of fadeouts in larger cities for all the three years. This indicates larger cities are the source of diffusion of the disease and the disease stay longer in large cities.

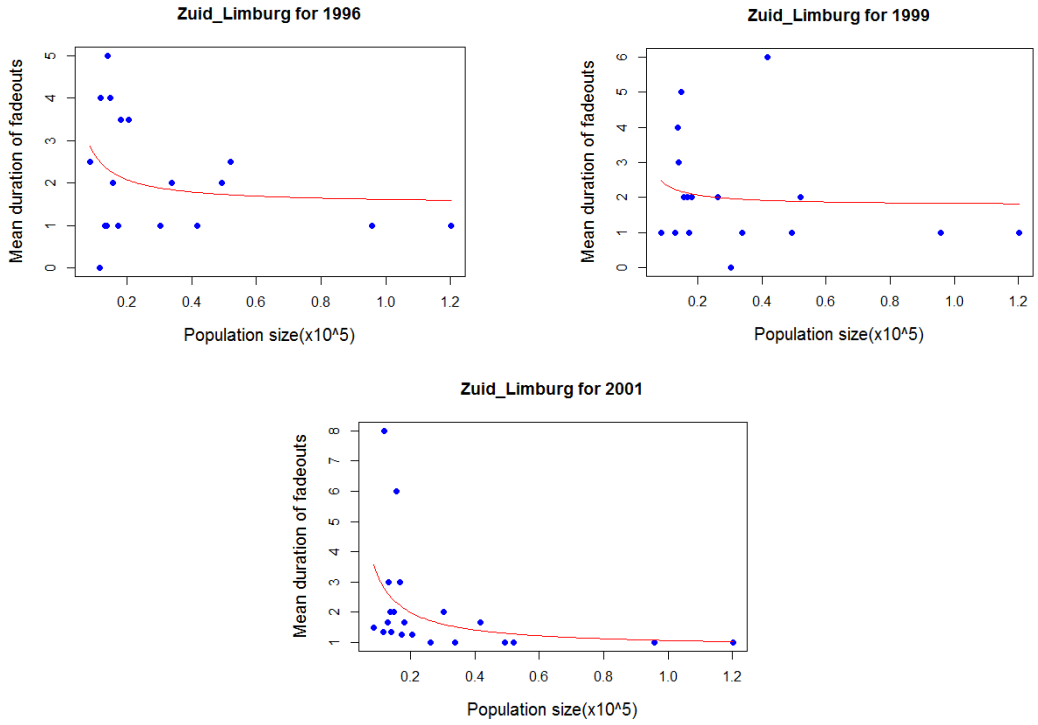


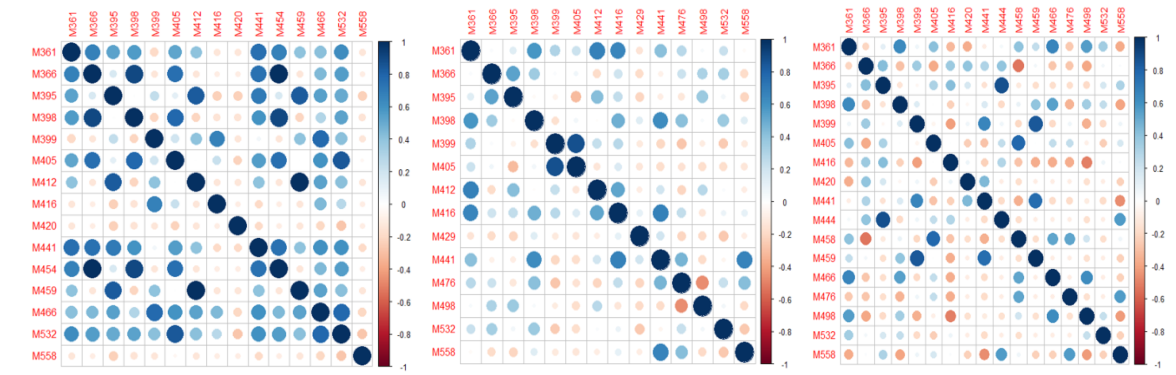
Figure 4.18: Hierarchical diffusion in a Zuid_Limburg urban zone using mean duration of fadeouts

The final plot of hierarchical diffusion was presented in Figure (4.18) which displays the hierarchical diffusion in a Zuid_Limburg urban zone using mean duration of fadeouts in relation to population size. This urban zone also shows a hierarchical diffusion of disease which indicates longer fadeouts in small cities and shorter in larger cities for all the three years. Similar to the above urban zones, larger cities are the source of the disease in this urban zone.

In general, at municipality spatial aggregation level, the graphical representation of the results indicates that long lasting fadeouts happened with smallest population sizes and, conversely, the largest population sizes were more subject to shorter fadeouts. In addition, from the graphical representations we can observe that the mean duration of fadeouts displays clear hierarchies of diffusion than the total number of fadeouts. Moreover, for large cities we found negative residuals from the regression of mean duration of fadeouts on the population size, indicates that these large cities have higher persistence of the disease. As a conclusion, larger cities were more likely to maintain the disease and these large cities are the source of the diffusion.

4.4.2. Result for synchrony

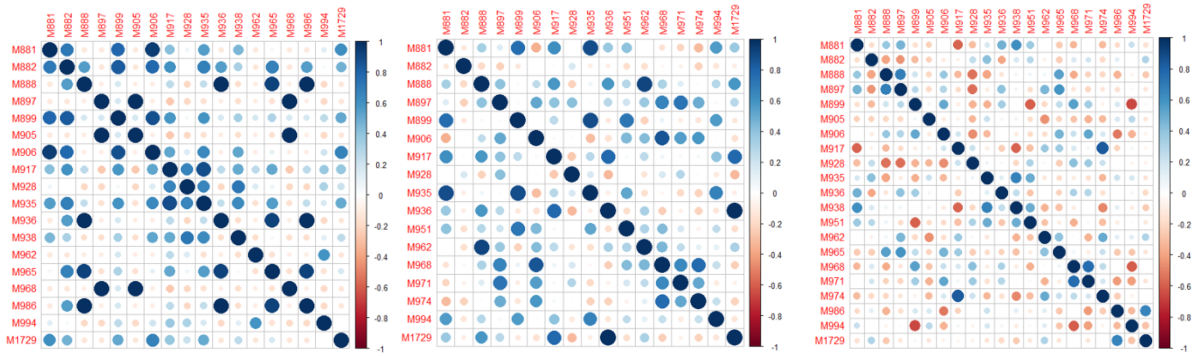
Results of synchrony of diffusion were analysed as described in the methods section (3.3.2), by using the frequency of the disease in each municipality and assessing based on pairwise Pearson correlation coefficients between municipalities within and between urban zones. The amount of synchrony was assessed by pairwise averaging to analyse synchrony in a region wide diffusion pattern and to quantify the synchrony statistically. The identified synchrony is presented below using graphical representations using the corrplot package in R. The plots given below present the patterns of synchrony at a municipality spatial aggregation level for the three years.



North_Randstad Holland 1996 North_Randstad Holland North_Randstad Holland 2001

Figure 4.19: Measuring of synchrony in North_Randstad Holland urban zone

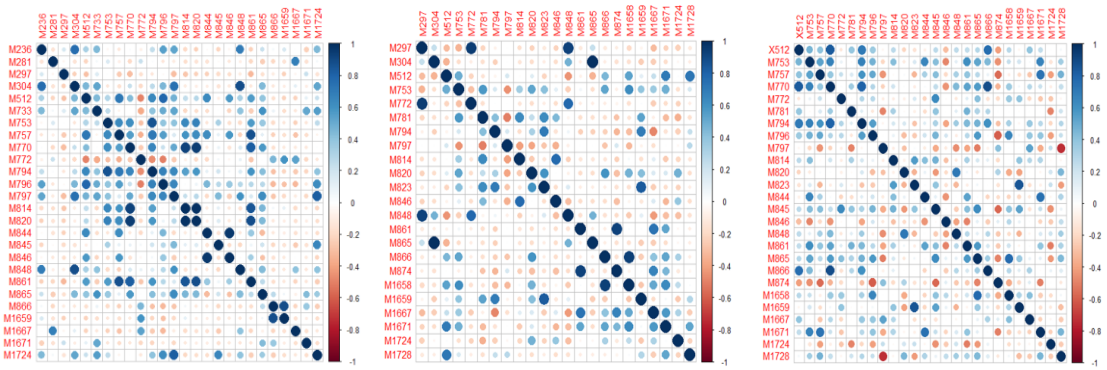
Figure (4.19) shows the relationships of disease frequency between municipalities in terms of Pearson correlation coefficients in the North_Randstad Holland urban zone. When examining calculated values of the correlation coefficients in this urban zone for the three years, the year 1996 shows relatively higher positive correlation coefficients. This also has higher pairwise average value (average= 0.232248). In year 1999, we see low correlation coefficient values (the average pairwise value for this year is 0.089138). The correlation coefficient in the year 2001 is also very low (average= 0.016071). In general this urban zone shows some degree of synchrony of diffusion in the year 1996.



Zuid-Limburg 1996 Zuid-Limburg 1999 Zuid-Limburg 2001

Figure 4.20: Measuring of synchrony in Zuid-Limburg urban zone

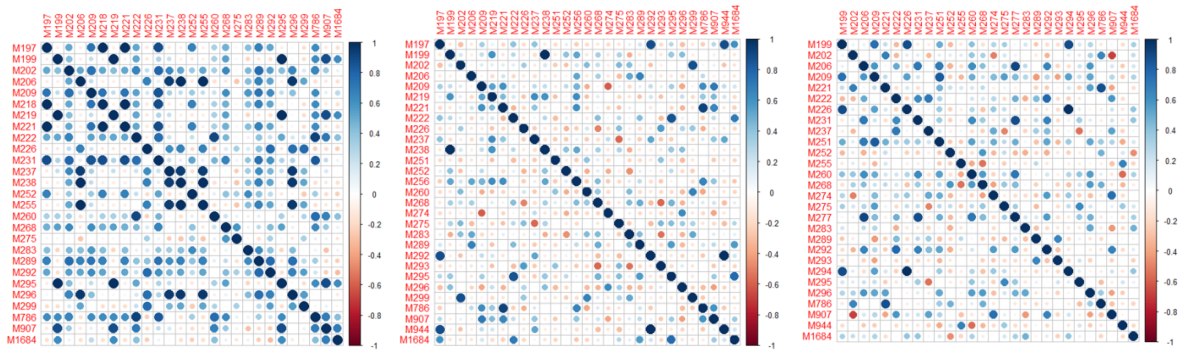
Figure (4.20) illustrates the result of spatial synchrony of disease frequencies between municipalities within Zuid-Limburg urban zone for three years. Clearly, all years showed a lower correlation in their synchrony of diffusion. The years 1996 and 1999 showed almost similar average pairwise correlation values (average for 1996=0.178709 and 1999=0.157639). In the year 2001 pairwise average correlations showed very low negative values (average=-0.00200). In this year the synchrony is different from the other urban zones also. It is very low and negative correlation. This means municipalities showed very low opposite patterns of synchrony of diffusion.



Brabantstad 1996 Brabantstad 1999 Brabantstad 2001

Figure 4.21: Measuring of synchrony in Brabantstad urban zone

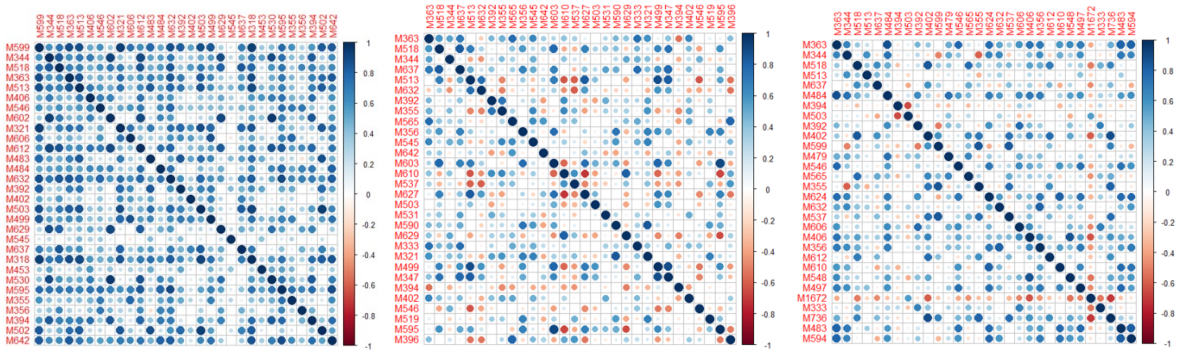
The plots of correlation coefficients of frequency disease in Brabantstad in Figure (4.21) indicate that, synchrony of diffusion is low in this urban zone for all years. The average pairwise correlations are (average 1996=0.09464, 1999=0.0132 and 2001= 0.081445) also lower in this urban zone. Similar to other urban zones year 1996 has relatively higher average pairwise values.



Arnhem-Nijmegen 1996 Arnhem-Nijmegen 1999 Arnhem-Nijmegen 2001

Figure 4.22: Measuring of synchrony in Arnhem-Nijmegen urban zone

Figure (4.22) showed the synchrony of the frequency of the disease between municipalities of Arnhem-Nijmegen urban zone for the three years. This urban zone showed higher synchrony of diffusion pattern in year 1996 (average = 0.264688). The year 1999 and 2001 showed very low synchrony of diffusion (average for 1999= 0.0273 and year 2001=0.0749). In general, in this urban zone synchrony of disease diffusion is low for all years.



Randstad Holland 1996 Randstad Holland 1999 Randstad Holland 2001

Figure 4.23: Measuring of synchrony in Randstad Holland urban zone

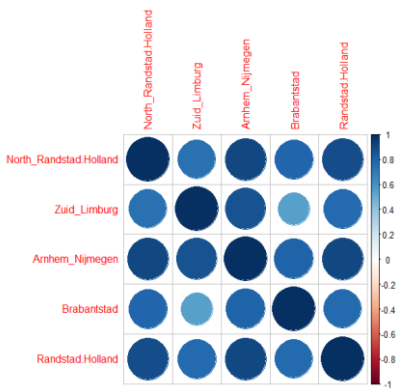
In Figure (4.23) synchrony of diffusion in the Randstad Holland urban zone was displayed and the calculated correlation coefficient clearly showed synchrony in the year 1996. The average correlation value is also higher for this year (average=0.503855). This pairwise average value in the year 1996, was the highest correlation value of this spatial scale when we compare it with the other urban zones. This highest correlation value may relate to the concept of synchrony between large cities because more number of large cities found in this urban zone. The next higher average values of synchrony in this urban zone were shown in year 2001 (average =0.213698). In year 1999, average pairwise value is very low (average = 0.129053). Table (6) summarizes the measured synchrony in each urban zone for three years at a municipality aggregation level (within the urban zone).

Table 6: Synchrony for three epidemic years at municipality (within urban zone) spatial scale

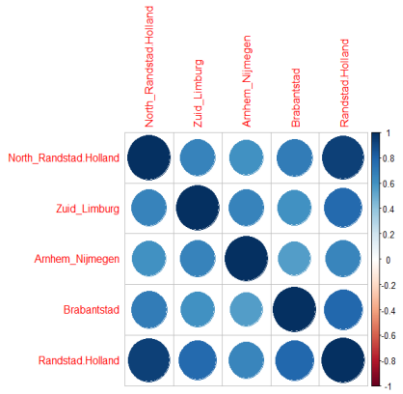
Urban zone name	1996	1999	2001
South of Brabantstad	0.2322482	0.0891375	0.0160707
Zuid-Limburg	0.1787087	0.1576391	-0.0020000
Arnhem-Nijmegen	0.2646880	0.0273000	0.0794000
Brabantstad	0.0946400	0.0132000	0.0814450
Randstad Holland	0.5038550	0.1290530	0.2136908

In general, there was a marked difference in the synchrony of disease frequencies of the three years. The Year 1996 showed relatively higher synchrony of diffusion between municipalities in all urban zones. Especially in Randstad Holland this year showed higher synchrony of diffusion. This higher synchrony in the year 1996 is due to that the disease was moving faster during this year and the epidemic outbreak was shorter for this year.

The next step during measuring synchrony was analysing synchrony at urban zone spatial aggregation level. In order to do this analysis all the methods were applied in a similar way at municipality spatial scale level, the only difference was the scale. We measure the synchrony of the total disease frequencies of the one urban zone with the total frequencies of the other urban zones and we found the results presented in Figure (4.24). From the figure, we can see that the plots show higher positive correlation coefficients for the three years. The pairwise average values are also higher at this spatial scale.



Synchrony of all urban zones in year 1996



Synchrony of all urban zones in year 1999

Synchrony of all urban zones in year 2001

Figure 4.24: Synchrony of diffusion at urban zone spatial aggregation level for three years

The correlation coefficients between urban zones showed an overall higher correlation in all year. Especially year 1996 showed the highest average correlation coefficient (0.802623). The year 1999 also showed next highest pairwise average (0.693528) correlation between urban zones. The year 2001 also showed medium pairwise average (0.571831) between urban zones. The urban zone varies in disease frequencies, but they are spatially synchronized, meaning that the average values of moving up and down together are higher for the three years. This implies that there is a synchrony of diffusion of disease between urban zones.

Overall, the analysis revealed that, the synchrony between diseases frequencies were higher at urban zones spatial aggregation level than at municipality level.

Additionally, the results of average pairwise synchrony between urban zones were analysed based on the distance from the largest urban zone. This analysis was performed by calculating the pairwise average values of each urban zone pairing with Randstad Holland and we compare these values to see the impact of the separation distance between urban zones. The results of these calculated average pairwise values are presented in Table (6) and the comparison was done based on this calculated average pairwise correlation values. For instance, in the year 1996, we compare the average value of North_Randstad which is the nearest urban zone for Randstad Holland with Arnhem_Nijmegen which is farther than North_Randstad from Randstad Holland. And we found higher average value in Arnhem_Nijmegen (North_Randstad= 0.8894164, Arnhem_Nijmegen=0.9056728). For the year 1999, we compare Arnhem_Nijmegen (average=0.6569353) and Zuid_Limburg (average=0.7753245) which is the farthest from the Randstad Holland and we found higher average pairwise value for Zuid_Limburg with Randstad Holland. In the year 2001 we found the lowest average pairwise correlation value for North_Randstad (average=0.4814358) which is the nearest to Randstad Holland. Therefore, from these findings, we can conclude that the synchrony of diffusion doesn't depend on the separation distance between the urban zones.

Table 7: Pairwise average values between Randstad with the other urban zones.

	North_Randstad	Brabantstad	Arnhem_Nijmegen	Zuid_Limburg	Year
Randstad	0.8894164	0.7740873	0.9056728	0.7791792	1996
Randstad	0.933614	0.7800073	0.6569353	0.7753245	1999
Randstad	0.4814358	0.8511065	0.6684875	0.6560271	2001

Moreover, during the analysis of synchrony, we carried out a measure of synchrony diffusion using railway network connectivity. The railway mobility data were used to assess synchrony at municipality spatial aggregation levels using the frequency of the disease for the three years. During this analysis, first 330 municipalities connected by railway network were selected and 30 largest municipalities were chosen from those 330 municipalities for ease of plotting in R scripting software. Map of these identified municipalities and their railway connectivity is presented below in Figure (4.25).

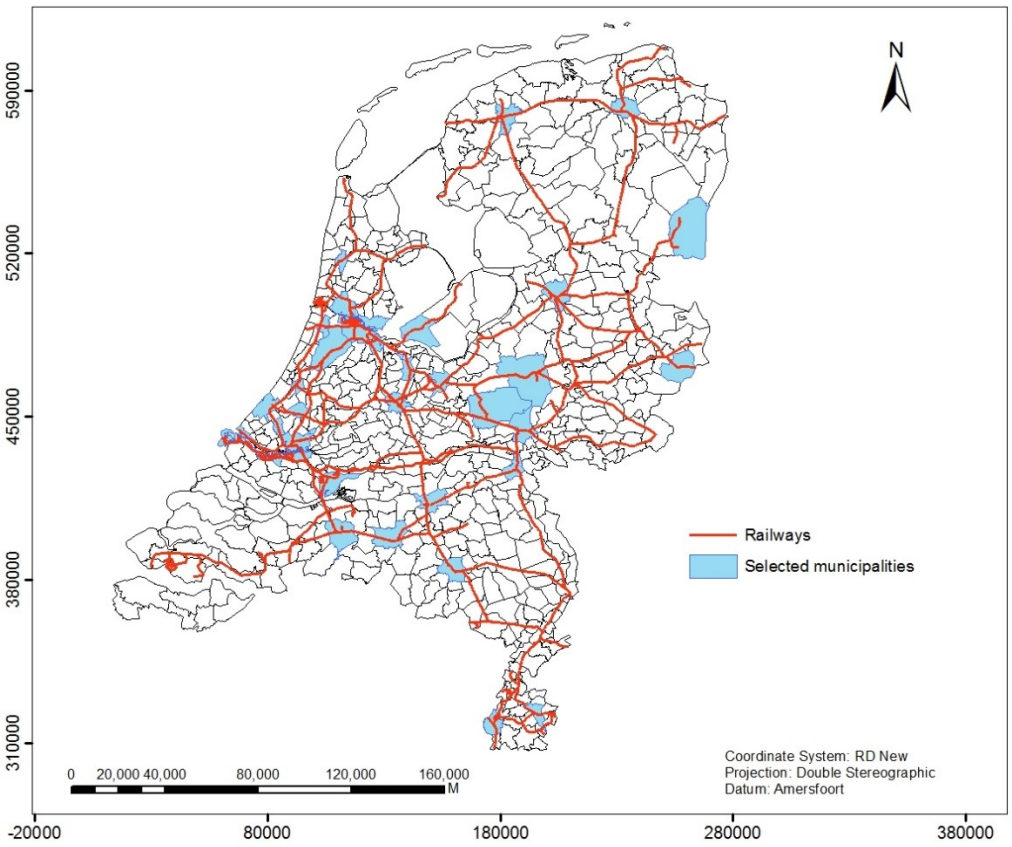
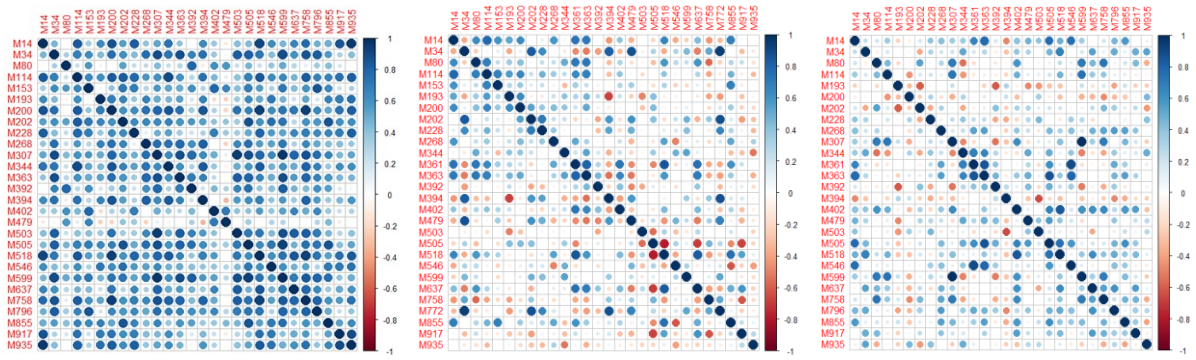


Figure 4.25: 30 selected municipalities and their railway connectivity

From those selected, 30 large municipalities, municipalities with a total sum of zero disease cases were removed and 28 municipalities were left for the analysis in each year. Having identified these large cities based on their railway connectivity, the synchrony of disease diffusion was characterized by their Pearson correlation coefficients and visualized using the corrplot package in R. Finally, we examined whether these cities were showing synchrony in their diffusion and the total amount of synchrony was quantified using pairwise averaging. The results of the synchrony of the selected 28 municipalities for the three years were illustrated in Figure (2.26).



1996 1999 2001

Figure 4.26: Measuring of synchrony in selected large cities connected by railway mobility data

The results of the correlation coefficient show that, these large municipalities clearly showed synchrony in the year 1996. The average correlation value is also higher for this year (average= 0.513605). Next the pairwise average values in the year 1999 were assessed and it shows very low average (average=0.0853337) value. In year 2001, average pairwise value is also very low (average = 0.148318).

In general, analysing synchrony for those large municipalities selected based on railway connectivity shows synchrony in the year 1996. The other two years shows a lower number of average pairwise correlations. Synchrony of diffusion based on railway connectivity was shown also only in the year 1996.

Furthermore, synchrony was measured by selecting municipalities using intercity train tracks. The results illustrate below in figures (4.27-4.30) shows the synchrony of disease diffusion of cities connected by intercity tracks. The identified synchrony was presented using graphical representations and quantified by pairwise averaging the Pearson correlation between those connected cities. The plots of those municipalities connected by intercity lines are based on municipality code. The reason plotting using municipality code is the disease data were provided based on municipality code.

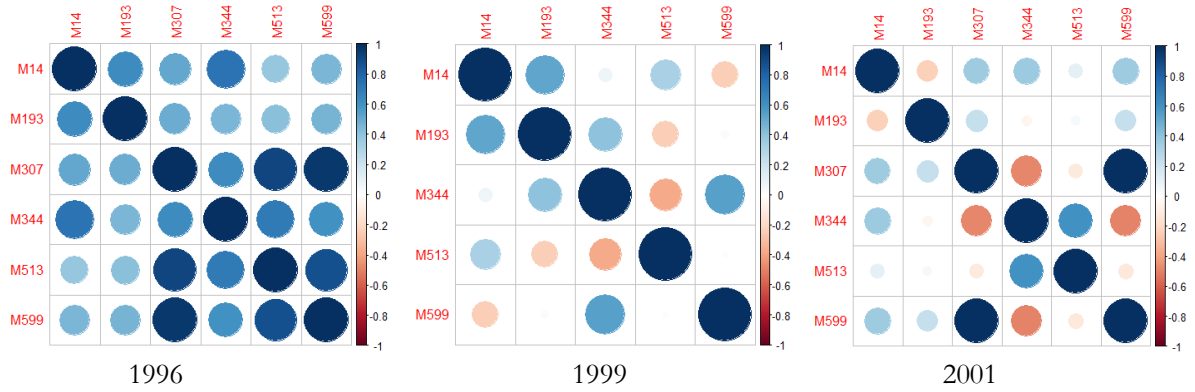


Figure 4.27: Synchrony of cities connected by intercity network line 1

Figure (4.27) shows the synchrony of the disease frequency between connected municipalities by line 1. This line connects cities from:
Rotterdam(599) =>Gouda(513) =>Utrecht(344) =>Amersfoort(307) =>Zwolle(193) =>Groningen(14). When examining calculated values of the correlation coefficients in this urban zone for the three years, year 1996 shows a higher synchrony of diffusion between connected cities with a higher average pairwise correlation coefficients (average = 0.61717). But this higher synchrony is not shown in the year 1999 (average = 0.101907) and 2001 (average = 0.1200057). Especially in year 1999, the correlation was very low, in addition, there were no disease cases in Amersfoort (307) and this city was removed from the analysis during this year. In general, in this intercity connectivity line only year 1996 shows higher synchrony of diffusion between connected cities.

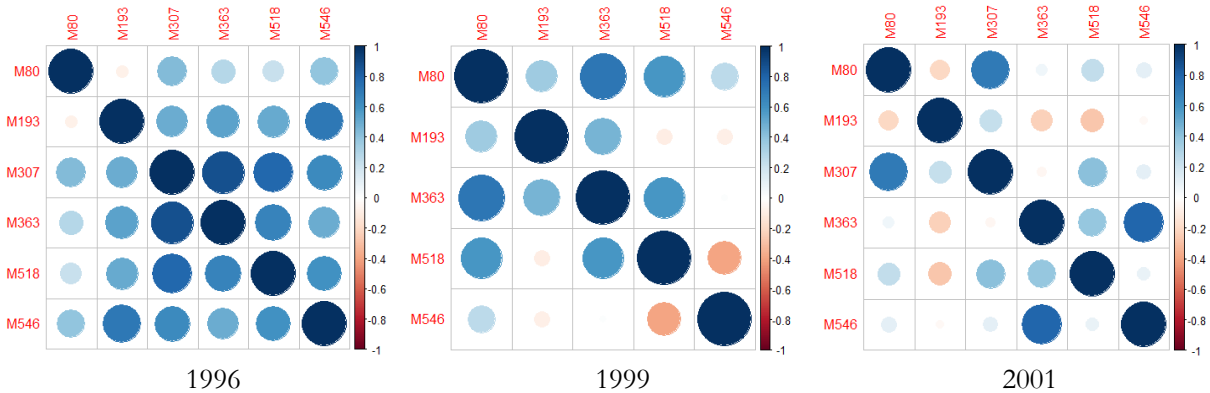


Figure 4.28: Synchrony of cities connected by intercity network line 2

Line 2 contains the intercity line from:
Den Haag (518) =>Leiden (546) =>Amsterdam (363) => Amersfoort (307) => Zwolle (193) =>Leeuwarden (80). The result presented in Figure (4.28) indicates the correlation between these connected cities. Similar to cities connected by intercity line 1, the results of this intercity connection show higher average pairwise correlations of synchrony in year 1996 (Average = 0.505062). The second higher value for this line was observed in the year 1999 (Average = 0.243072). During year 200l, lowest average correlation (0.158234) was observed. These cities connected by line 2 show clear synchrony of diffusion in the year 1996.

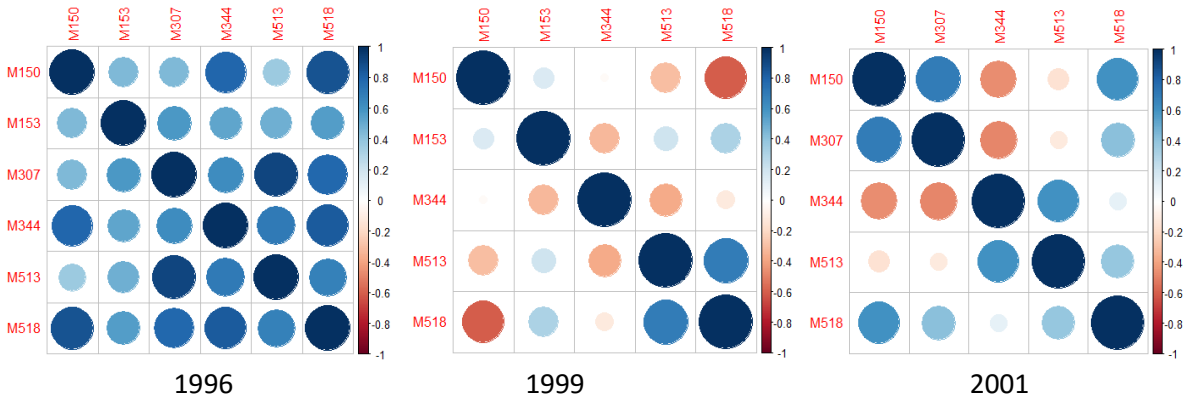


Figure 4.29: Synchrony of cities connected by intercity network line 3

The result displayed in the Figure (4.29) shows the synchrony of diffusion between cities connected by an intercity line from:
Den Hague (518) =>Gouda(513) =>Utrecht(344) =>Amersfoort(307) =>Deventer(150) =>Enschede(153). The results show synchrony was higher during year 1996 (average= 0.640114) as well. In year 1999 the average pairwise correlation was very low (average=0.03841) and there were no disease cases in Amersfoort (307). The year 2001 also shows a low value of average pairwise correlation (average=0.158656). There was also no disease case in Enschede (153) during this year. For this intercity connectivity line higher synchrony was observed during the year 1996 like other cities connected by intercity lines.

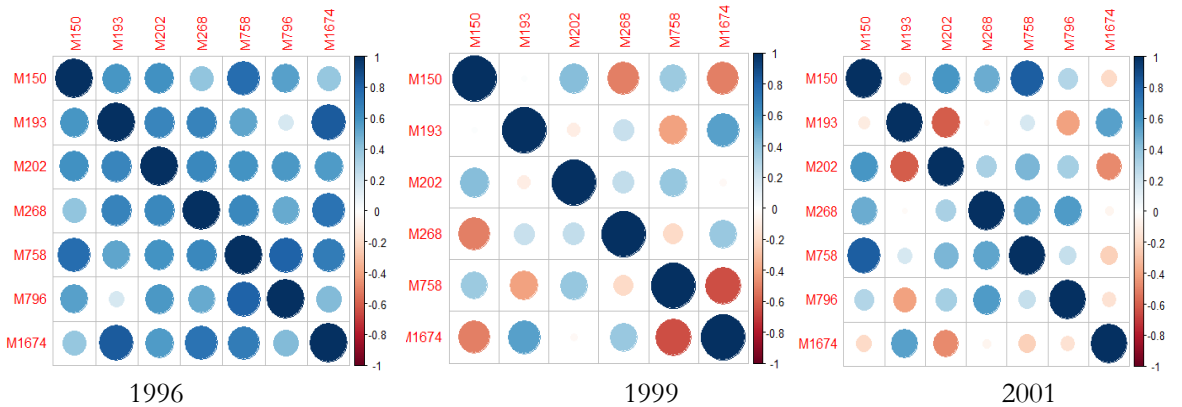


Figure 4.30: Synchrony of cities connected by intercity network line 4

Figure (4.30) present the results found by intercity connection line which contains cities from:
Roosendaal(1674)=>Breda(758)=>Den Bosch(796)=>Nijmegen(268)=>Arnhem(202) =>Deventer(150) =>Zwolle(193). The results of this line also show highest average pairwise correlation (0.588147) in the year 1996. In year 1999, very low average pairwise correlation (Average= 0.013699) was found and no disease cases were found in Den Bosch (796). For this line year 2001 also shows low average pairwise

correlation (0.148078513). The year 1996 shows similarity with the other intercity connection lines which shows the highest synchrony of diffusion between connected cities in the year 1996.

Overall, the highest average pairwise synchrony was observed during the year 1996 for all large cities connected by intercity line. The result found during the year 1996 confirmed that there is a higher synchrony of diffusion between more connected cities than in less connected cities. But this result is not found in the year 1999 and 2001. During these two years no clear synchrony of diffusion was observed. This may indicate that, we need to include other factors to analyze the diffusion patterns of infectious disease.

4.4.3. Disease diffusion in non-urban areas

The results of the urban to non-urban diffusion was analysed using the timing of the peak month during the epidemic year. The rankings of means of timing of the peak for both urban and non-urban areas were calculated yearly and plotted together to visualize the difference in diffusion patterns between urban and non-urban zones. From the plots we can see that the means of ranks of months containing a peak disease case of urban areas are a bit lower than the non-urban areas for all the three epidemic years. Lower ranks of means of months with a peak value indicate that disease reaches the peak earlier in the urban areas.

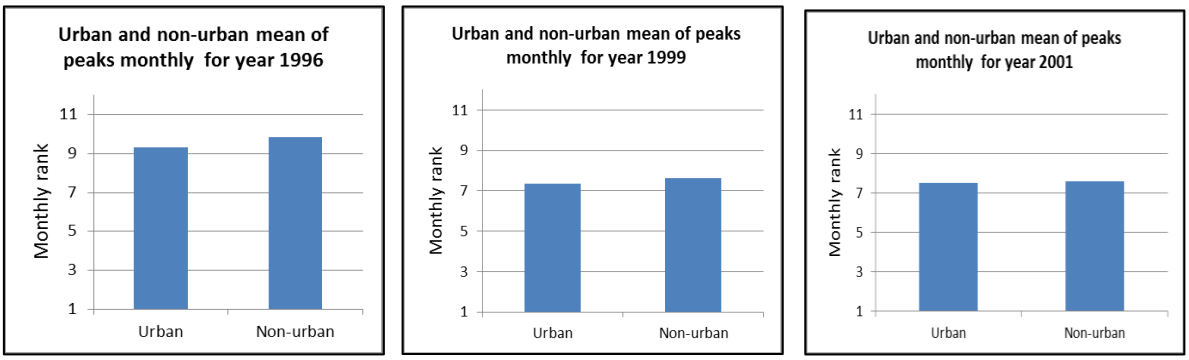


Figure 4.31: Mean rank of months with peak disease cases in urban and non-urban areas

The number of municipalities used for non-urban areas differ based on the number of peaks (maximum disease cases) identified in the epidemic year. Numbers of non-urban municipalities which have a peak disease cases were identified by removing municipalities with zero disease cases and peak value of 1. Using these criteria, for the year 1996, 63 non-urban municipalities were identified and based on these numbers of non-urban municipalities; the first large 63 urban zone municipalities which have a peak value during the epidemic year were selected. Ranks of months of in the urban and non-urban during the epidemic year were assigned and means of ranks of months with a peak value were 9.33 and 9.87 respectively. For the year 1999, 64 non-urban and 64 urban zone municipalities were selected by using similar methods used for the year 1996. The means of the ranks of months with a peak value were 7.34 for urban zones and 7.64 for non-urban zones. Means of rank of months with a peak disease case were lower in the urban areas for the year 1999 also. For the year 2001, 79 non-urban and urban municipalities were identified. The means of the ranks of months with a peak value were 7.51 for urban zones and 7.58 for non-urban zones. These mean values also showed a little bit differently between urban and non-urban areas. The mean values for all the three years indicate that urban zone disease case reaches the peak first and non-urban zones next to urban zones. This result may suggest that the disease cases arise first in the urban areas and diffuse to the non-urban areas. Table (8) given below summarizes the descriptive statistics of ranks of months, which have a peak value for all the three years.

Table 8: Descriptive statistics of mean ranks of months for urban and non-urban areas

Name	N	Mean	Std.Deviation	Minimum	Maximum
Non-urban_1996	63	9.87	1.68	3	12
Urban_1996	63	9.33	1.626	4	12
Non-urban_1999	64	7.64	2.698	1	12
Urban_1999	64	7.34	2.75	1	12
Non-urban_2001	79	7.58	3.169	1	12
Urban_2001	79	7.51	3.084	1	12

5. DISCUSSION

5.1.1. Initial data exploration

The results of initial disease data exploration indicate most of the municipalities have fewer than 50 disease cases for all epidemics in all three epidemic years. Very few municipalities have a higher number of disease cases. As can be observed from the plots in section (4.1) the highest peak in disease cases (962) happened in the year 1996 during October and all the urban zones have similar peaking time. This indicates that the disease was moving faster during this year and the epidemic outbreak was shorter for the years 1996. The timing of the peaks for year 1999 and 2001 vary between urban zones.

5.1.2. Comparison of different mobility data

Different mobility data were used to group areas in to different urban zones. The reason why we need to group areas to urban zone is because of the dense social-contact networks of urban areas characterize them to form a perfect match for fast, uncontrolled disease diffusion (Eubank et al., 2004). In addition, the movement of most individuals is limited within particular urban areas (Kang et al., 2012). For this reason, the intention was given to identify urban areas and three mobility datasets (distance between road layer intersection points, urban agglomeration networks and commuting distance) were used to split the country into different urban zones.

The amount of overlap of the urban zones generated by each of these methods was large. Among these mobility data distance between road network intersection points using the percolation method was the chosen and suitable for the analysis of disease diffusion. The percolation method used to group areas based on density of road network. Grouping using the percolation method is also supported by the research findings of other scholars who used percolation method based on distance between road layer intersection points to split urban areas of Britain (Arcaute et al., 2015).

5.1.3. Analysis of disease diffusion

5.1.4. Measuring hierarchical diffusion

Hierarchical diffusion was measured by analysing the relationship between population size and disease fadeouts and by fitting a non-linear regression model using two different datasets: duration of fadeouts and the number of fadeouts at two spatial aggregation level (within urban zone and between urban zones).

Within the urban zones, the results of measuring hierarchy of diffusion using the total number of fadeouts in relation to population size didn't lead to clear patterns of diffusion except for Randstad Holland. Possible explanation for this is that the larger cities are relatively small and therefore still experience a considerable number of fade-outs. In Randstad Holland the cities are larger (e.g. Amsterdam) and no more fadeouts were found.

The results of measuring hierarchy using mean duration of fadeouts within urban zones indicate that, pertussis diffuses hierarchically from cities with large population sizes to cities with small population size. This was observed in all urban zones. The identified hierarchical diffusion shows a similarity to the hierarchies of diffusion found by Broutin et al. (2004). The results show that the disease persists longer in larger populations, indicating that the larger the population, the shorter duration of fadeouts. This also in agreement with observed by Rohani et al. (2000) who found the larger the locality, the shorter the period of the disease extinction using pertussis time series in 60 cities in England and Wales. In conclusion, this

study suggests that the spatial hierarchical patterns of pertussis were from large to small cities depending on contact, which occurs more on large cities.

Hierarchical diffusion between urban zones shows for both numbers of fadeouts and duration of fadeouts a hierarchical diffusion. As the total population of the urban zone is much larger compared to the population of a single municipality, it is not surprising that the number of fadeouts is very low (almost none). All the urban zones meet the CCS and the disease persists in all the urban zones.

The way the urban zones were determined is unlikely to affect the results. The percolation method used to determine group areas based on distance between road intersection points was successful in group areas based on the density of road networks. Since, when we visualize the road layer, the more the density of the road network, the larger the connected urban zone identified. In addition, the areas grouped based on the connectivity of the road network are overlapped with the urban agglomerations identified by the national government in the Netherlands (van VROM, 2004) and areas identified by the commuting distance.

The method used to measure hierarchical diffusion does not provide a quantitative measure (degree of hierarchy) but is totally driven by visual interpretation of the results. And therefore, it could be good to quantify the degree of hierarchy using other methods such as using a Bayesian hierarchical diffusion model which allows accounting for individual difference and commonalities.

Overall, results hierarchical diffusion show a clear diffusion pattern when we use mean duration of fadeouts and also a there is difference in in the two aggregation levels (within urban zone and between urban zones). No significant difference was observed between the three epidemic years during the measuring of hierarchical diffusion.

5.1.5. Measuring synchrony of diffusion

During this research synchrony was measured using the frequency (number of disease cases per 100000 inhabitants) of the disease for each municipality and by assessing the Pearson correlations between them. Similarly to hierarchy, synchrony was measured at two aggregation levels (within urban zones and between urban zones for three epidemic years.

Different type of mobility data have been applied to measure synchrony at different spatial aggregation levels. The results shown that different aggregation levels lead to differences in the result of synchrony. The results of synchrony at the municipality (within urban zone) level were very low. Only Randstad Holland during the year 1996 shows a medium synchrony (average pairwise synchrony=0.503855). This urban zone contains more number of large cities. There was also higher synchrony at the regional level between identified urban zones than within urban zones for all years. This is in line with the concept of synchrony exists as part of a hierarchical diffusion pattern i.e. cities at the same level of the hierarchy. The results are in a way similar as expected epidemics in large centers are highly synchronized (Grenfell et al., 1994).

Additionally, the results of average pairwise synchrony between each urban zone and the largest urban zone Randstad Holland were calculated and compared based on their distance from the largest urban zone (Randstad Holland). The results show that the average pairwise values didn't depend on the separation distance. This indicates that a separation distance between pairs doesn't matter the synchrony of diffusion.

The other way applied to measure synchrony was using railway mobility data. The railway data was used to select municipalities connected by railway network and to measure synchrony between large cities

connected by the railway network. The results of this municipalities connected by railway mobility data also show lower synchrony between municipalities. Only the year 1996 shows higher average (0.513605) value of synchrony. There was a significant variation between year 1996 and the other two years 1999 (0.0853337) and 2001 (0.148318). But no significant variation was observed between the year 1999 and 2001. During the year 1996, there was higher value (0.513605) of the average pairwise synchrony, which indicates similarity in the diffusion pattern of the disease. The correlation between the synchrony of diffusion and connecting by the railway mobility network is not showing higher values in all year. This could link to other factors which cause variations in the disease causes and human activities.

Moreover, synchrony was measured by selecting municipalities connected by intercity network. The results show that, there was higher synchrony of diffusion in more connected large cities during the year 1996. This diffusion during year 1996 is an indication of, higher synchrony of diffusion of the disease in more connected cities than in less connected cities. This result is in agreement with Grenfell et al. (1994) who observe that epidemics in large centres were highly synchronized. This is also the characteristic of hierarchical spread (synchrony between large cities). For the years 1999 and 2001 no synchrony of diffusion was observed. We shall suggest that, further analysis is needed because they were not showing consistent results.

5.1.6. Disease diffusion in non-urban areas

Results of diffusion analysis for urban versus non-urban areas indicate that disease case of pertussis reached the peak in urban areas sooner than in the non-urban areas. However, the difference between the mean timing of the peak was not significant. It is around one month for all the epidemic years. During this analysis, it is difficult to determine the exact diffusion of the disease from urban to non-urban areas based the timing of the peak.

In conclusion, disease diffusion patterns were observed in highly connected cities, for instance, when we look at the measuring hierarchy of diffusion using the total number of fadeouts in relation to population size, hierarchy of diffusion was clearly shown in the Randstad_Holland urban zone. This urban zone has a higher number of large cities with a large number of commuting trips taking place between these cities. In addition, when we look at synchrony during the year 1996, at the municipality level, only Randstad_Holland showed higher synchrony of diffusion this is also an indication for synchrony is higher between highly connected large cities. Moreover, synchrony is getting higher when we analysed the 30 largest cities connected by railway. The measure of synchrony again increased when we analysed more connected cities connected by intercity network. All the analysis performed for the year 1996 shows there is a correlation between disease diffusion and the connectivity of large cities. The reason why the year 1996 shows higher synchrony of diffusion between highly connected cities is, the highest peak disease case was happening in this year and most of the municipalities have a similar peaking time. The other reason is that the disease diffused very fast during this year. For the years 1999 and 2001 further analysis is needed because they were not showing consistent results. Results during these two years were variable when the spatial aggregation varies. On the contrary, some results show low (negative average pairwise) values of synchrony when connectivity increases. Especially, the year 1999 shows very low average pairwise correlation for most of the results found during the analysis. The reason for this result is that there are large cities without disease cases or a large city with very low number of disease cases. For instance, Rotterdam has very few numbers of disease cases (only 3 disease cases) during year 1999 and no disease cases in Eindhoven during year 2001.

6. CONCLUSION AND RECOMMENDATIONS

Different mobility data and analysis approaches have been presented in this research to meet the objective of the study and answer the research questions. This section presents how those research questions were answered to address the objectives of the study and recommendations for further analysis and future work.

6.1. Conclusions

The main objective of this research was to compare different mobility variables at different spatial-temporal scales to determine their relevance as indicators for measuring the diffusion pattern of infectious diseases at different scales. The aforementioned objective was addressed by answering the following research questions:

1. What are the mobility indicator variables that can be used as proxies for identifying urban zones/lines with high commuting?

To answer this question, a literature review was performed and we found mobility has several indicator variables including distance, road networks, urban networks, mobile phone data, highway networks, railway networks and traffic indexes.

Different experiments were conducted to identify urban zones including the percolation method using a road network, urban agglomerations, commuting distance and railway network connection of large cities. During this research, both the road network and the railway network were selected to create subsets of municipalities. Areas with a high density of roads were selected as urban zones (assuming a high commuting between municipalities inside these zones), and both the within and between urban zone disease diffusion was analyzed. A selection of line based commuting was made based on both the highway and railway network, in order to select highly connected cities (highly commuting lines). From these, subsets of municipalities were selected using intercity train tracks, to analyze disease diffusion between cities with direct links.

2. Which mobility indicator variable is more appropriate to identify urban zones/lines in the context of disease diffusion at a certain spatial-temporal scale?

Differences were found between analysis within and between the urban zones, and between the urban zone analysis and large cities and railway tracks. Where hierarchy was mainly found within the urban zones, synchrony was found between urban zones. Again synchrony was found when we use intercity lines. This indicates that a multi-scale analysis using different mobility data is needed to get a good understanding of the complex diffusion pattern.

3. What type of disease variable can be used to study the disease in different zones/commuting lines?

In the literature, several disease variables such as, disease cases, frequency, first time of infection, direction of spread, number (duration) of fadeouts, timing of the peak disease cases, incidence etc. were used as a disease variable to analyse disease diffusion. In this study, the total number and mean duration of fadeouts were calculated and used as an indicator variable to measure hierarchical disease diffusion at the municipal spatial (within urban zone) scale. These disease variables were used in combination with population size. A

characteristic of hierarchical diffusion is that the disease maintains itself in larger population centres (cities) and spreads to smaller communities close by. In these communities the population is too small to maintain the disease. This should lead to a small number of fadeouts in both small and large settlements and a higher number of fadeouts in medium size communities. Hierarchical spread is also characterized by a long duration of fadeouts in small communities and a short duration of fadeouts in large communities. Within urban zones, the analyses revealed a pattern typical for hierarchical spread, although in many cases the number of fadeouts of larger cities was still high. Even though, both mean duration and the total number of fadeouts show hierarchical patterns, results using the mean duration of fadeouts seem to be more suitable. We found a clear hierarchy of diffusion within all urban zones when we use the mean duration of fadeouts. As a result, this study suggests, mean duration of fadeout is the better disease indicator variable to analyse hierarchical diffusion. The other disease variable used in this research was frequency (number of disease cases per 100000 inhabitants) to measure synchrony of diffusion. This disease variable was also suitable to make comparison between population groups. Other disease indicators such as time of first infection, speed and direction of spread could have been used, yet were not included in this study but might be useful for analyzes of disease diffusion.

4. Is there a hierarchy of disease diffusion and how does this hierarchy vary between more and less connected areas?

In this research, the relationship between disease number of fadeouts and mean duration of fadeouts were analysed to see the hierarchical diffusion patterns of the disease on the selected urban zones. The results of this analysis show that there is a hierarchy of diffusion in more connected cities. For instance, when we observe the result of total number of fadeouts in relation to population size, hierarchy of diffusion was not found within less connected urban zones. Hierarchy of diffusion was found only in Randstad Holland, which contains a higher number of connected large cities. In this urban zone both the number of fadeouts and duration of fadeouts showed hierarchical diffusion. In addition, this urban zone is the largest urban zone, which is identified by the density of connected road network intersection points. This indicates that there is a hierarchy of diffusion in more connected areas.

5. Is there a synchrony of disease diffusion and how does this synchrony vary between more connected and less connected areas?

Synchrony refers to the fact that different cities, located further away from each other, are infected at the same time. This can be caused by commuting. It is often observed between cities that are farther apart, but have the same population size. In this research synchrony was checked between cities within an urban zone, but also between urban zones, and between cities connected by railways. The results of synchrony at the municipality (within the urban zone) level were very low. Only Randstad Holland, during the year 1996, shows a medium synchrony (average pairwise =0.503855). This urban zone contains a higher number of large cities. There was also higher synchrony between the identified urban zones than within urban zones for all years. This concept of synchrony exists as part of a hierarchical diffusion pattern (synchrony between larger cities at the same level in the hierarchy). During the year 1996, the results of municipalities connected by railway mobility data also show higher synchrony (average=0.513605) between municipalities. Again, during the year 1996, synchrony increased when we analysed municipalities connected by the intercity network. This diffusion during the year 1996 is an indication of, higher synchrony of diffusion of the disease in more connected cities than in less connected cities. But for the years 1999 and 2001 no synchrony of diffusion was observed. This could be linked to other factors (besides mobility) which cause variations in the disease diffusion.

6. Are the epidemic years showing similar or different diffusion pattern on the identified zones/lines?

For a hierarchical diffusion pattern analysis the three epidemic years (1996, 1999 and 2001) were not showing significant differences in diffusion hierarchies. On the other hand, for the analysis of synchrony of disease diffusion, there was a significant difference between the epidemic years. Higher synchrony of diffusion was observed in the year 1996. The reason why the year 1996 shows higher synchrony of diffusion between highly connected cities is, the highest peak disease case was happening in this year and most of the municipalities have a similar peaking time. The other reason is that the disease diffused very fast during this year. But for the years 1999 and 2001 no synchrony of diffusion was observed. Especially, the year 1999 shows very low average pairwise correlation for most of the results found during the analysis. During these years there are large cities without disease cases or with very low numbers of disease cases. For instance, Rotterdam has very few disease cases (only 3 disease cases) during year 1999 and no disease cases were recorded in Eindhoven during the year 2001. Results during these two years were variable when the spatial aggregation varies therefore, further analysis is needed to analyse what makes the difference.

7. What is the relationship between the disease diffusion of urban versus non-urban areas?

Diffusion analysis of urban versus non-urban areas show that, disease case of pertussis reached the peak in urban areas sooner than in the non-urban areas. However, the difference between the timing of the peak between the two groups was not significant and it is difficult to determine the exact diffusion of the disease from urban to non-urban areas based on the timing of the peak. This indicates another disease indicator variable such as time of first infection is needed to get a better understanding of urban versus non-urban diffusion pattern.

6.2. Recommendation

In this research recommendations are made based on the result of the analysis and the overall findings of the research work. The results of this study could be further analyzed by including the following recommendations:

- For future study, we recommend extending the scale to the national level, to see the diffusion pattern at a countrywide pattern of diffusion.
- In this research, we have only tried to analyze the relationship between population size and disease fadeout in order to measure hierarchical diffusion. Since the disease is complex and influenced by different factors, it is important to include other variables and their effects on the spatial-temporal diffusion pattern will be needed to see a clear diffusion pattern of infectious disease. One way could be by investigating the diffusion patterns with other factors such as vaccination coverage, seasonality, number of immigrants, the growth rate of the population etc.
- During the analysis of hierarchical diffusion the pattern of the fitted model sometimes flipped; the fitted model could be improved to see a better pattern of hierarchy of diffusion.

- When we measure synchrony of diffusion we see a clear synchrony only in the year 1996. No synchrony was found in the years 1999 and 2001. This may be linked to other factors and needs to be further analysis.
- Since the number of commuters varies based on age groups, it is recommended to study the relationship between human mobility and disease diffusion based on age groups.

- In this research, different patterns of diffusion were found for different mobility proxies. This indicates that a multi-scale analysis using different mobility data is needed to get a good understanding of the complex diffusion pattern. Hence, we recommend using other proxies besides the ones we used.
- The mobility proxies used in this research, such as distance have no impact on the diffusion pattern of disease and are not showing a consistent pattern of diffusion for all years. Therefore, analysis of disease diffusion by getting real commuting data instead of using proxies is also recommended.
- Measuring hierarchies and synchrony to analyze diffusion patterns did not lead to clear results at both scales. Hence, we suggest to also using other methods to analyze disease diffusion patterns such as contagious diffusion and travelling waves.

LIST OF REFERENCES

Arcaute, E., Molinero, C., Hatna, E., Murcio, R., Vargas-Ruiz, C., Masucci, P., & Batty, M. (2015). Regions and Cities in Britain through hierarchical percolation. *Physics and Society*, *1*(30), 11. Physics and Society.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(51), 21484–9. doi:10.1073/pnas.0906910106

Balcan, D., Gonçalves, B., Hu, H., Ramasco, J. J., Colizza, V., & Vespignani, A. (2010). Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science*, *1*(3), 132–145. doi:10.1016/j.jocs.2010.07.002

Balcan, D., & Vespignani, A. (2011). Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nature Physics*, *7*(7), 581–586. doi:10.1038/nphys1944

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747–3752. doi:10.1073/pnas.0400087101

Bartlett, M. S. (1960). The critical community size for measles in the U.S. *J R Stat Soc A*, *123*(for 1940), 37–44. doi:10.2307/2343186

Bell, I. R., Schwartz, G. E., Boyer, N. N., Koithan, M., & Brooks, A. J. (2013). Advances in Integrative Nanomedicine for Improving Infectious Disease Treatment in Public Health. *European Journal of Integrative Medicine*, *5*(2), 126–140. doi:10.1016/j.eujim.2012.11.002

Bharti, N., Djibo, A., & Ferrari, M. (2010). Measles hotspots and epidemiological connectivity. *Epidemiology and Infection*, *138*(09), 1308–1316. doi:10.1017/S0950268809991385

Bjørnstad, O., Ims, R., & Lambin, X. (1999). Spatial population dynamics: analyzing patterns and processes of population synchrony. *TREE*, *14*(11), 427–432.

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*(7075), 462–5. doi:10.1038/nature04292

Broutin, H., Elguero, E., Simondon, F., & Guégan, J.-F. (2004). Spatial dynamics of pertussis in a small region of Senegal. *Proceedings. Biological Sciences / The Royal Society*, *271*(1553), 2091–8. doi:10.1098/rspb.2004.2847

Broutin, H., Mantilla-Beniers, N., & Rohani, P. (2007). Ecology Of Infectious Diseases: An Example with Two Vaccine-Preventable Infectious Diseases. *Encyclopedia of Infectious Diseases: Modern Methodologies*, 189–198.

Charaudeau, S., Pakdaman, K., & Boëlle, P. (2014). Commuter mobility and the spread of infectious diseases: application to influenza in France. *PloS One*, *9*(1). doi:10.1371/journal.pone.0083002

Choisy, M., & Rohani, P. (2012). Changing spatial epidemiology of pertussis in continental USA. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4574–4581. doi:10.1098/rspb.2012.1761

Chowell, G., Hyman, J., Eubank, S., & Castillo-Chavez, C. (2003). Scaling laws for the movement of people between locations in a large city. *Physical Review E*, *68*(6).

Cliff, A. D., Haggett, P., Ord, J. K., & Versey, G. R. (1981). *Spatial Diffusion: An Historical Geography of Epidemics in an Island Community*. New York: cambridge university press.

Cliff, A., & Haggett, P. (2004). Time, travel and infection. *British Medical Bulletin*, *69*(1), 87–99. doi:10.1093/bmb/ldh011

Cullen, R. (2003). Seasonality and critical community size for infectious diseases. *The ANZIAM Journal*, *44*, 501–512.

De Melker, H. E., Conyn-van Spaendonck, M. A., Rümke, H. C., van Wijngaarden, J. K., Mooi, F. R., & Schellekens, J. F. (1997). Pertussis in The Netherlands: an outbreak despite high levels of immunization with whole-cell vaccine. *Emerging Infectious Diseases*, *3*(2), 175–8. doi:10.3201/eid0302.970211

Dobesova, Z. (2011). Programming language Python for data processing. In *2011 International Conference on Electrical and Control Engineering* (pp. 4866–4869). IEEE. doi:10.1109/ICECENG.2011.6057428

Essam, J. W. (1980). Percolation theory. *Reports on Progress in Physics*, *43*(7), 833–912. doi:10.1088/0034-4885/43/7/001

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, *429*(6988), 180–184. doi:10.1038/nature02541

Frias-Martinez, E., Williamson, G., & Frias-Martinez, V. (2011). An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* (pp. 57–64). IEEE. doi:10.1109/PASSAT/SocialCom.2011.142

Ge, E., Lai, P. C., Zhang, X., Yang, X., Li, X., Wang, H., & Wei, X. (2015). Regional transport and its association with tuberculosis in the Shandong province of China, 2009–2011. *Journal of Transport Geography*, *46*, 232–243. doi:10.1016/j.jtrangeo.2015.06.021

Gong, Y. . W., Song, Y. R., & Jiang, G. P. (2014). Epidemic spreading in metapopulation networks with heterogeneous infection rates. *Physica A: Statistical Mechanics and Its Applications*, *416*, 208–218. doi:10.1016/j.physa.2014.08.056

Gonzalez, M., Hidalgo, C., & Barabasi, A. (2008). Understanding individual human mobility patterns. *Nature*, *453*(5). doi:10.1038/nature06958

Grenfell, B. (1997). (Meta)population dynamics of infectious diseases. *Trends in Ecology & Evolution*, *12*(10), 395–399. doi:10.1016/S0169-5347(97)01174-9

Grenfell, B., Kleczkowski, A., Ellner, S. P., & Bolker, B. M. (1994). Measles as a Case Study in Nonlinear Forecasting and Chaos. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *348*(1688), 515–530. doi:10.1098/rsta.1994.0108

Grimaldi, C., & Balberg, I. (2006). Tunneling and Nonuniversality in Continuum Percolation Systems. *Physical Review Letters*, *96*(6), 066602. doi:10.1103/PhysRevLett.96.066602

Grunsky, E. (2002). R: a data analysis and statistical programming environment–an emerging tool for the geosciences. *Computers & Geosciences*, *28*, 1219–1222.

Hennemann, J., Kohl, C. D., Smarsly, B. M., Metelmann, H., Rohnke, M., Janek, J., … Wagner, T. (2015). Copper oxide based H2S dosimeters – Modeling of percolation and diffusion processes. *Sensors and Actuators B: Chemical*, *217*, 41–50. doi:10.1016/j.snb.2015.02.001

Jacobs, J. H., Viboud, C., Tchetgen, E. T., Schwartz, J., Steiner, C., Simonsen, L., & Lipsitch, M. (2014). The association of meningococcal disease with influenza in the United States, 1989-2009. *Pl0S One*, *9*(9). doi:10.1371/journal.pone.0107486

Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and Its Applications*, *391*(4), 1702–1717. doi:10.1016/j.physa.2011.11.005

Koomen, E., & Groen, J. (2004). Evaluating future urbanisation patterns in the Netherlands. In *paper for the 44th congress of the European Regional Science Association August 25–29, 2004*. Porto, Portugal.

Li, D., Zhang, Q., Zio, E., Havlin, S., & Kang, R. (2015). Network reliability analysis based on percolation theory. *Reliability Engineering & System Safety*, *142*, 556–562. doi:10.1016/j.ress.2015.05.021

Lloyd, A. (2001). Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theoretical Population Biology*, *60*(1), 59–71. doi:10.1006/tpbi.2001.1525

Lloyd-Smith, J., & Cross, P. (2005). Should we expect population thresholds for wildlife disease? *Trends in Ecology & Evolution*, *20*(9), 511–519. doi:10.1016/j.tree.2005.07.004

Mei, S., Chen, B., Zhu, Y., Lees, M. H., Boukhanovsky, A. V., & Sloot, P. M. A. (2015). Simulating city-level airborne infectious diseases. *Computers, Environment and Urban Systems*, *51*, 97–105. doi:10.1016/j.compenvurbsys.2014.12.002

Merler, S., & Ajelli, M. (2010). Human mobility and population heterogeneity in the spread of an epidemic. *Procedia Computer Science*, *1*(1), 2237–2244. doi:10.1016/j.procs.2010.04.250

Merler, S., & Ajelli, M. (2010). The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings. Biological Sciences / The Royal Society*, *277*(1681), 557–565. doi:10.1098/rspb.2009.1605

Meyers, R. A. (2012). *Mathematics of Complexity and Dynamical Systems*. Springer Science & Business Media. Retrieved from https://books.google.com/books?id=iXEmLUcXAPcC&pgis=1

Ministerie van VROM. (2004). Summary National Spatial Strategy: creating space for development. Retrieved from https://www.rijksoverheid.nl/

Molinero, C., Arcaute, E., Smith, D., & Batty, M. (2015). The Fractured Nature of British Politics, 13. Physics and Society. Retrieved from http://arxiv.org/abs/1505.00217

Ortúzar, J. de D., & Willumsen, L. G. (2011). *Modelling Transport* (Fourth ed.). John Wiley & Sons.

Plata, S. (2006). A note on Fisher's correlation coefficient. *Applied Mathematics Letters*, *19*(6), 499–502. doi:10.1016/j.aml.2005.02.036

Poletto, C., Tizzoni, M., & Colizza, V. (2012). Heterogeneous length of stay of hosts' movements and spatial epidemic spread. *Scientific Reports*, *2*, 476. doi:10.1038/srep00476

Poletto, C., Tizzoni, M., & Colizza, V. (2013). Human mobility and time spent at destination: impact on spatial epidemic spreading. *Journal of Theoretical Biology*, *338*, 41–58. doi:10.1016/j.jtbi.2013.08.032

Remais, J., Akullian, A., Ding, L., & Seto, E. (2010). Analytical methods for quantifying environmental connectivity for the control and surveillance of infectious disease spread. *Journal of the Royal Society, Interface / the Royal Society*, *7*(49), 1181–1193. doi:10.1098/rsif.2009.0523

Riley, S., Eames, K., Isham, V., Mollison, D., & Trapman, P. (2014). Five challenges for spatial epidemic models. *Epidemics*, *10*, 68–71. doi:10.1016/j.epidem.2014.07.001

Rohani, P. (1999). Opposite Patterns of Synchrony in Sympatric Disease Metapopulations. *Science*, *286*(5441), 968–971. doi:10.1126/science.286.5441.968

Rohani, P., Earn, D. J. D., & Grenfell, B. T. (2000). Impact of immunisation on pertussis transmission in England and Wales. *Lancet*, *355*(9200), 285–286. doi:10.1016/S0140-6736(99)04482-7

Saberi, A. A. (2015). Recent advances in percolation theory and its applications. *Physics Reports*, *578*, 1–32. doi:10.1016/j.physrep.2015.03.003

Sattenspiel, L. (2009). *The Geographic Spread of Infectious Diseases: Models and Applications*. Princeton University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=jtGP_qwD1MgC&pgis=1

Schanzer, D., Langley, J., Dummer, T., & Aziz, S. (2011). The geographic synchrony of seasonal influenza: a waves across Canada and the United States. *PLoS ONE*, *6*(6). doi:10.1371/journal.pone.0021471

Statistics Netherlands. (2015). *Transport and Mobility 2015*. Statistics Netherlands. Retrieved from http://download.cbs.nl/pdf/2015-transport-and-mobility.pdf

Susilo, Y. O., & Maat, K. (2007). The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation*, *34*(5), 589–609. doi:10.1007/s11116-007-9129-5

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., … Colizza, V. (2014). On the Use of Human Mobility Proxies for Modeling Epidemics. *PLoS Computational Biology*, *10*(7). doi:10.1371/journal.pcbi.1003716

Van der Laan, L. (1998). Changing Urban Systems: An Empirical Analysis at Two Spatial Levels. *Regional Studies*, *32*(3), 235–247. doi:10.1080/00343409850119733

Vazquez-Prokopec, G. M., Bisanzio, D., Stoddard, S. T., Paz-Soldan, V., Morrison, A. C., Elder, J. P., … Kitron, U. (2013). Using GPS Technology to Quantify Human Mobility, Dynamic Contacts and Infectious Disease Dynamics in a Resource-Poor Urban Environment. *PLoS ONE*, *8*(4), 1–10. doi:10.1371/journal.pone.0058802

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science (New York, N.Y.)*, *312*(5772), 447–51. doi:10.1126/science.1125237

Wallace, R., & Wallace, D. (1999). Deindustrialization, inner-city decay, and the hierarchical diffusion of AIDS in the USA: how neoliberal and cold war policies magnified the ecological niche for. *Environment and Planning A*, *31*(1), 113–139. doi:10.1068/a310113

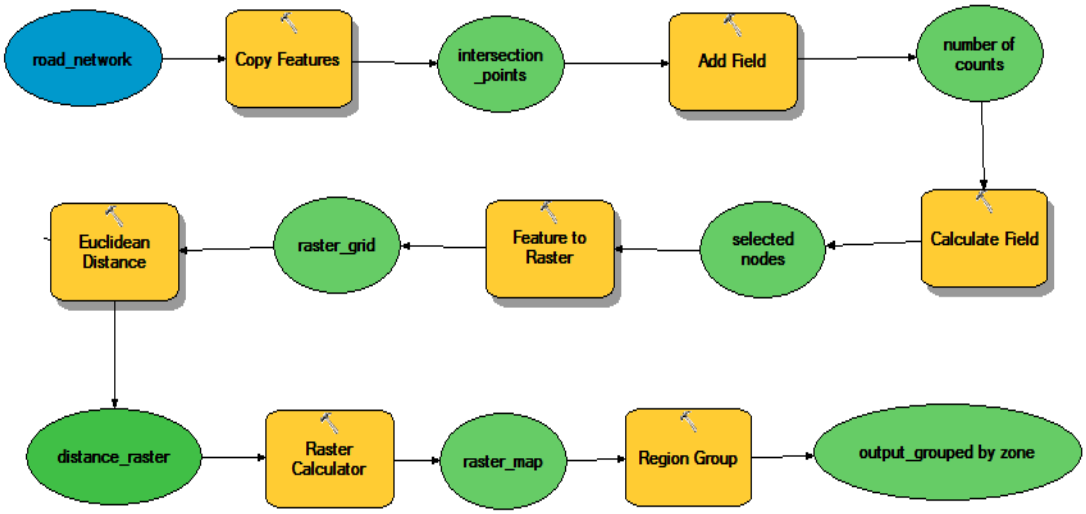
Wallinga, J., & Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, *160*(6), 509–16. doi:10.1093/aje/kwh255

Wu, U.I., Wang, J.T., Chang, S.C., Chuang, Y.C., Lin, W.R., Lu, M.C., … Chen, Y.C. (2014). Impacts of a mass vaccination campaign against pandemic H1N1 2009 influenza in Taiwan: a time-series regression analysis. *International Journal of Infectious Diseases : IJID : Official Publication of the International Society for Infectious Diseases*, *23*, 82–9. doi:10.1016/j.ijid.2014.02.016

APPENDICES

Appendix (A)

1. The percolation model used for identification of urban zones



2. **Maps of the percolation method at some distance thresholds using the road intersection points**

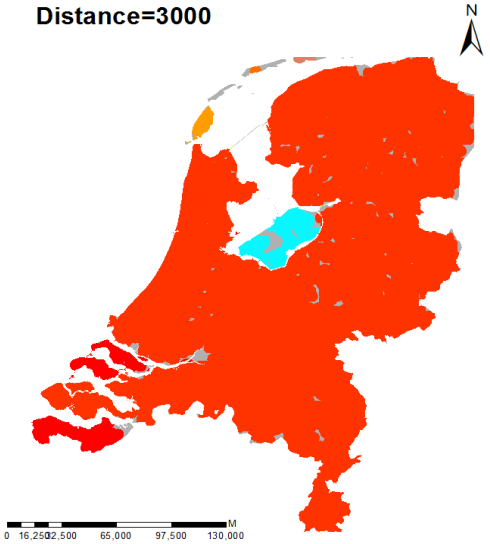
Distance=5000

Distance=4500

Distance=4000

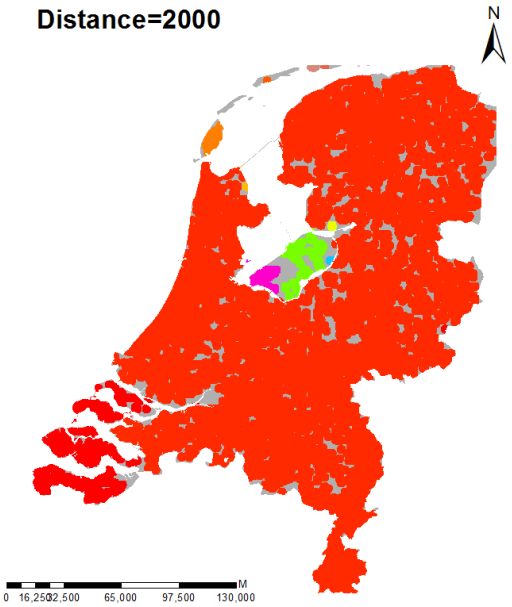
Distance=3500

Distance=3000

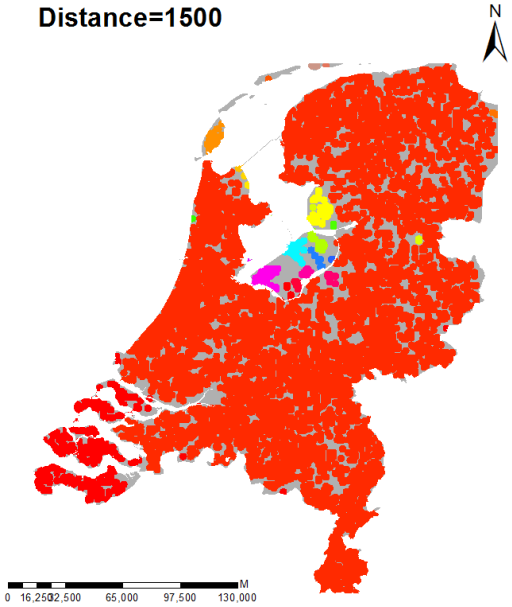


Distance=2500

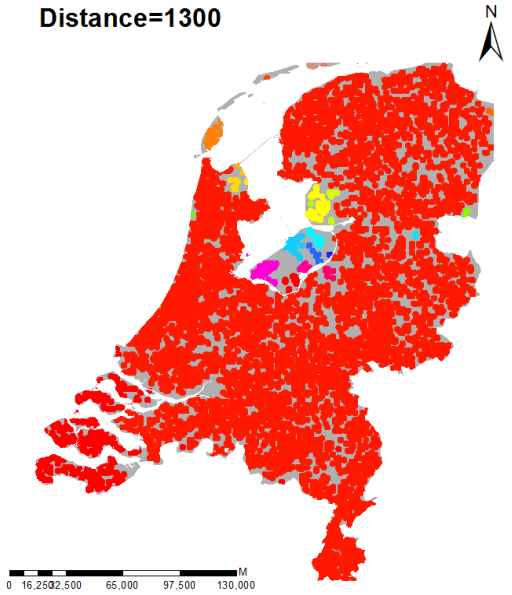
Distance=2000



Distance=1500



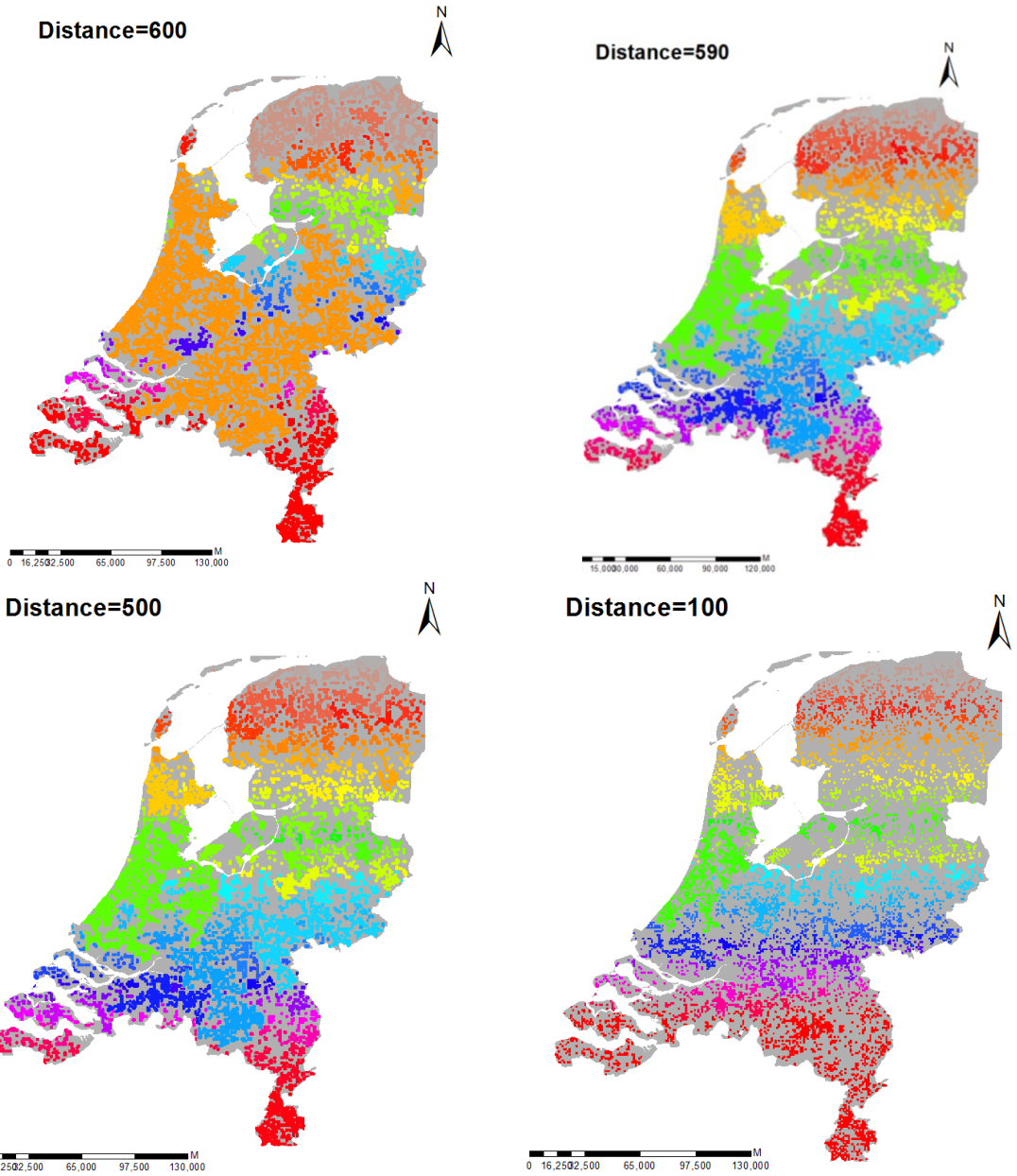
Distance=1300



Distance=1000

Distance=800

Distance=700



Appendix (B)

Python codes

Data preparation code (Python codes):- The disease number of fadeouts and duration of fadeouts and peak of timing of the disease cases were calculated using a python code.

1. Python code to calculate duration of disease fadeouts

```
import csv
import numpy as np
#Load .CSV file
f = open("96_19.csv")
r = csv.reader(f)
```

```
np_arr = np.empty((96, 12))
i = 0
for row in r:
  if i == 0:
    i += 1
    continue
  np_arr[i-1, :] = np.array([int(x) for x in row[1:13]])
  i += 1
f.close()
def count_next_zeros(row):
  count = 0
  for num in row:
    if num == 0:
      count += 1
    else:
      return count
  return count
agg_counts = []
for row in np_arr:
  idx = 0
  counts = []
  for num in row:
    if num != 0:
      count = count_next_zeros(row[idx+1:])
      if count != 0:
        counts.append(count)
      idx += 1
      if idx == row.size-1:
        agg_counts.append(counts)
    else:
      idx += 1
      if idx == row.size-1:
        agg_counts.append(counts)
      continue
print(agg_counts)
```

1. **Python code to calculate peak timing of the disease**

```
import csv
import numpy as np
f = open("Year_96.csv")
r = csv.reader(f)

np_arr = np.empty((11, 12))
i = 0
for row in r:
  if i == 0:
```

```
    i += 1
    continue
  np_arr[i-1, :] = np.array([int(x) for x in row[1:13]])
  i += 1

f.close()
months = ["Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"]
max_cases = []
for row in np_arr:
  max_case = int(np.max(row))
  max_cases.append((months[np.where(row==max_case)[0][0]], max_case))
print(max_cases)
```

Appendix (C)

R scripts used

1. Scripts used for plotting boxplot of the three years together

```
library(gstat)
library(maptools)
library(rgdal)
#setwd("d:\\12")
#load data
b= read.csv("all_years.csv")
boxplot(b$X1996,b$X1999,b$X2001,horizontal=TRUE,
       names=c("1996","1999","2001"),
       col=c("turquoise","tomato", "green"),
       #tag=("median"),
       xlab="disease cases per municipality", ylab="Year",
       main="Number of disease cases per municipality")
```

2. Scripts used for plotting disease case in relation to population size

```
library(gstat)
library(maptools)
library(rgdal)
#setwd("d:\\12")
#load data
pop_dis= read.csv("p&d_all.csv")
xdata = c(pop_dis$x)
ydata = c(pop_dis$y)
plot(xdata,ydata, main="Disease cases versus population size  for 2001",
 cex=1.5, cex.lab=1.25, pch=20, col= "blue",
 xlab="Disease cases per municipality",
  ylab="Population size(x10^5)")
```

3. Scripts used for analysing hierarchical diffusion

```
library(gstat)
library(maptools)
library(rgdal)
#setwd("d:\\12")
#To analyze hierarchical diffusion using total numeber of fadeouts and fitting a non-linea regression line
#load data
Mean_BR2= read.csv("Randstad_Holland_1996.csv")
Population_size = c(Mean_BR2$population_size)
Total_fadeout = c(Mean_BR2$T_fadeout)
#Plot population versus total numeber of fadeouts
plot(Population_size,Total_fadeout, main="Randstad_Holland for 1996",
cex=1.5, cex.lab=1.25, pch=20, col= "blue",
xlab="Population size(x10^5)", ylab="Total number of fadeouts" )
# Calculate standard deviation of x
sd(Mean_BR2$population_size)
#Calculate standard deviation of y
sd(Mean_BR2$T_fadeout)
## Calculate the correlation between x and y
cor(Mean_BR2$population_size,Mean_BR2$T_fadeout)
#Calculate the initial parameter p2
p2=(cor(Mean_BR2$population_size,Mean_BR2$T_fadeout)*(sd(Mean_BR2$T_fadeout)/sd(Mean_BR2
$population_size)))
#calculate mean of x
mean(Mean_BR2$population_size)
## Calculate mean of y
mean(Mean_BR2$T_fadeout)
# calculate the inital parameter p1
p1=(mean(Mean_BR2$T_fadeout)-(p2*mean(Mean_BR2$population_size)))
#Fitting a non-linear regression model
fit = nls(Mean_fadeout ~ p1*log(p2*(Population_size)), start=list(p1=p1,p2=p2))
#Summarize
summary(fit)
new = data.frame(Population_size = seq(min(Population_size),max(Population_size),len=200))
 #Do fitting the line
lines(new$Population_size, predict(fit,newdata=new), col = "red")
#Residuals
resid(fit)
```

4. Scripts for analysis of hierarchical diffusion using mean duration of fadeouts and fitting a non-linear regression line

```
library(gstat)
library(maptools)
library(rgdal)
##setwd("d:\\12")
```

```
#load data
     Mean_BR2= read.csv("Mean_AN_1999_fit.csv")
     Population_size = c(Mean_BR2$Population_size)
    Mean_fadeout = c(Mean_BR2$Mean_fadeout)
#Plot population versus total numeber of fadeouts
    plot(Population_size,Mean_fadeout, main="Arnhem-Nijmegen for 1996",
     cex=1.5, cex.lab=1.25, pch=20, col= "blue",
     xlab="Population size(x10^5)",
     ylab="Mean duration of fadeouts" )

#calculate the inital parameter p2
p1=(mean(Mean_BR2$Mean_fadeout)-(p2*mean(Mean_BR2$Population_size)))
#Calculate standard deviation of x
sd(Mean_BR2$Population_size)
#Calculate standard deviation of y
sd(Mean_BR2$Mean_fadeout)
#Callculate the correlation between x and y

ccor(Mean_BR2$Population_size,Mean_BR2$Mean_fadeout)

#calculate the inital parameter p2
p2=(cor(Mean_BR2$Population_size,Mean_BR2$Mean_fadeout)*(sd(Mean_BR2$Mean_fadeout)/sd(Me
an_BR2$Population_size)))
#calculate mean of x
mean(Mean_BR2$Population_size)
#Calculate mean of y
mean(Mean_BR2$Mean_fadeout)
# Calculate the inital parameter p2
p1= mean(Mean_BR2$Mean_fadeout)-(p2*mean(Mean_BR2$Population_size)))

#Fitting a non-linear regression model

fit = nls(Mean_fadeout ~ p1*exp(p2*Population_size) , start=list(p1=p1,p2=p2))

#summarize
summary(fit)
new = data.frame(Population_size = seq(min(Population_size),max(Population_size),len=200))

# Do fitting the line
lines(new$Population_size,predict(fit,newdata=new), col = "red")
#Residuals
resid(fit)
```

5. Scripts for analysis of synchrony of diffusion

```
library(gstat)
library(maptools)
library(rgdal)
library(corrplot)
#setwd("d:\\12")
#load data
cross_8= read.csv("s_96_8.csv")
#Calculate the Pearson correlation coefficients
cros <- cor(cross_8, method = "Pearson", use = "pairwise")
#Plot the correlation coefficients using corrplot package
corrplot(cros, method="circle")
#Correlation matrix
cros
```