REVEALING PATTERNS: SPATIO-TEMPORAL PATTERN DETECTION AND REPRODUCTION

Ellen-Wien Augustijn-Beckers

Graduation committee:

Chairman/Secretary

Prof.dr.ir. A. Veldkamp	University of Twente

Supervisors

Prof.dr. M.J. Kraak	University of Twente
Prof.dr. R. Zurita Milla	University of Twente

Members

Prof.dr.ir A. Stein	University of Twente
Prof.dr.ir. M.F.A.M. van Maarseveen	University of Twente
Prof.dr.I. Benenson	Tel Aviv University
Dr. D.J. Karssenberg	Utrecht University

Referee

Dr. A.N. Swart	RIVM

ITC dissertation number 323
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4578-5
DOI 10.3990/1.9789036545785

Cover designed by
Printed by ITC Printing Department

REVEALING PATTERNS: SPATIO-TEMPORAL PATTERN DETECTION AND REPRODUCTION

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the graduation committee,
to be publicly defended
on 11 July 2018 at 12:45 hrs

by

Petronella Wilhelmina Maria Augustijn-Beckers

born on 21 March 1964

in Geldrop, The Netherlands

This thesis has been approved by
Prof.dr. M.J. Kraak, supervisor
Prof.dr. R. Zurita-Milla, supervisor

Preface

Patterns are everywhere. In everyday life we often see them but we do not immediately recognise them as having any scientific meaning. Patterns are generally observed from the angle of beauty. We might stop to admire a nicely woven spider web or a snowflake. We know that patterns of self-similarity are frequently observed in flowers and leaves of plants. Perhaps we also reflect on the ribbles in the sand of a beach or wonder why half-circular patterns emerge when many people try to get through the same entrance, but we rarely link these patterns to the processes that produce them.

Patterns are the footprints left behind by nature and human beings and their scientific value is significant as different patterns are an indication that the processes responsible for these patterns are different. When developing simulating models, patterns are often used for validation. When the model is able to re-produce the patterns this is an indication that all essential components of the system have been captured.

The pattern that we are trying to reproduce is not always an end point, but as simulations are dynamic in space and time, so are the patterns. A pattern may consist of a number of states linked together in a fixed sequence like succession of ecosystems. The steps we see in the temporal dimension are linked to different spatial patterns. For example where an infectious disease may start as a local outbreak, over time, it can diffuse into different directions and can ultimately cover the complete area.

For some systems, the patterns that they produce have long been known, and the challenge is not so much in the detection of patterns but in reproducing them. This type of pattern is normally linked to processes that leave visible marks on the earth surface like urbanisation. For other systems, we do not know if they produce patterns that are robust and stable as no permanent imprint is visible to the naked eye.

Disease diffusion belongs to the last category. Although epidemics are known to produce temporal patterns, we do not always know if clear spatial-temporal patterns exist. In such a case analytical techniques have to be developed that enables the detection of patterns, and proof is needed of the stability of these patterns before they are applicable in spatio-temporal modelling.

It is not always easy to reproduce patterns as there is not a single way to model a system and new implementations can lead to new understanding of the systems around us. Agent-based modelling techniques have provided us with the possibility to combine both spatial and social processes in a single

model. This is important as the pattern that emerges can be an interplay between natural processes and the actions of human beings. This is especially true for urbanisation where the landscape dictates the suitability for building but people take the decision to build. It equally applies for epidemics, where a disease is transferred from an infected to a susceptible human in a natural way but the actions of people determine if two people are at the same location at the same time.

This study shows for a number of case studies how patterns can be reproduced using agent-based modelling, yet it also aims at developing new analytical methods to detect patterns in empirical data.

Acknowledgements

It can be difficult to finish a PhD when having a full time appointment as a PhD student or AIO, yet, it is an even more daunting undertaking when trying to complete a PhD parallel to a job and family. Completion of this work would not have been possible without the support of many friends and colleagues. To all those people who shared with me the good moments on this path, and who were there to listen to my doubts and support me when the task seemed endless, I owe you all a lot of thanks.

Many thanks go to Alfred Stein, Menno-Jan Kraak and Raul Zurita-Milla for stimulating me during the years I worked on this research. Alfred, you were one of the first to advise me to start a PhD and further aroused my interest in health applications. Menno-Jan, thank you for giving me the opportunity to finish my PhD research. Especially during the final phase of my work you gave me the time to make the necessary last steps to bring this undertaking to a good end. Raul, you always encourages and inspired me, thanks for the joined visits to conferences, your tedious detailed feedback on many of my manuscripts, the joined struggle to learn new geo-computational methods, and the work together with your MSc and PhD students. I still feel that completion of this work would not have been possible if you would not have moved into the next door office.

Progress over the past years has been unevenly spread but my two sabbaticals, the first in 2013 in Indiana USA and the second in 2017 at home in the Netherlands have been times when real progress was made. My thanks therefore go to Dr. Suresh Rao and Dr. Tatiana Filatova for receiving me during these sabbaticals and supporting me in my seemingly impossible pursuit.

Furthermore I would like to thank all my colleagues and former colleagues in the department of Geo-Information Processing (GIP) of the University of Twente. My special thanks go to Corné van Elzakker who supported me during my initial steps on the road of research. Thank you Corné for teaching me the basic academic skills I needed to conduct independent research and for your listening ear and valuable advise during my complete research period. Thanks also to Rolf de By for helping to finish my first article and his willingness to read and provide feedback on my writing. To Bas Retsios who supported me and my students in our first attempts to code agent-based models. And to all other colleagues and former colleagues in the GIP department, Parya, Andre, Frank, Ton, Wim, Willy, Richard, Jolanda, Rob, Javier, Barend, Lyande and all the others, thanks for your support.

I was also very lucky to find many colleagues from other departments encouraging me to continue and providing important suggestions and

motivation to continue. Thank you all for believing in me and a special thanks to Sherif Amer for his encouraging words and sharing his own PhD experiences. To Wietske Bijker for sending cartoons and lending me books about pitfalls for PhD students and already warning me about the post-PhD dip. The many lunch walks were also very inspirational. Thanks also to Johannes Flacke for sharing my interest in agent-based modelling and the joined work on modelling informal settlements. Frank Osei, for the collaboration on the cholera study for Ghana. Your expert knowledge about the local situation proofed to be essential and to Zoltan Vekerdy for his coffee and listening ear.

I would like to thank the Female Faculty Network Incentive Fund of the University of Twente for their financial support. RIVM and Nicoline van der Maas for providing me with pertussis data. To Mirjam Bakker and Ente Rood from KIT for the many courses we organized together in which they shared their expertise on health.

I always believed that the best way of learning is to teach a particular topic. Therefore I am grateful to all my students, and especially my MSc students that via their research inspired me and contributed to my understanding and skills. Thanks to you Shaheen Abdulkareem for sharing my interest in ABMs and your valuable friendship throughout the years. A special thanks also to Juliana Usya, and Tom Doldersum for their work on the cholera model and to Sietske Tjalma for letting me use your pertussis model for the experiments of chapter 6.

My most sincere thanks go to my family Denie and Else for their loving support during these many years. Now that my PhD has ended, I hope to have more time for both of you and the many interesting projects you think up. I intend to devote the coming period to try to realise your plans and fantasies. I also would like to thank my parents Heleen and Herman Beckers for motivating me during my earlier life and for their courage to let me go, even when this meant moving to places far away. I think this was the only way to bring me back to them.

Table of Contents

List of figures

List of tables

Chapter 1 Introduction

Patterns are an important source of information for humans; we use them for their own merit but also as guiding elements in the construction and validation of various kinds of models.

Despite their importance, scientists still struggle to develop algorithms to detect patterns and to build models that reproduce these patterns. Numerous factors hinder the recognition of patterns, such as the difficulty of analysing large volumes of data and the fact that incomplete datasets make it difficult to detect patterns. What we observe is often not a single pattern but the result of multiple processes leading to a variety of patterns that emerge simultaneously and this makes detection difficult. A further complicating factor is the fact that not all patterns are meaningful.

Many disciplines, both technical and application oriented, work on patterns. A technical domain like data mining concerns itself with finding patterns in large datasets, geo-computation tries to develop learning algorithms to reveal patterns, cartographers try to visualize them, and modellers try to reproduce them. Application domains such as ecology, epidemiology, climatology, urbanism and many others use and study spatio-temporal patterns.

Despite the difficulties listed above, patterns are the only footprints left behind by processes that have great importance to humans, and we have no alternative besides trying to reveal the useful information they contain. This PhD thesis aims at making a small contribution to this huge scientific endeavour. In particular, this thesis focuses on the problem of understanding geospatial patterns from a technical perspective. The coming sections of this introductory chapter contain a reflection on geospatial patterns (1.1), clustering for patterns recognition (1.2) and how patterns are used in agent-based modelling (1.3). This general overview is illustrated by examples from two application domains: urbanism and epidemiology. These application domains were chosen to link to the papers discussed in the following chapters. The chapter ends with the research objectives and the thesis outline (1.4 and 1.5).

1.1 Geospatial patterns

A geospatial pattern is a regular, recognizable (repeatable) arrangement of phenomena on the earth surface. In this thesis, when we refer to a spatial or spatio-temporal pattern, we in fact mean a geospatial pattern. Spatial patterns are the traces left behind by self-organizing systems. Self-organization means that systems evolve to a steady organized state based on local interactions. Many systems in biology and ecology show this type of regularity. Self-organization is thought to be the result of bottom-up interactions (driven by

the lowest level). Spatial pattern formation is defined as the ability of systems to self-organize into spatially structured states from initially unstructured or spatially homogenous states.

Spatial Patterns

Spatial patterns are just as much concerned with the space between objects as with the objects themselves. A pattern of building blocks for example is determined by the blocks but perhaps even more by the space between blocks. Spatial patterns occur in many natural objects around us (Figure 1-1) and come in many different forms. In a simple classification, spatial patterns are grouped in point, line and area (polygon) patterns. This classification refers more to the geographic data types than to the actual grouping of the phenomena. Why could a pattern not consist of a combination of points, lines and polygons? Actually, if we closely look at the moss and bark in Figure 1-1, we see areas are delineated by lines and it is hard to determine if these are area patterns or line patterns as, without the lines, the areas would not be there.

Figure 1-1 Mirco patterns on a sea shell (left), moss (middle) and tree bark (right)

When studying the spatial distribution of points there are three possible patterns: uniform, random and clustered (Figure 1-3). The distance between the points plays a key role in the classification of the patterns. In uniform patterns, the points are spaced evenly with the distance between objects roughly the same. This is not the case in random patterns where there is no structure between the points. Clustered patterns show a grouping of points that are closer together compared to the other points in the same space.

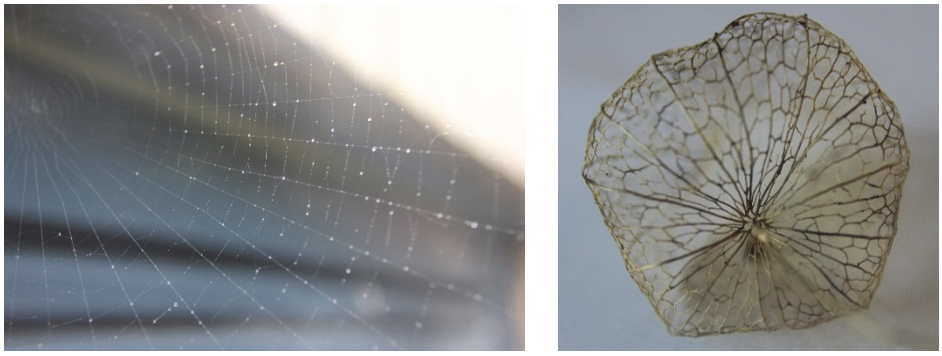

Figure 1-2 Radial Line patterns.Spider web (left) Seed of a Chinese Lantern, Physalis alkekengi *(right)*

Line patterns like point patterns can be uniform, random or clustered. What differentiates lines from points is the fact that they have a length, shape, and orientation. All of our networks (rivers, roads, electricity lines) form linear patterns. Networks have an additional property as they show connectivity. The endpoints of lines touch each other. Many linear patterns have a specific type of organization. Uniform linear patterns are grouped as patterns of straight lines, curved lines or radial patterns (Smythies, 1957, Wilson, 1986). Straight lines can be parallel lines, chessboard line patterns like the Manhattan road structure, or herringbone patterns. Curved line patterns can be spirals, whirlpools, ring patterns or sinus waves. Spider webs are a good example of a radial pattern (Figure 1-2).

Figure 1-3 Different patterns of dispersion. Ranging from clustered (left) to linear deposition (middle left) to homogeneous (right).

Two types of area patterns are found, patterns based on the shape of the areas and patterns using administrative units to display incidence rates of a phenomenon. When we visualize this type of data, we automatically see the spatial areas which are not related to the data we are studying.

Point patterns are applied in both spatial epidemiology, e.g. for the occurrence of diseases (Gatrell et al., 1996), but can also be applied on the distribution of cities or buildings when modelled as points (Pissourios et al., 2012).

Table 1-1 Patterns of points, lines and areas in epidemiology and urbanisms

	point	**line**	**area**	
			incidence	**shape**
Epidemiology	Clustering of disease	Linear clusters along rivers	Area clustering of incidence	
Urbanism	Clusters of cities or services e.g. business districts within cities	Infrastructural patterns Metrics referring to city boundaries	Clustering of areas based on population density	Shapes of building blocks, or of the total urban area.

When a disease is water born like e.g. cholera, linear patterns around rivers may exist. Urban areas contain many different linear patterns, for example in their road structures. These can be ring patterns (e.g. in Amsterdam following the canals) or spider web structured with a dense road network in the centre of the city and longer roads at the outskirts.

Urban shapes can easily be captured via satellite images but this is not possible in spatial epidemiology. Where a city consists of building blocks, roads and other tangible spatial objects, a disease has no spatial footprint of its own and no permanent appearance (See Table 1-1). Thus, where spatial patterns in urban systems are easily detectable, spatial patterns in epidemiology are far more difficult to find.

Change in space and time – Spatio-temporal patterns

Patterns are regarded as spatio-temporal when a spatial pattern develops, changes or disappears over time. We can differentiate between two types of spatio-temporal patterns: patterns that are *transient* and patterns that *stabilize*. Transient patterns appear, and disappear over time whereas stable patterns remain for a long time once they are formed. Epidemics can be regarded as periodic outbreaks, leading to patterns that appear and disappear again, followed by a re-appearance in the next epidemic. Urban systems lead to patterns that tend to be stable over longer periods.

Examples of processes leading to spatio-temporal patterns or change in these patterns are movement, growth, and diffusion. Spatial phenomena can be described via their location, shape, size and their attribute condition or state. Change in spatio-temporal pattern can be related to change in one of these four elements or their combinations.

Claramunt and Thériault (1996) provide a typology for change in spatio-temporal features including three categories: Evolution of a single entity, functional relationships between entities and spatio-temporal processes between several entities.

In many systems, the formation of patterns require an activator (driving force) and an inhibitor (repressor). To produce the ripples on the sand of a beach we need the wind and water as driving forces and the sand as the inhibitor. As wind speed, wind direction and the beach surface (position of the sand particles) are constantly changing the pattern of the sand is also changing over time and can be regarded as a spatio-temporal pattern (Figure 1-4).

Systems with multiple driving processes are considered to be complex systems. Such complex systems are characterized by the fact that they show a form of self-organisation, also referred to as spontaneous order. In this thesis, we simply say that patterns emerge. Order arises from local interactions. We call this order spatial complexity when multi-scale spatial patterns emerge.

Driving forces in urbanism include population growth, land markets, economic factors etc. The landscape itself can be seen as a repressor because unfavourable building conditions occur for steep slopes, or wetlands. In the case of epidemics, the spatial distribution of the population (spatial population separation), human mobility and non-spatial factors like replenishment of susceptible hosts after epidemics (Grenfell et al., 2004) can be seen as drivers.

Figure 1-4 Pattern on a beach, the result of an interplay of different systems of wind, water and sand.

To get a better grip on spatio-temporal patterns, we need to understand the temporal component of the processes that produce these patterns. The trajectory of a system is the order in which the systems moves from one state into another. In principle a chronological system is a system that progresses from an initial state to a final state without loops (cyclic trajectory) or parallel trajectories (branching system).

Within the group of chronological systems, we can differentiate between a linear system and a sequential system. This is done based on the progression in time between the states. The time between events can progress evenly leading to gradual change in the system. If we move forward double the amount of time, the change in the system will be double. This is a completely linear system.

In sequential systems, a number of states are visited in a fixed order, yet the time transition is uneven. No, change may happen over a certain period of time, and large changes may appear in short time frames (state transition). Sequential patterns emerging in time were described by Hägerstrand (1953) as *primary stage*, the *diffusion stage*, *condensing stage* and *saturation stage*. Each of these stages have different spatial patterns.

In both epidemiology and urbanism we find sequential patterns. For spatial epidemiology, the primary stage can be compared to the index case in epidemics, the diffusion stage is the period when the epidemic spreads, the condensing state is the stage where the disease declines and finally, in the saturation stage, the disease disappears.

In urban growth, we can recognize a primary stage (small rural settlement), followed by a period of fast growth, however, gradually other processes will start to appear like infill (condensing stage) and urban sprawl.

Complex systems normally display both gradual and abrupt changes. These are also referred to as smooth and discontinuous changes (Petraitis, 2013). Discontinuous changes are linked to transition in state of a system, the so-called phase transitions. The critical point at which the phase transition occurs is often referred to a as the tipping point, e.g. by Gladwell explaining the three important elements of tipping points as: contagiousness (1) the fact that small changes can have big impact (2) and that change happens at one dramatic moment (3) (Gladwell, 2000).

In self-organized criticalities (SOCs), a system is critical if its state changes dramatically given some small input. A dramatic state change can be compared with the collapse of a sand pile. To explain this more clearly we need to go back to the sand-pile example of Bak et al. (1989). When sand is added to the pile first small avalanches will occur but when we keep on adding sand, also large avalanched emerge. The system evolves and patterns start to appear at different scales (small and large avalanches).

Both epidemics and urban systems can be described as SOCs (Rickles et al., 2007, Chen and Zhou, 2008). Epidemics have been referred to as self-organized criticality (SOC) by Rhodes et al. (1977). When the number of people

infected with a certain pathogen in a small area reaches a threshold value, this can inevitably lead to an epidemic or even pandemic. Batty (2005) describes the transition from industrial to a post-industrial city as an example of a phase transition in an urban system.

Before and after a phase transition the complex system has a clear order that can be linked to two other concepts: power laws and self-similarity (fractals). A power law can best be explained by an example. In languages, small words are used more frequently compared to longer words or we have many small earthquakes compared to a few serious ones. This is also referred to as scale-free as there is no real scale at which we can describe these phenomena.

Self-similarity or fractals indicate that things look the same at different scales. A fern leave (Figure 1-5) is composed of a number of sub-leaves. Each of the sub-leaves again is a composition of smaller leaves. The shape of the sub-leave is identical to the shape of the complete leave. Both power-laws and fractals refer to a scale-free state.

Cities show self-similarity or spatial invariance across different scales. This can be seen in roads (Batty, 2012) where we recognize that main road patterns repeat themselves in local roads. Cities show hierarchical patterns in respect to business districts. The main center is often surrounded by a number of local business centers.

Besides the fact that we can recognize examples of complexity in a single city, we can also see complexity when we evaluate characteristics of a large number of cities. City size distributions obey the rules of power laws, in many different countries there are few large cities and many smaller cities and larger cities seem to be further apart compared to smaller cities (Hsu et al., 2014). Cities power laws apply to population, rank and areas of cities (Chen and Zhou, 2008).

Figure 1-5 Self-similarity in plants

Self-similarity can be found in both space and time. When evaluating long-term epidemic curves of a certain disease, we often see that similar patterns repeat with every outbreak. Scale-free power-law fractal behavior in time was

found by Jose and Bishop (2003), who determined different scaling regions in time-series for rotavirus dynamics.

An example of spatial fractal like patterns (self-similarity in space) was determined by Philippe (1999) for childhood leukemia in the San Francisco area. When evaluating the spatial distribution of the disease cases over seven scales it was well fitted by a power-law function.

Realizing that complex systems will produce patterns with self-similarity can be very useful in pattern detection and pattern reproduction. If we know that epidemics might be self-similar this means that we can expect that epidemics occurring at different moments in time, might show similar spatio-temporal diffusion patterns. These patterns might not be visible constantly, especially in a diffusion process with transient patterns that go through a number of diffusion phases. For the comparison of patterns, this temporal dimension might become important. Can we align epidemics so that we can compare the same stage in the diffusion process? This might also be valid in space, different areas in a country might show similar diffusion patterns during a particular disease outbreak. Yet, how do we delineate the areas that we are comparing?

1.2 Clustering to reveal spatio-temporal patterns in disease data

There are many analytical methods to detect patterns in empirical data. In this thesis, we concentrate on clustering, which is a method used to group similar objects. Clustering can be applied on many different types of input data. As we are interested in spatio-temporal patterns, we focus on time-series clustering. What makes time series clustering different from regular clustering is that the elements in a time series have a fixed order (in time) and this order should be maintained.

Time series characteristics may include trends and seasonality. Many statistical analyses start with the decomposition of the time-series, to remove long-term changes (trends) and seasonal components. There are however drawbacks to this approach. With the decomposition of the data, important information might be lost. Decomposition is important in prediction studies, yet less relevant in clustering. As we are not interested in prediction or long-term trends but in similarity and hierarchy detection within and between epidemics, decomposition is less relevant for this work.

Clustering can be performed using a range of techniques based on statistics or machine learning. Many good overview papers of time-series clustering exist including Liao (2005) and Aghabozorgi et al. (2015). These techniques can be divided into model-based approaches, feature-based approaches, shape-based

approaches and multi-step approaches (Aghabozorgi et al., 2015). This research mainly focuses on shape-based approaches which try to match the shapes of two time-series. Examples of this group of techniques are Self-Organizing Maps (SOMs), Dynamic Time Warping (DTW) and Shape Based Distance (SBD). It is not easy to select the most suitable technique for a particular purpose based on technical specifications. The most suitable clustering technique depends on the input data and aim of the clustering and can in many cases only be selected based on a set of experiments.

In section 1.1 we identified that patterns resulting from complex systems show self-similarity. For a single area, self-similarity can exist between different epidemics (i.e. self-similarity in time). At the level of an epidemic, similarity may also exist between time series of different spatial units (i.e. self-similarity in space). Therefore clustering should be applied in time and in space to try to locate/identify these self-similarities.

In shape-based clustering approaches, two time series can be similar if they have the same shape but also if they are in the same phase or have the same amplitude. Time-series of different areas within a single epidemic can be similar in shape even if they differ in amplitude (less or more disease cases) or phase (occur earlier or later).

Amplitude and phase differences are important for systems that exhibit **Hierarchical diffusion** where the diffusion occurs through an ordered sequence of classes or places (Cliff et al., 1981b). Hierarchical diffusion is often seen in epidemics, when infection trickles down from large cities to smaller towns and villages. Differences in phase and amplitude can be linked to hierarchies in the diffusion process (Viboud et al., 2006). Large cities are infected earlier compared to smaller cities and villages (phase) and will have a larger number of infections due to the larger population (amplitude). Cities with similar population sizes can show similarity in time of infection (phase synchrony).

According to Liebhold et al. (2004), **spatial synchrony** refers to *"coincident changes in the abundance or other time-varying characteristics of geographically disjunct populations"*. We can also say that two locations are aligned in time (peaks in a time series occur simultaneously). Spatial synchrony can be the result of hierarchical diffusion yet, this is not necessarily the case. When multiple cities are connected to a larger city that is infected with a contagious disease, they may all get infected simultaneously leading to spatial synchrony in disease infection. But, when Lilac trees in the US flower simultaneously with similar trees in Europe we observe spatial synchrony that is not due to hierarchical diffusion but to the temporal convergence of phenological events.

In principle there are three types of measures for spatial synchrony: correlation of abundance, phase synchrony and peak coincidence (Liebhold et al., 2004). Where synchrony refers only to the phase of the time-series, **self-similarity** can also refer to shape and amplitude. Time series are similar based on all three elements of shape, phase and amplitude.

Table 1-2 Mapping of self-similarity, synchrony and hierarchy to phase, shape and amplitude of time series for disease diffusion

		Self-similarity	**Synchrony**	**Hierarchy**
Phase or Timing	Single epidemic		Synchrony in phase between places of the same order.	Differences in phase, as e.g. a disease diffuses between places of different order (cities versus villages)
	Between epidemics	Robust pattern: spatial area is assigned to the same cluster over multiple epidemics based on time of infection (early, late)	Spatial areas are synchronized over multiple epidemics	-
Shape	Single epidemic	Similarity between spatial locations that are at the same order in the hierarchy.		Differences in shape of the time series between places of different order (disease disappears in small places – fade out and maintains itself in larger places
	Between epidemics	Similarity in shape between time series for the same location during multiple epidemics		-
Amplitude	Single epidemic	Similarity between spatial locations of similar hierarchical order		Difference between spatial locations of different hierarchical order
	Between epidemics	Similarity between time series for the same location during multiple epidemics		-

A full overview of how the concepts of self-similarity, synchrony and hierarchy relate to the phase, shape and amplitude of time series is provided in Table 1-2.

When clustering time-series of disease data we can limit ourselves to a single epidemic. However, we can also split larger time series containing multiple epidemics into sub-sequences . By clustering time-series that include multiple epidemics per spatial area, we can observe self-similarity, or robustness of the observed patterns. However, this analysis requires the that the time series are aligned in phase. How this can be done is one of the research questions of this thesis.

When clustering time series, one must realize that each time-series represents a certain spatial area and that the delineation of these areas is crucial. For instance, we know that hierarchical systems show different patterns at different levels, and that a diffusion process is mainly driven by areas with high population. Hence clustering analysis should focus on highly populated areas. The spatial delineation of the clustering input will also be addressed in this research.

Various factors might hamper the detection of spatio-temporal patterns in disease data, including the fact that notification data is often incomplete, and diffusion processes happen at a fast speed. Moreover, it is unclear how robust epidemic patterns are in space and time. Will similar diffusion patterns be found in different epidemics or does every epidemic follow its own diffusion path? Regularities are needed as only robust patterns can be used in modelling. If robust diffusion patterns are found in empirical data, can these patterns be reproduced via agent-based modelling? Or are there still missing pieces in our understanding of the complex process that produces these patterns?

1.3 Agent-Based Modelling

Where clustering focusses on the detection and recognition of spatio-temporal patterns, agent-based modelling is concerned with the reproduction of these patterns. Modellers want to reproduce patterns because they provide information about the system being modelled. If a model is unable to reproduce observed spatio-temporal patterns, it is very likely that the system that produced these patterns is not correctly represented in the model. Modelling techniques that can reproduce emergent behaviour, e.g. Agent-based modelling (ABMs), can be used to model and reproduce complex patterns.

Agent-based models are based on individual-based and bottom-up modelling approaches. They assume that systems are emergent and that by applying a

limited number of rules (behaviour) at the individual level (agent), complex systems can be modelled.

Emergence is not the result of what was modelled (programmed in) explicitly but of what is not imposed. Railsback and Grimm (2012) identify three essential points:

a. Emerging properties are not the sum of the properties of the agents.
b. They are not individual level properties.
c. They cannot be easily predicted or intuitively derived from the properties of the model components.

The level of emergence varies per ABM model. No emergence is an indication that the model could also be implemented as a mathematical model, and too much emergence leads to chaos which cannot be interpreted. To be able to detect patterns in model output, it is important that the scale at which these patterns occur matches the scale of the model. Thus, the need to represent a given pattern dictates the minimum level of detail for the model. In many studies multiple patterns are used simultaneously in the design and implementation of the model as simultaneous fulfilment of multiple patterns is non-trivial (Wiegand et al., 2002) .

Pattern-Oriented Modelling (POM) was introduced as a framework to develop agent-based models (Grimm et al., 2005). POM is an integrated approach, starting with the model development and leading to a better validation by two means: it makes the model structurally realistic and therefore less sensitive to parameter uncertainty (Grimm et al., 2005) and it helps in the calibration process as parameters can be fitted to multiple patterns. Hence, POM addresses two challenges of ABM: complexity and uncertainty.

A key characteristic of POM is that it always uses multiple patterns. One of the reasons for this is that although a weak pattern cannot be used on its own, a combination of weak patterns may be useful. Another reason is that multiple patterns can be linked to different hierarchical levels. The use of multiple patterns linked to alternative hypothesis allows for comparison of these hypothesis (Grimm et al., 2005).

ABM models that simulate systems that show self-organized criticalities (SOCs; section 1.1.) should simulate state transitions. State transitions should be emergent and patterns before and after the tipping points will be different. The total duration of the simulation, in respect to the rate of transitions, is crucial to determine whether a state transitions will occur. Thus, simulations should run long enough so that transitions can be generated.

Constructing models that reproduce observed patterns is not enough. We also need analytical tools to detect patterns in empirical data and to compare

patterns in empirical and simulated data. This comparison is often conducted via similarity metrics (e.g. clustering distances). However, no single metric can capture the typical complexity of spatio-temporal patterns. In addition, empirical and simulated data are structurally different. Empirical data is typically incomplete and ABM models produce complete datasets. The comparison between simulated and empirical patterns is one of the research questions of this thesis.

There are several examples of ABM in epidemiology. ABM models exist for influenza (Lee et al., 2008, Mniszewski et al., 2008, Yahja, 2006, Germann et al., 2006), smallpox (Chen et al., 2004) and malaria (Linard et al., 2008). The validation of the observed patterns is typically done using the R_0 metrics, which represents the number of follow up infections caused by a single initial infection. Nevertheless, this metric cannot capture the complexity of spatio-temporal diffusion pattern. Being a purely temporal metric, R_0 does not capture the spatial dimension of the diffusion pattern. Even as a temporal measure, R_0 captures only the explosiveness of the infection and has no ability to differentiate between time series with single and multiple peaks. An important research question is if an alternative metric or approach can be developed so that we can get a better representation of the spatio-temporal patterns and also use it to compare patterns in empirical and simulated data.

Spatio-temporal patterns of disease diffusion are influenced by the distribution of infectious individuals and by their mobility, which also has spatio-temporal patterns by itself. This raises the following question: can links be found between disease diffusion and mobility patterns by means of ABM simulation? Disease models are normally age structured. There are also differences in mobility patterns between age groups. Can links be made between a disease model and age-structured mobility?

There are also several examples of ABMs in urban applications. One of the first examples is the urban segregation model of Schelling (Schelling, 1969). This model simulates agents that decide, based on the fraction of neighbours that belong to their own group, if they want to stay or move away. Although the model is simple, it is able to generate integrated, segregated and mixed patterns and it is still used by many more complex models (Hatna and Benenson, 2015).

Cities are the result of individuals that decide to settle somewhere or move away. ABM facilitates the implementation of this kind of behaviour, enabling agents to decide on the suitability of sites. Urban growth can then be simulated by letting agents optimize their state based on a number of environmental and socio-economic factors. This location choice process was described by Benenson (2004) as including the following stages: assessment of one's

residential situation, decision to attempt to leave, investigation of the available alternatives, their utility estimated compared to that of the current location. Besides individual decision making human mobility is also important aspects of urban growth that can be integrated in ABMs (Huang et al., 2013). Most urban models are made for developed economies and little attention is given to simulation of informal settlements in developing countries. Moreover, most urban models are developed in the raster domain overlooking the fact that ABMs can capture a level of spatial realism that can produce a richness in pattern that cannot be represented as raster cells. In this PhD thesis, an effort is made to fill this gap and produce a model with detailed realistic spatial outputs. This type of output can help when comparing simulated data to empirical data from informal settlements.

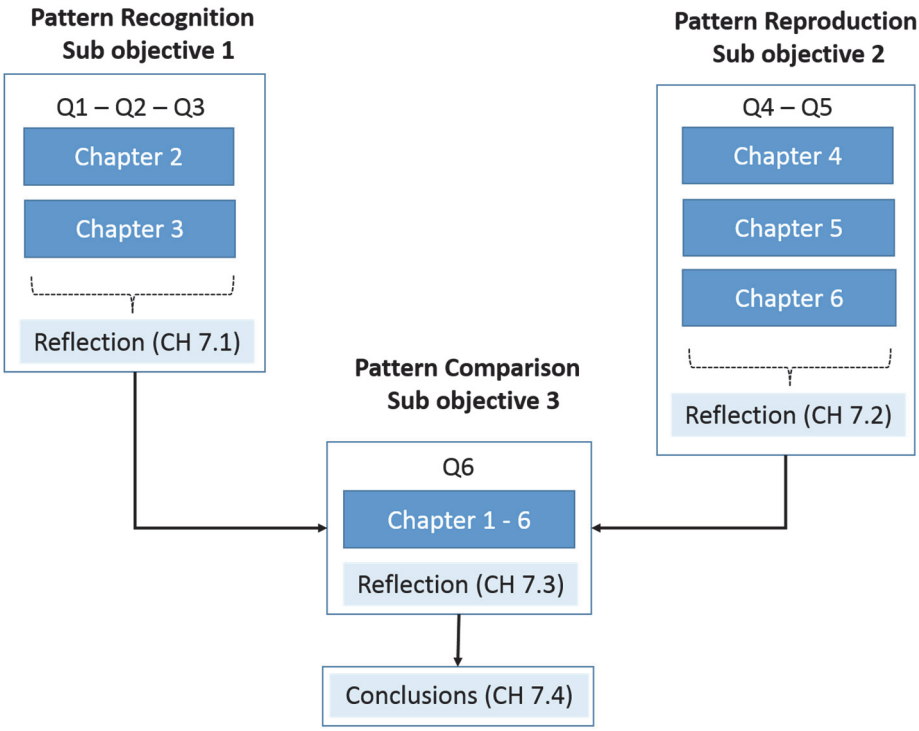

Figure 1-6 Conceptual model of relationships between the components of study

Although at first glance epidemiology and urban growth seem to be different disciplines, they share common grounds because the distribution of people and their mobility play an important role in the emergence of complex spatio-temporal patterns in both domains. Perhaps links can be identified in pattern recognition and POM for epidemiology and urbanism. in this PhD thesis we develop methods to detect spatio-temporal patterns in empirical and in

simulated data. These methods can enhance modelling approaches, and ensure a good validation of ABMs.

1.4 Research Objective and Research Questions

POM cannot be applied unless clear patterns have been discovered. In this era of growing datasets, new techniques are needed for empirical pattern detection. These techniques should be suitable for self-organizing complex spatial systems and must be able to detect robust spatio-temporal patterns. Hence, the main objective of this PhD is to:

"To design new approaches to the use of spatio-temporal patterns for building and validating ABMs".

This main objective splits into three specific objectives, which are operationalized by means of 6 research questions, and illustrated with examples from two scientific domains (epidemiology and urban studies):

a. To develop and evaluate pattern detection techniques that can recognize (robust/self-repeating) spatio-temporal patterns that can be used to build and validate ABMs.

   Q1. How can time-series clustering be used to identify diffusion hierarchies and spatial and spatio-temporal self-similarity?
   Q2. *How should the data be spatially aggregated to maximize the likelihood of obtaining meaningful clusters?*
   Q3. *How should time series be aligned to be able to compare epidemics using clustering(-based) approaches?*

b. To develop and evaluate methods to use these patterns when building geographically explicit ABMs.

   Q4. How can models that generate more detailed (vector based) simulation outputs help to compare simulated and empirical data?
   Q5. How important is the use of spatio-temporal patterns, compared to temporal and spatial patterns, when building geographic ABMs?

c. To develop and evaluate methods for the comparison of simulated and empirical patterns so that ABMs can be validated.

Q6. What are the factors that hinder validation of agent-based models based on spatial-temporal patterns and how can the comparison of empirical and simulated patterns be improved?

An Overview of the main methodology is provided in Figure 1-6.

1.5 Thesis outline

This PhD thesis consists of seven chapters. After this *Introduction*,
Chapter 2 focuses on the detection of spatio-temporal diffusion patterns in measles epidemics on Iceland. In this chapter, we develop a Self-Organizing maps (SOMs)-based method to detect diffusion patterns and use the Sammon projection to compare spatio-temporal diffusion patterns between epidemics.

In ***Chapter 3*** *we discuss the use of* time series clustering to find similar diffusion patterns in disease data. When comparing outbreaks from different years, aligning the data is important to be able to find similarity. When a time series is slightly misaligned along the time axis (e.g. the epidemic starts in a different month), many traditional methods find a low correlation with other time series for the same area. Here we use a shape-based clustering approach to identify Dutch urban areas with similar pertussis patterns. The percolation method was used to identify the main urban areas in the Netherlands.

In ***Chapter 4*** we develop an ABM for informal settlements in Dar-es-Salaam, Tanzania using a vector based implementation. This model shows how the organization of building patterns changes based on a few simple rules on site selection and alignment of buildings.

In ***Chapter 5*** *we develop* another ABM model. This time to test a hypothesis of run-off water from open dumpsites as a mechanism of cholera diffusion in Kumasi, Ghana. This model evaluates both spatial and temporal patterns using a limited input dataset.

In ***Chapter 6*** the results of an ABM for pertussis in the Netherlands are compared to empirical datasets. Different simulated patterns are created by varying the starting point of the infection and the mobility model. Model and empirical patterns are compared using the method developed in chapter 2.

In ***Chapter 7*** we provide a synthesis of the results found in this PhD thesis, including the answers to the research questions. In this chapter we also reflect on future research directions in both pattern detection and agent-based modelling.

Chapter 2 Self-Organizing Maps as an approach to exploring spatiotemporal diffusion patterns[1]

2.1 Background

Spatiotemporal analysis of epidemic waves can reveal important information on anomalies and trends, and provide inside into the underlying diffusion patterns (Viboud et al., 2006, Cliff et al., 1981b). These patterns are categorised as contagious spread, hierarchical spread, or mixed diffusion. Contagious spread depends on direct person to person contact and results in centrifugal patterns from the source outward (Cliff et al., 1981b). Hierarchical spread refers to disease transmission through an ordered sequence of geographic locations (normally based on their size) (Viboud et al., 2006) and it can be related to the movement of people, carrying a disease to a new centre of population via long distance travel. Due to this, hierarchical spread is typically characterized by the display of synchrony among locations that have similar size but that are geographically apart (Cliff et al., 1981b). Two or more locations are synchronized when they exhibit a parallel development in the number of disease cases.

The search for synchrony is not unique to epidemiology but originates in innovation diffusion and ecology, and it occurs in many other disciplines (Andrew Liebhold et al., 2004). Hence, multiple methods exist to quantify and to map synchrony (Bjornstad et al., 1999). Among these methods wavelets are frequently used as they also allow to study non-stationary (trends) in time series (Cazelles et al., 2007). Wavelets analyse disease diffusion in the frequency domain where synchrony can be identified via the coherence in the phase of the number of diseases cases at each geographic location (Grenfell et al., 2001).

Besides synchrony, another important property of spatiotemporal disease diffusion is the trajectory of wave propagation. This trajectory captures the step by step diffusion by describing the speed and direction of spread (Cliff et al., 1981b). As waves of infectious diseases are normally a combination of contagious and hierarchical spread (Cliff et al., 1981b), this trajectory is not a single and continuous line (as a trajectory representing human movement) but a reflection of a moving front or fronts. Methods for capturing this movement

[1] *This chapter is also a paper co-authored with R. Zurita-Milla "Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns." International Journal of Health Geographics 12(1): 60.*

range from different calculations of front velocity (Cliff et al., 2008), to methods that capture the direction of diffusion as related to clusters of human population, network distance or travel distance (Brockmann et al., 2006, Hufnagel et al., 2004).

In this paper, we propose using self-organizing maps (SOMs) to study disease diffusion in space and time. SOMs are a well-known data-mining method, used to cluster and visualize high dimensional data by projecting it into a low-dimensional (typically 2D) space (Kohonen, 2001). This projection makes it easier to understand spatiotemporal datasets and the patterns that they might contain (Andrienko et al., 2010, Wang et al., 2013). In spatial-epidemiology, SOMs are mostly used as a non-linear analytical method to study multivariate patterns (Wang et al., 2011, Basara and Yuan, 2008, Koua and Kraak, 2004) but here we show that they also enable the integrated analysis of both synchrony and diffusion trajectories. Moreover, this data mining methods is advantageous because it does not require transforming the data to a new "data space" (like wavelets). This greatly facilitates the interpretation of the results as the shape of the epidemiological curve (number of cases as a function of time) is preserved so that the time of infection and intensity (persistence) can be studied for each geographic location.

The detection of synchrony using SOMs is based on the fact that they maintain the topological characteristics of the input data. This ensures that locations with a high level of synchronisation in the timing and intensity are mapped near to each other forming clusters. The study of diffusion trajectories can be achieved by further applying the Sammon's projection to the previous SOM results. In short, in this study we illustrate the following issues: identification of locations (spatial units) with similar diffusion processes – synchrony (1) and characterization of spatial temporal diffusion patterns – diffusion trajectories (2).

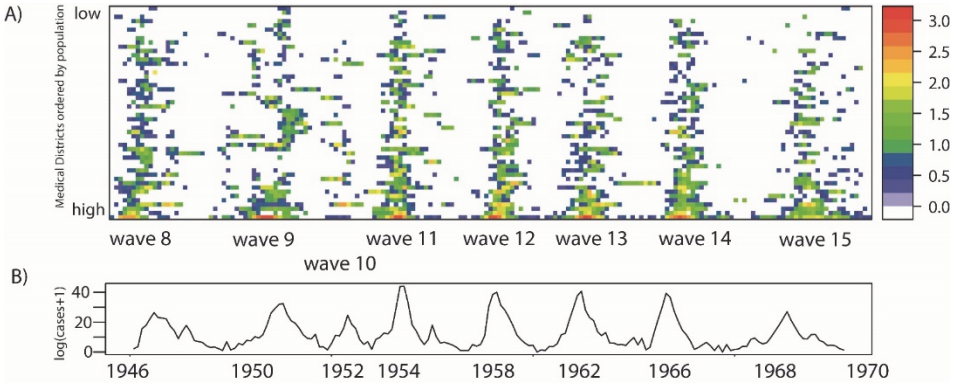

Figure 2-1 test Measles dataset. Measles cases in Iceland (1946-1970). ***(A)*** *Spatial distribution of log10 (cases +1) per medical district, with medical districts arranged in ascending order of population size and color representing epidemic intensity.* ***(B)*** *Time series of total number of notified cases for the complete country (log10 (cases +1)).*

2.2 Methods

Disease data

To illustrate this study, we used data on eight historical Measles epidemics in Iceland (Cliff et al., 1981a). The epidemics span the period November 1946 (wave 8) to December 1970 (wave 15). Prior outbreaks took place (waves 1-7), but are not included in this research for reason of data incompleteness and re-organization of medical districts. After 1970, outbreaks have different characteristics due to the introduction of mass vaccination.

The data reports monthly Measles cases for each of the 50 medical districts of the country. Figure 2-1a shows the log transformed Measles cases ($\log_{10}$ (cases +1)) for all of the eight epidemics per medical district, with the medical districts sorted in ascending order of population size. The colour representing the epidemic intensity shows that for all waves, there are only few medical districts with high intensity and that these are always the centres with the higher population (bottom of the graph). The number of cases per epidemic outbreak ranges between 6000 cases for wave 9 and less than 1900 for wave 10 (Figure 2-1b) (Cliff et al., 1981a). Because of the low number of cases wave 10 is excluded from further analysis. Inter wave periods are evenly distributed showing no significant changes in pattern over the studied time period. All presented analyses are performed on log transformed input data to ensure a normal distribution and are scaled (0-1).

This dataset was selected because Iceland has proven to be an excellent study-example for disease diffusion processes for a number of reasons including: the isolation of the country which creates a self-contained system with few external

influences; the stability of the population and spatial structure (medical districts), and length of the available time series. This has led to the well-documented and extensively studied Measles dataset (Cliff et al., 2009a, Cliff et al., 1981b).

Self-Organizing maps (SOMs)

SOMs are a type of un-supervised artificial neural network used to cluster high dimensional data by projecting it onto a low-dimensional lattice. This lattice consists of neurons that are trained iteratively to extract patterns from the input data. These patterns are generalizations of the input data and are referred to as codebook vectors. At the start of the training phase, each neuron is assigned a codebook vector that is updated at each iteration, in such a way that topological properties in the input training data are preserved.
We used the Kohonen R package (Wehrens and Buydens, 2007) to train several SOMs following these steps:

a. The size (number of neurons, including number of rows and columns) and type (rectangular or hexagonal) of the SOM lattice were chosen.
b. Each neuron was assigned a random vector of weights or codebook vector (m_k) with the same dimensionality as the input data.
c. Data samples were iteratively presented to the low-dimensional lattice to identify the best matching unit, BMU, which is the neuron that contains the codebook vector that minimizes the Euclidean distance with the data sample at hand. This iterative process is known as training the SOM and each iteration (t) is used to update the codebook vector of the BMU and the neighboring neurons according to:

$$m_k(t+1) = m_k(t) + \alpha(t) h_{ck}(t)(x(t) - m_k(t)), \tag{2.1}$$

in which m_k is an n-dimensional codebook vector, α (t) is the learning rate, $h_{ck}(t)$ is the neighbourhood kernel of the BMU neuron and *x* is a randomly chosen input vector from the training dataset.

For the training of the SOM we used a hexagonal SOM lattice, using a standard linearly declining learning rate from 0.05 to 0.01 over 1000 iterative updates. The radius of the neighbourhood kernel uses the starting value of 2/3 of all unit-to-unit distances using a square neighbourhood.

After the training, a secondary clustering can be performed on the SOM lattice, using visual analytics or a different clustering algorithm. Especially when the training lattice is large (larger than the number of clusters needed), a secondary clustering is known to outperform the initial SOM (Vesanto and

Alhoniemi, 2000). Secondary clustering can be performed via visual analytics or by using a second clustering algorithm.

A relatively simple way of identifying SOM clusters is by using the U-matrix. The U-matrix displays the Euclidean distance between the codebook vectors of neighbouring SOM neurons. High values in the U-matrix visually separate clusters. However, this method has proven to be difficult, especially with complex datasets. Therefore, several authors have proposed graph-based technics to enhanced the U-matrix for cluster interpretation (Tasdemir and Merenyi, 2012). Here we used an enhanced U-matrix as proposed by Hamel and Brown (Hamel and Brown, 2011). In this method, the centres of the lattice neurons are used as the vertices of a planar graph (a graph without crossing edges). The edges in the graph connect nodes to the neighbouring node with the maximum gradient. In this way, subgraphs are created, that indicate the clustered neurons. When displaying the graph on top of the U-matrix, an easy visual interpretation of the number and composition of the clusters is possible.

Besides the visual identification of clusters based on the U-matrix, a different clustering algorithm can be used for secondary clustering. A range of options exist including k-means and hierarchical clustering (Vesanto and Alhoniemi, 2000). In hierarchical clustering, neurons are first assigned to their own cluster, the distance between clusters is calculated and then, iteratively, the most similar clusters are joined. A disadvantage of these methods is that user has to decide the number of clusters to be found.

After training the SOM and performing the secondary clustering, the third step in the SOM process is the mapping of the data onto the trained SOM, identifying for each input vector the BMU neuron and cluster. The training dataset and mapping dataset can be the same, subsets, or mapping data may consist of new data not included in the training sample. Here, we trained the SOMs with the complete dataset to ensure that all existing patterns are represented in the codebook vectors of the lattice. However, different subsets of the training data are mapped back onto the lattice for evaluation. These subsets correspond to single epidemic waves, making it possible to compare the mapping of the total dataset to the mapping of the individual waves.

The standard way to quantify error for trained SOMs is the *quantization error,* which measures the distance between the mapping data and the codebook vector. In this research, the quantization error is used to evaluate the "goodness of fit" of the mapped data. The smaller the quantization error, the better the mapping.

When applying SOMs for spatiotemporal analyses, the data used for training and mapping needs to be considered in a dual fashion: from a spatial

perspective and from a temporal perspective (G. Andrienko et al., 2010, Wu et al., 2013). A data organisation of the type space over time (SxT) allows the detection of spatial units (medical districts) that show similar behaviour over time; that are synchronized. Here, we would like to find both synchronies over the total time series and over single waves. Therefore, two different types of space over time datasets are used: the space over time (SxT) dataset where T includes the complete time series (Figure 2-2a) and the space over wave (SxW) dataset where T covers a single wave (W) (Figure 2-2b). To study the diffusion of the disease over time, a time over space (TxS) data organization is used (Figure 2-2c).

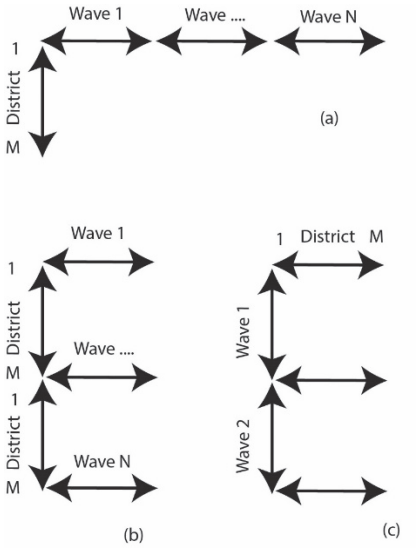

Figure 2-2 Data organization. Data organization Space in Time (SXT) SOM (A), Space over Wave (SXW) SOM (B) and Time in Space (TXS) SOM (C).

Finding clusters of synchronized codebook vectors

Finding synchronies based on SOMs makes use of the combined ability of the SOM method to produce a generalized prototype vector from the input data and to order these vectors topologically onto a training lattice. Input data vectors that map to the same neuron are synchronised, vectors that map to neighbouring neurons might also be synchronised. This can be identified by performing a second clustering on the SOM lattice grouping neighbouring neurons with similar codebook vectors.

Detection of synchronies between medical districts is based on an SxT data organisation. The training of the SOM (lattice size 3x4) is followed by a secondary partitioning based on hierarchical clustering (See Figure 2-3).

However, the exact number of clusters is unknown. This is why the clustering is confirmed using the Component Planes of the temporal SOM (Figure 2-3, step a) with a lattice size of 7x7. Component Planes are slices of the codebook vectors that represent the status of a variable for all the neurons in the SOM lattice. The correlation among variables becomes visible via similar patterns in their component planes. Methods for using Component Planes for correlation hunting have been described previously (Vesanto, 1999, Himberg, 1998).

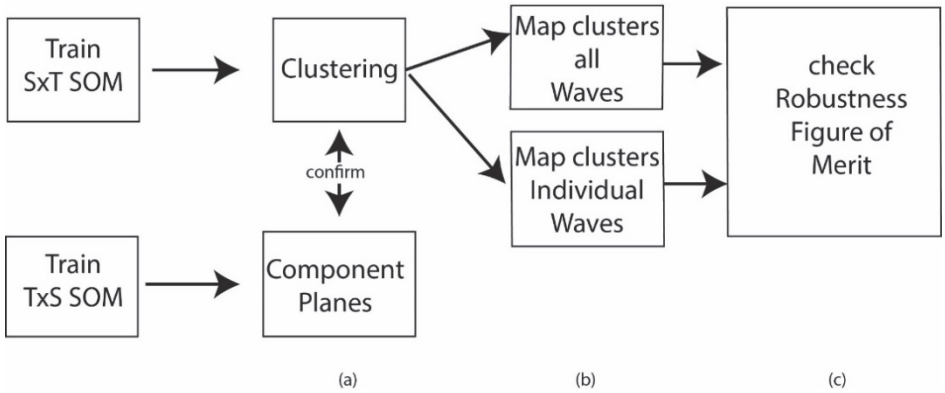

Figure 2-3 Flow diagram synchrony. Flow diagram showing the steps to identify synchrony. A. - clustering, B - mapping of the dataset, C - check on the robustness of the clusters

We can re-organise our dataset in order to construct a dataset were each data vector represents one month (Figure 2-2 – data organisation (c)). This data organisation is also referred to as Time in Space (TxS). We trained this SOM using a lattice size of 7x7 neurons. When using the TxS SOM, variables represent spatial locations. A component plane in this case, is a representation of all the neurons a medical district has been mapped to, including the frequency. Two spatial locations with identical or similar component planes are correlated. We compared the clustering found with the SxT SOM with the component planes of the TxS SOM to verify the number of clusters. This was done by grouping the component planes of the medical districts per cluster.

After confirmation of the clustering, both the complete dataset and the individual waves are mapped back onto the SxT SOM lattice to determine the BMU (Figure 2-3 step b). In SOMs, training vector and mapping vector should have an equal length. However, a single wave subset is much shorter than the complete time-series. Thus, subsets of input vectors were created by combining a Nodata matrix with subsets of the scaled input data. This is possible because SOMs allow for "missing data".

When training the SOM with the complete dataset and mapping back single waves, clustering found in the complete dataset may differ from the clusters of individual waves. The robustness of the clusters was checked via the so-

called figure of merit (Levine and Domany, 2001). The figure of merit (M(v)) measures the extent to which the clustering for the subsamples (individual waves) corresponds to the clustering of the complete dataset for the variable or variables v, in our case the disease incidence. Mapping can be presented as an *N x N* mapping matrix τ_{ij} in which τ_{ij} = 1 when two medical districts are mapped to the same neuron or cluster, and τ_{ij} = 0 when mapped to different neurons or clusters. The figure of merit is based on the comparison of mapping matrices of the resamples $\tau^{(\mu)}$ and the original matrix τ per subset (w), in our case a wave:

$$M(v) = \frac{\sum_w (\frac{\delta \tau_{ij}, \tau_{i,j}^{\mu}}{\tau_{i,j}})}{w} \quad (2.2)$$

M(v) is used to compare the mapping of all the subsamples with the mapping of the complete dataset by counting the number of times the same mapping occurs in both samples and dividing this by the total number of subsets. M(v) = 1 indicates a perfect score.

SOMs for Identifying diffusion patterns

SOMs can also be used to identify diffusion trajectories. For this, we followed two steps: First, we grouped epidemics based on their diffusion pattern. Next, we visualised the synoptic states and created diffusion trajectories (Figure 2-4).

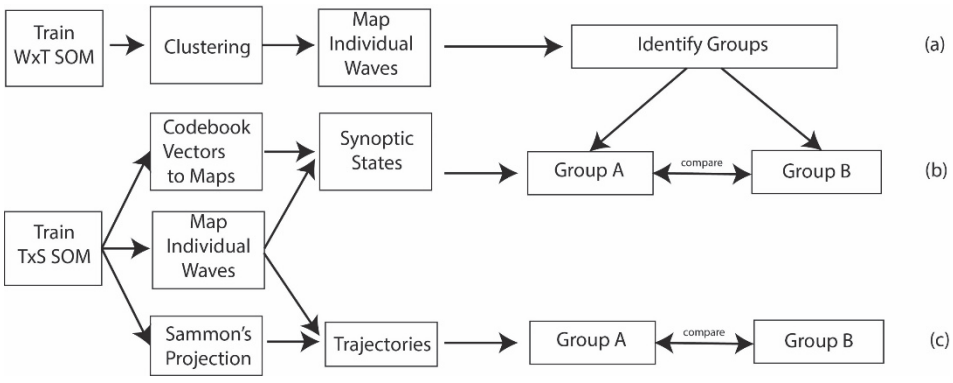

Figure 2-4 Flow diagram spatial diffusion. Flow diagram showing the steps for the identification of diffusion. A - grouping of similar waves. B - comparison of synoptic states of the groups identified under A. C - mapping of diffusion trajectories of all waves onto the Sammon's Projection, and comparison of the trajectories of the groups identified under A.

Grouping waves with similar diffusion patterns

Grouping of epidemic waves with similar characteristics is done based on the SxW SOM (Figure 2-2 – data organisation (b)). For the SxW SOM, codebook vectors represent a medical district during a single epidemic. This SOM maintains the epidemic curve in the purest form, and allows for a high level of topological consistency. After training (lattice size 7x7 neurons), a secondary grouping is performed using the enhanced U-matrix method, and the individual waves are mapped back onto the SOM lattice (Figure 2-4 (a)).

A limitation of the SOM algorithm is that all input vectors should be equal in length. As this data organisation is per wave, and waves cover different time periods, the vectors are aligned at the beginning of the wave, and zero values are added to shorter outbreaks, to ensure equal vector length.
Grouping of waves is found by comparing the mapping of the individual waves on the SOM lattice.

Sequence of synoptic states

A synoptic state is a pattern that spatially characterises a diffusion state. Each wave can be represented as an ordered sequence of synoptic states. The number of states per wave is variable. This sequence provides information on the speed and on the direction of spread. The workflow for generating trajectories of synoptic states is shown in Figure 2-4b.

In order to retrieve synoptic states, a SOM is trained using the TxS SOM (Figure 2-2 – data organisation (c)). In this case each data vector represent one month (variables represent the medical districts). After training the SOM, codebook vectors are translated into a GIS map (Figure 2-4 – (b)). This can be done because the variables of each codebook vector represent a sequence of medical districts. By transposing the codebook vectors (to SxT) and visualising them in a GIS, each neuron (codebook vector) of the SOM lattice can be shown as a GIS map.

The data is now mapped back onto the SOM. For each month a mapping to a codebook vector is determined. After grouping successive months with the same mapping (within the same wave), a sequence of maps is retrieved, this composes the ordered sequence of synoptic states per wave.

States differ in duration. The speed is represented as the number of states and the duration of each synoptic state (the number months mapped to the state). The direction of spread can be derived from the maps via visual comparison of the states.

Sammon's Trajectories

Alternatively, the direction of spread can be evaluated by mapping trajectories on the Sammon's projection (Figure 2-4 – (c)). The method of combining the analysis of synoptic states and Sammon's trajectories has been previously used by Zurita-Milla et al. (Zurita-Milla, 2013). Yet, here we applied it in combination with mapping back subsets on point data.

This trajectory is constructed on a "TxS" SOM, in which each vector in the input dataset represents an epidemic month. This is the same SOM and the same mapping as used for the synoptic states.

To describe the diffusion path, the Sammon's projection is used to visualise the SOM codebook vectors in 2-D space. The Sammon's projection aims to minimize the following error function:

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \qquad (2.3)$$

in which d^*_{ij} is the Euclidean distance between the vectors *i* and *j* in the input space (the codebook vectors), and d_{ij} is the corresponding distance in the output space (the Sammon's coordinates).

Like this, each codebook vector in the SOM lattice can be projected to a 2D space. The diffusion trajectory is the vector that depicts the "sequence of movement" over the SOM lattice. Arrows connect neurons in the order in which they are mapped and the shapes of different epidemic vectors can be compared to reveal (dis)similarity between diffusion patterns of different waves.

2.3 Results

Spatial Synchrony

Identification of spatial synchrony was performed as described in methods - section "spatial synchrony". After training, the SOM lattice was partitioned into five clusters (Figure 2-5a - Lattice with clusters) identifying neuron 12 as cluster 1, neurons 8, 9 and 11 as cluster 2, neuron 10 as cluster 3, neuron 7 as cluster 4 and neurons 1-6 as cluster 5.

The identified clusters were verified using the component planes of the TxS SOM (Figure 2-6). When comparing the component planes of the clusters, it can be confirmed that districts mapped to the same cluster show good

correlation. In our experiment, including one extra class in the hierarchical clustering would lead to neuron 6 being identified as a separate class. When visualising the component planes of this class (Figure 2-6 – cluster 5), it turns out that the patterns in this group are not very prominent and the group is not very homogeneous compared to the other classes.

The mapping of both the complete dataset and of the individual waves was visualised in a GIS (Figure 2-5b and 2-5c). The results of the mapping for the complete time series, revealed that cluster 1 represents the medical district in which Reykjavik (the most dominant city in the process) is located, and a group of medical districts mapped to cluster 2, surrounding Reykjavik, are highly synchronised. On the SOM lattice this cluster is adjacent to cluster 1. In the north of Iceland three more medical districts were identified (mapped to clusters 3 and 4) that are potential regional diffusion centres. They have a relatively high incidence rate but are topologically further away from Reykjavik on the SOM lattice. On further examining it is revealed, that these correspond to the areas of Ísafjörðar and Akureyri. By examining the codebook vectors of the SOM lattice, neurons 1-6 are grouped into one large cluster (cluster 5). The codebook vectors of these neurons show that these represent medical districts with lower frequencies.

When comparing the total mapping with the mapping of the individual waves (Figure 2-5) we see that in each of these waves more local medical districts are mapped to clusters 1-4 (indicating a role in the diffusion process). This shows that there is a group of medical districts that are important in all outbreaks, but also medical districts that play a role in the diffusion process of single epidemics. For incidental mapping to cluster 2 (highly synchronised with Reykjavik) we see several additional mappings in the northern parts in almost all waves. Incidental mapping to clusters 3 and 4 may indicate (second level) diffusion synchrony with local northern and north western centres or a different diffusion pattern. Especially wave 9 has many medical districts mapped to cluster 4 throughout the island. This can be an indication of a different direction of diffusion.

Table 2-1 Quantization error

wave(s)	all	8	9	11	12	13	14	15
SxT SOM	77.11	46.89	76.72	60.26	50.90	58.80	51.20	54.38
TxS SOM	11.40	11.76	11.81	10.29	12.67	12.86	9.04	10.88
WxT SOM	6.08	4.20	10.40	3.88	3.96	6.50	4.87	8.91

The quantization error is the average distance between the input vector and the BMU. The results are shown in Table 2-1. Quantization error for the complete dataset is high (77.11). After mapping the individual waves, this value improves to 42.33-73.56. The error for the complete dataset is high

because it is difficult to match a vector over the total length of the time series. Values improve however (become smaller) when matching only parts of the time frame.

To test the robustness of mapping back of individual waves the figure of merit M(v) was calculated (Table 2-2). This figure expresses the number of medical districts that were mapped to the same cluster for the mapping of the complete dataset and also for the mapping of the individual waves. The average over the combined waves was 0.82 (scale 0-1), indicating robust clusters over the temporal period.

Table 2-2 Figure of merit

wave	8	9	11	12	13	14	15
m(v)	0.77	0.7	0.94	0.72	0.77	0.94	0.9

Spatiotemporal diffusion and trajectories

Grouping waves

The grouping of epidemic waves was performed as described in methods - section "Grouping waves". This experiment produced a SOM with a high level of topological consistency between the neurons (Figure 2-7a). Visually, interpretations like "early" (top right), "middle" (top left), "late" (bottom right), and "high" (edges) or "low" (middle) intensity can be given to the neurons. By enhancing the U-matrix, ten clusters were identified (Figure 2-7b).

The partitioning of the SOM lattice into clusters can be translated into a heat key and the data can be visualised in a GIS. Where red colours represent "early", yellow colours "middle" and green colours represent "late". Figure 2-8 shows the mapping of the individual waves using this heat key. Comparison of the GIS maps revealed that waves 11 and 14 are fast (early) developing waves (primarily red coloured), waves 12 and 15 are primarily yellow, meaning their spread is of medium speed, and wave 9 is a late developing wave (green coloured). However, interpretation of these maps is "intuitive" (it depends on the human interpretation of the colour scheme). Hence, it is easier to evaluate the results by mapping directly onto the SOM lattice.

This leads to the results shown in Figure 2-9. After mapping the complete dataset, each wave was mapped back individually. Most waves do not have mapped samples for all neurons or clusters, but the mapping is grouped to a particular area of the lattice. Similar waves should be projected to the same clusters of the lattice.

Two groups of epidemics were identified. These are the fast developing (early) epidemic waves 8, 11, 13 and 14 (Group A) that have many medical districts

mapped to the upper right hand of the lattice, and the slow developing (late) epidemics, waves 9, 12 and 15 (Group B) that show a mapping to the lower and left part of the SOM lattices. The quantization error for this experiment, when mapping back the complete dataset, is 6.08 (Table 2-1). This shows that for this experiment, the distance between the codebook vectors and the data vectors was much smaller compared to the mapping of the synchrony experiment. This was to be expected as single waves were used.

Synoptic states

This experiment was conducted on a Time in Space (TxS) SOM as explained in methods - section "Synoptic states". The trained lattice is shown in Figure 2-10a. For this type of SOM, each codebook vector represents one specific spatial pattern, so the SOM lattice can be translated into a lattice of GIS maps (see Figure 2-11).

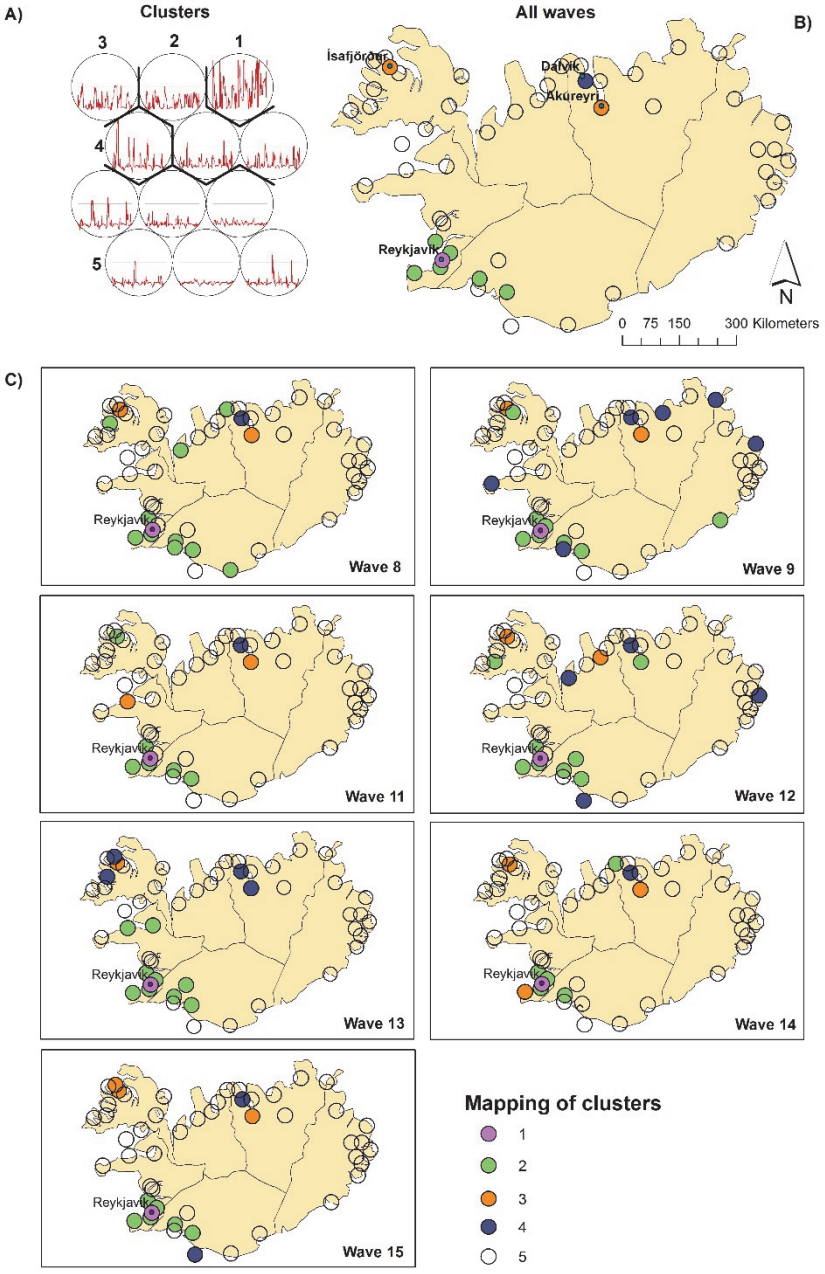

Figure 2-5 Results Synchrony. (A) SOM lattice showing hierarchical clusters (black lines). Numbers indicate the order of the neurons. (B) GIS mapping of the complete dataset (all waves), and the individual waves. (C) Numbers in the legend correspond to the numbers of the neurons in the SOM lattice.

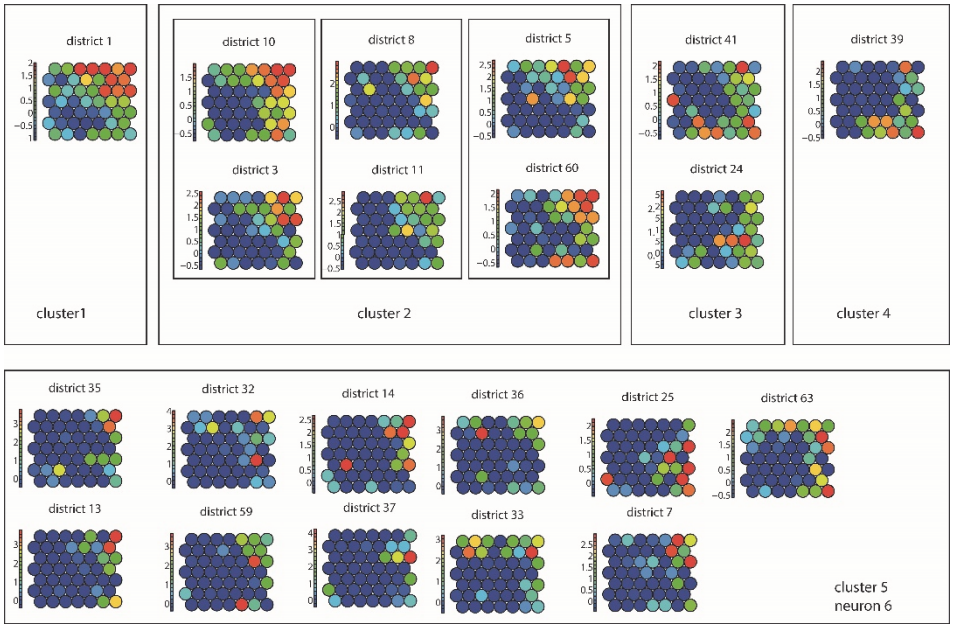

Figure 2-6 Component planes Component planes of the (TXS) SOM, organized by hierarchical clusters.

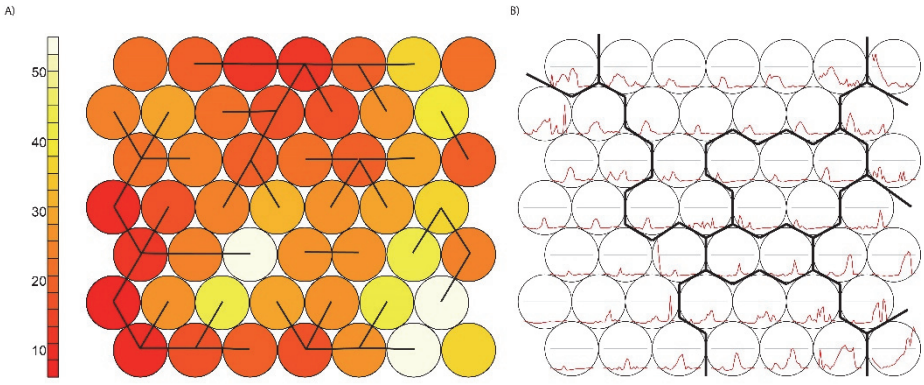

Figure 2-7 Clusters SxW SOM Enhanced U-matrix, with light background colors indicating high values, dark colors indicating low values and graph lines indicating the clusters (A). Lattice with codebook vectors and cluster lines (B).

When we conducted an interpretation of these maps, it turns out that neuron 23 represents a static state of almost no disease occurrence. Around this neuron, we identified synoptic states that represent disease occurrence in certain (combinations of) compass points, with the top right of the lattice corresponding to infection in the north and south-western parts of Iceland, and the lower left of the SOM lattice corresponding to infection in the eastern and northern parts of the island.

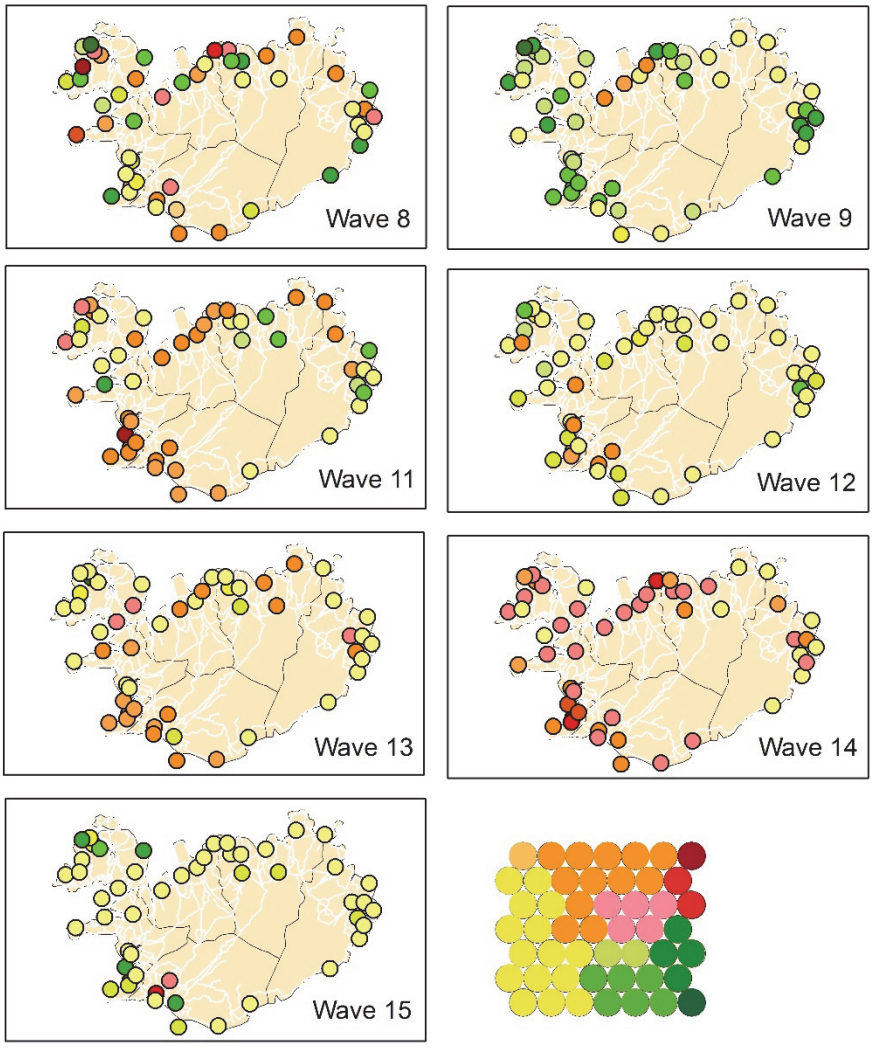

Figure 2-8 GIS mapping SxW SOM. GIS mapping using color coding for the clusters

Next, the codebook vectors of each wave were mapped as an ordered synoptic spatial time series in a GIS (Figure 2-12). This way, a sequence of maps per epidemic wave can be constructed to reflect the spatio-temporal patterns found in each epidemic wave. When comparing the patterns of Group A – consisting of waves 8, 11, 13 and 14 – we noticed that these patterns are all about equal in length, and are spatially very similar. This group seems to consist of fast developing waves. Group B – waves 9, 12 and 15 – shows more diversity in number of synoptic states and in diffusion pattern. Information about the number and duration of synoptic states can also be found in Table 2-3. Number of states for Group A ranges from 9-10, for Group B from 11-14. Duration of each state ranges from 1-6 months. Group B waves have a longer duration of the first two states.

Sammon's Trajectories

For a further analysis of the diffusion direction, the diffusion trajectories were projected as a Sammon's projection (Figure 2-13).This experiment is described in section "Sammon's Trajectories" of the methods section. The data organisation and the SOM lattice (see Figure 2-10a) were the same as for the previous experiment. The Sammon's projection of the lattice is shown in the same figure (Figure 2-10b). In this projection the distance between the vectors is explicitly mapped, but the topological relationships are not necessarily maintained. Numbers in the Sammon's projection refer to the numbers of the neurons. As can be observed, the Euclidean distance between the neurons in the top right hand of the SOM lattice (numbers 35, 42, 48, 49) is relatively large.

For each wave, the "trajectory of diffusion" was mapped onto the Sammon's projection (Figure 2-13) and the trajectories were compared. Waves 8 and 14 both have a trajectory starting in neuron 23, moving in a circular fashion to the right hand side of the figure (reaching neuron 35 as one of their peak stages) to return to neuron 23. Their diffusion pattern is strikingly alike. This diffusion corresponds to the sequence of "infection in the Reykjavik area" followed by "infection in the Reykjavik area and the north", and returning via infection for example infection in the east back to neuron 23.

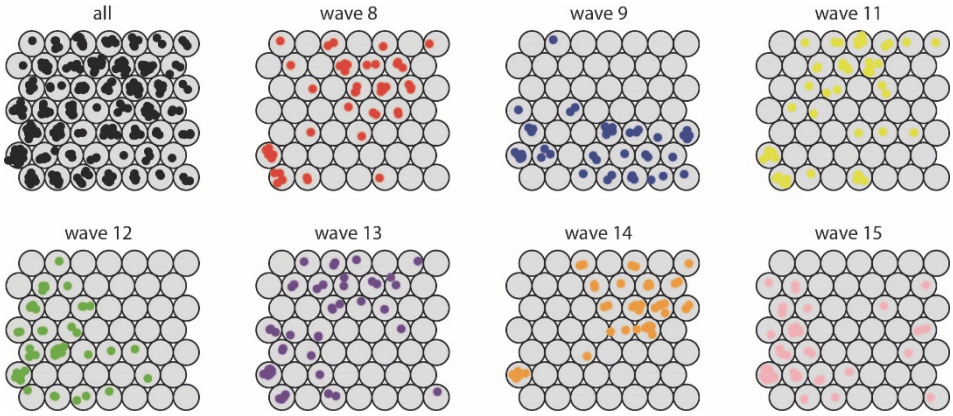

Figure 2-9 Mapping SxW SOM on SOM lattice.. Mapping of the Medical Districts on the SOM lattice. Each dot representing one medical district.

Waves 11, 12 and 13 have trajectories with similar characteristics compared to the previous group. Their trajectories also follow a circular path from neuron 23 to the upper-right hand side of the graph indicating similar diffusion. However the lines of waves 11 and 13 show cross-overs and wave 12 shows an opposite direction. Cross-overs occur when a fast spread (or decline) to all areas takes place. Where waves 8 and 11 decline to the north, the opposite direction of wave 12 is triggered by a decline to the north and south of the island.

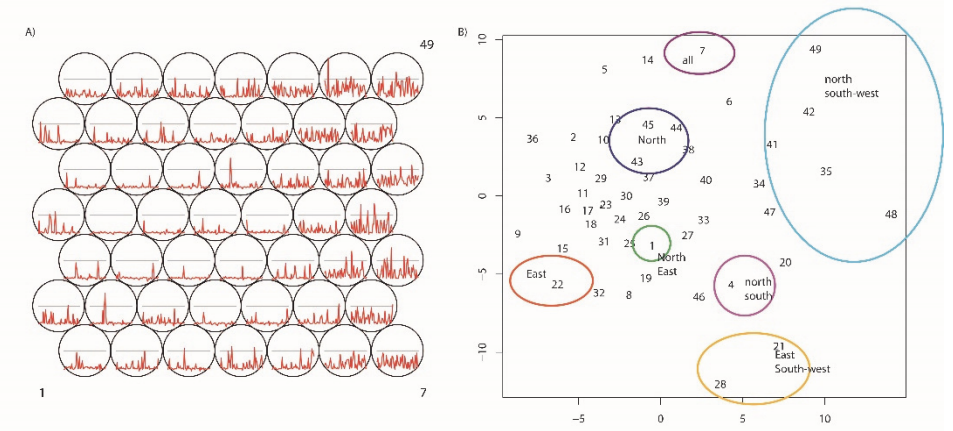

Figure 2-10 Codebook vectors and Sammon's Projection TXS SOM. Lattice showing codebook vectors, with neuron 1 in the lower left corner, numbering in rows, from left to right and bottom to top, ending with neuron 49 in the top right hand corner (A) and Sammon's Projection with interpretation (B).

However, these diffusion trajectories do not differ significantly from the trajectories of wave 8 and 14.

Wave 9 and 15 have the strongest deviation from the general pattern. For the interpretation of these results we mapped the compass directions onto the Sammon's projection (Figure 2-10b). Wave 9 and 15 show strong vertical trajectories. These correspond to a different spatiotemporal diffusion pattern. For Wave 9 this is a pattern from neuron 45 to 28 (spread starting in the north) and for Wave 15 from (1, 14, 27) from the east, via the north to the western parts of the island.

2.4 Discussion

The use of SOMs, combined with Sammon's projection, has enabled us to identify synchrony between locations (medical districts), and to map diffusion trajectories of a time series of epidemic waves revealing their spatiotemporal diffusion patterns. This integrated approach was carried out using three different data organisations (in space and time). Training the SOM on the complete time series and mapping back individual waves has shown to be a simple but effective way to compare general spatiotemporal diffusion patterns for a complete time series with patterns of individual waves. Results found are consistent with results found for the same epidemics, using different methods.

The synchrony experiment revealed a number of medical districts that form the diffusion structure for all of the waves and, additional medical districts that only play a role in some waves. The medical districts that were identified as forming the stable structure all have a large number of inhabitants and the centres in the north and north west are connected to Reykjavik via domestic air travel. The identified medical districts show a great similarity to the structure found by Cliff et al. (Cliff et al., 2009b), in their quarterly lag maps of geographical spread, first and second quarter. They are also consistent with epidemiological theory (hierarchical diffusion models).

The research identified two different groups of epidemics with fast developing (group A) and slower developing waves (Group B). These groups were identified using the SxW SOM but the synoptic state experiment, based on the TxS SOM revealed the same grouping. Similar results were found by Cliff et al. (Cliff et al., 1981b) who found for Group A mean lag time in months of respectively 8.42, 9.05, 10.54 and 5.85 months and for Group B a mean lag time of 15.45, 11.29 and 13.32 months.

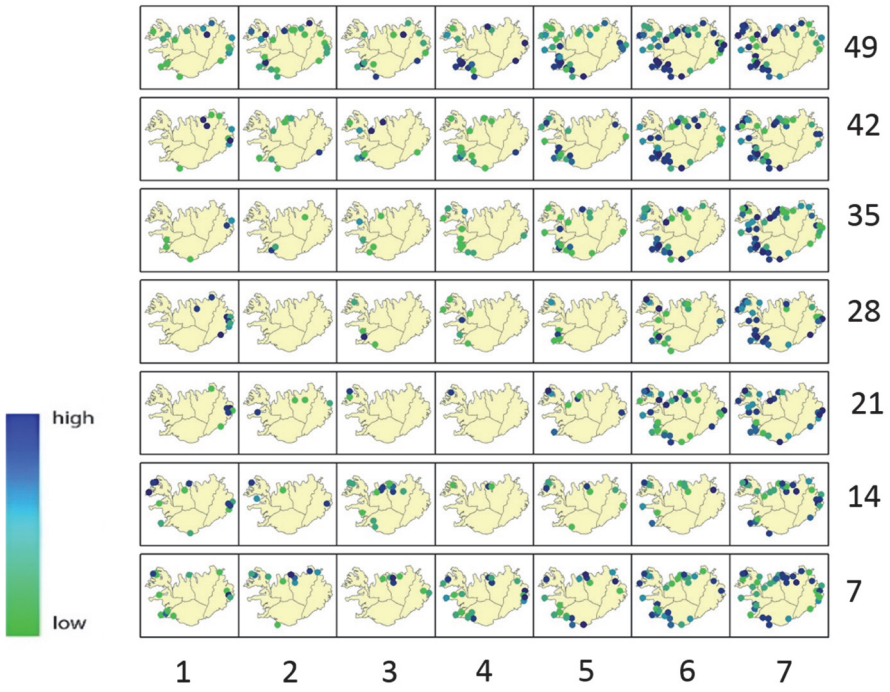

Figure 2-11 Lattice converted to GIS maps.. Lattice showing the codebook vectors as maps of synoptic states. Numbering from left bottom corner (1) to the top right (29), in rows from left to right.

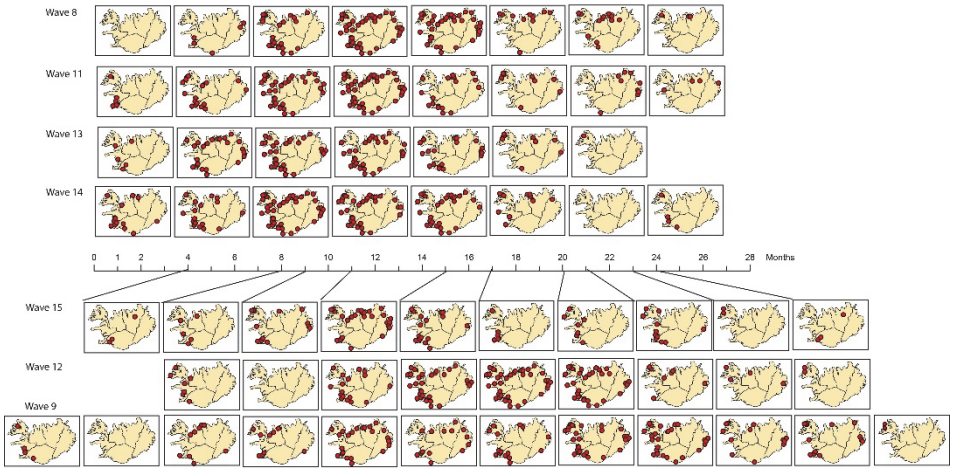

Figure 2-12 Trajectories of synoptic states. Representation of the waves as a sequence of synoptic states.

Experiments also showed that the fast developing waves (Group A) show considerable similarity in their spatiotemporal patterns. They all spread from Reykjavik to the north-eastern areas. Group B waves cannot be characterized by a single direction of spread; yet, the diffusion trajectories show that the spread of wave 9 and 15 does not follow the Reykjavik pattern. Findings were compared to Cliff et al. (Cliff et al., 1981b) that describe the spread of wave 9 as being confined to the northern parts of the island. For wave 15 the same source notes that it was slow moving, that it started in the south, but the difference in diffusion pattern, was not reported.

Table 2-3 Synoptic states. Number and duration (in months) of synoptic states per wave.

			number of months per state													
wave	**group**	**# states**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
8	A	9	1	2	1	2	1	3	3	3	6					
9	B	14	4	1	1	3	2	1	2	3	1	2	3	1	2	1
11	A	10	1	3	2	2	2	2	2	2	4	3				
12	B	11	4	2	2	1	1	2	3	1	2	2	1			
13	A	9	2	2	3	1	4	2	3	8	1					
14	A	9	2	2	3	1	4	2	3	8	1					
15	B	12	3	4	1	2	4	2	3	1	2	1	1	3		

Two methods were used for the clustering of the trained SOMs: the enhanced U-matrix and hierarchical clustering combined with a validation based on component planes of a temporal SOM. Although both methods lead to reliable results, the enhanced U-matrix is advantageous because it does not require the user to determine the number of clusters. However, other enhancement methods for the U-matrix exist, for example methods including the second best matching neuron (Tasdemir and Merenyi, 2012). These may be worth further exploring.

The proposed method was used on a time series of seven epidemic waves and a relatively small number of spatial locations. Spatially this method is scalable without any problems. However, there are probably a minimum number of waves needed to come to a reliable mapping of the individual outbreaks. If the diversity in the training dataset is too small this may lead to problems. The synchrony experiment was tested with shorter time series including fewer epidemics. A series of 4 outbreaks still gave a reliable result for our case study, but this may depend on the complexity of the dataset.

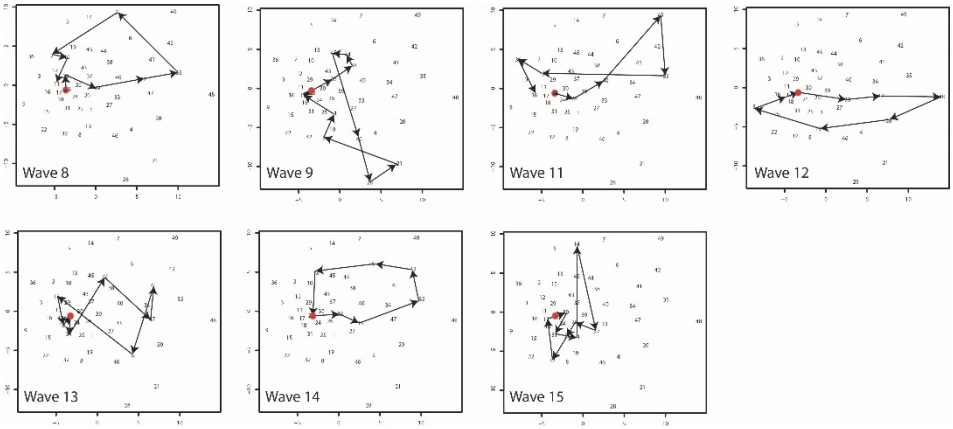

Figure 2-13 Trajectories on Sammon's Projection. Projection. Trajectories of spatiotemporal diffusion on the Sammon's Projection

Iceland is a small island with only one larger city and it is clear that this city is the "motor" of the diffusion mechanism. In this regard, the selected dataset is ideal for testing new methods (also because there are seven epidemic waves available). However, it would be interesting to test the proposed method in a much more heterogeneous environment (with more large cities and a more complex diffusion pattern). For this study all medical districts were included, but some of these centres represent areas with low population. When working with larger datasets, removal of sparsely populated areas may be an option.

SOMs are relatively easy to train and combine with other visualisation methods to enhance their analytical possibilities. However, results are very sensitive to values of input parameters (size and shape of the training lattice, number of iterations, type of initialization), and deriving meaningful information can be challenging (Wendel and Buttenfield, 2010). The method as applied here, therefore focusses more on comparison then on absolute characterisation of patterns.

Besides Measles, this method is potentially useful to explore and understand spatiotemporal diffusion patterns of other infectious diseases (e.g. Influenza, Pertussis) as SOMs can deal with large datasets as well as with missing data. This understanding might lead to new paradigms of modelling and validating spatially explicit disease models based on reproducing observed diffusion patterns.

This method can also be used for real time disease mapping. As data from partial outbreaks can be mapped back, comparison of diffusion patterns with previous waves may lead to early indications of the characteristics of an epidemic and, thus, help to design intervention actions. When linked to

systems for Volunteered Geographic Information, web-based monitoring networks a fully automated analysis may be an interesting option.

2.5 Conclusions

In this paper we proposed a SOM-based method to analyse spatiotemporal diffusion of infectious diseases. The method is based on training a SOM for a larger time-series (including multiple waves) and mapping back individual outbreaks for characterisation and comparison. Via a number of experiments we showed how this method can be applied for finding synchronies between spatial locations and for comparing spatiotemporal diffusion patterns of different epidemics.

We also demonstrated how different types of data organisation (in space and time) can help to reveal different information. Several types of secondary clustering (hierarchical, enhanced U-matrix and Component planes) were shown, that can be used to improve the SOMs performance. The integration of SOMs with other visualisation techniques, especially Sammon's Projection and GIS was used to detect, interpret and visualise spatial temporal patterns.

Results of the method are consistent with diffusion patterns found using other methods; this makes SOMs an interesting alternative, worth further exploring. For instance, by applying it to a larger dataset in a more dynamic geographic environment, by coupling it to a spatially-explicit disease model or by using it for near-real time disease monitoring.

Chapter 3 Using time series clustering to delineate pertussis reservoirs in the Netherlands

3.1 Introduction

Many emerging diseases can be described as systems where pathogens are permanently present in an infection reservoir (source) in which the disease has become endemic and from which the pathogen can spread to other areas that are directly or indirectly connected. An example is the re-emergence of pertussis in the Netherlands. Despite the high levels of childhood vaccination, the disease has become endemic at a country level with periodic epidemics every 3 to 4 years.

Although literature suggests that public health interventions can be based on knowledge about the location of the infection reservoirs (Haydon et al., 2002), little research has been conducted on the spatial identification or delineation of these areas for infectious diseases. The delineation of reservoirs is important to detect environmental determinants (Eisenberg et al., 2007) because the (re)-emergence of several infectious diseases has been linked to environmental changes like urbanization, ecological changes like (de)forestation, or societal changes like commuting patterns and other changes in human mobility (Morse, 1995, Patz et al., 2004).

The most common method to identify host areas for infectious diseases is to use the Critical Community Size (CCS). CCS is usually defined as the threshold population size above which a disease will persist once a population has been infected – or contrarily – below which a disease cannot persist in the long term (Bartlett, 1957). It is a useful yet somewhat classical measure that assumes that a "community" can easily be identified and is not globally linked to many other communities.

The value of CCS for different infectious diseases can be determined via empirical observations, analytical expressions and computer simulations (Viana et al., 2014). When CCS is determined based on empirical observations, (in most cases) no attention is paid to the spatial location of the reservoir but the only purpose it to determine the population size above which the infection should become endemic. The spatial scale to determine the CCS varies, e.g. Choisy and Rohani (2012) used data aggregated per American state, Bartlett (1960) applied the CCS to American cities with population above 200000 Keeling and Grenfell (1997) used towns in England and Wales and Metcalf et al. (2013) applied it at a country level. When CCS is calculated for large and heterogeneous areas like a state or a country, the delineated area may contain

both host and target populations. When a city is used to determine the CCS, this city may be functionally connected to surrounding suburbs or smaller towns that act as one reservoir of infection together with the central city.

Disease data is mostly registered per administrative unit and analysis using these areas may seem obvious though the shape and boundaries of these areas are arbitrary. When delineating the reservoirs of infection we need to be careful with the granularity of the input areas to retrieve a reservoir that is as geographically representative as possible. The input areas should be large enough to have a stable signal, yet small enough so that they will not contain parts that are less populated and do not contribute to maintaining the infection. A reservoir of infection is not a single community but a combination of multiple communities. It is not necessary that each community on its own can maintain the infection (is above the CCS) but the maintenance community as whole can. When cities are relatively close together they may together form one reservoir of infection.

When the host areas contains multiple communities, proximity and connectivity between them become important. Diseases spread via direct contact between infected and susceptible individuals and the spatial distribution of these contacts will influence the spatio-temporal diffusion patterns of the diseases. While local outbreaks are triggered by contacts in the direct neighborhood (e.g. within household) regional and country level propagation of diseases is linked to multi-scale human mobility. Regional spread of influenza in the United States was linked by Viboud at al. (2006) to work related daily commuting. According to Riley (2007) disease diffusion of infectious diseases is highly correlated to human movement at different spatial and temporal scales and lack of availability of detailed data on human mobility has resulted in a range of spatial-transmission models proving the importance of movement patterns of hosts.

Patterns of spatio-temporal disease diffusion can provide vital information on the differentiation between source population and target population. Infection of target populations will start with a spillover from the host area leading to an initial case (index case). This primary case can lead to a so called "stuttering transmission chain", caused by secondary infection in the target population, but as there is limited transmission in the target population the size of outbreaks will remain limited (Blumberg and Lloyd-Smith, 2013).

Time series analysis has been widely used in environmental epidemiology for example in regression studies (Bhaskaran et al., 2013). Clustering of time series has recently gained more attention within the health domain and has been applied by Tignor et al. (2017) for clustering of mobile health app data, by Ghassempour et al. (2014) for clustering multivariate time series of health

trajectories and by Augustijn et al. (2013, Augustijn et al., 2015) for clustering measles and pertussis data.

In this paper we identify areas that together interact as a coherent unit that maintains the disease and acts as one population in the diffusion process which we will call a *Critical Community Region* (CCR). We will delineate the CCR using time series clustering and examine the suitability of different distance measures and clustering techniques. We will illustrate the proposed method based on a synthetic dataset and a case study of pertussis in the Netherlands.

3.2 Methods

Case study and data

The Netherlands is a highly urbanized country in which boundaries of cities touch each other and complete regions can be characterized as urban. The country is small and has a high commuting level for work, school and social activities.

Pertussis childhood vaccination was introduced in 1952 with a very high average participation in previous decades (van der Maas et al., 2013). Despite increased awareness and changes in the vaccination program the number of pertussis cases in adolescents and adults has increased over the last 20 years leading to several epidemics.

For this study, weekly notifications of pertussis from the National Institute for Public Health and the Environment (RIVM) of the Netherlands at municipality level for the period 1996-2013 were used. Time series subsequences were extracted for epidemic periods leading to 6 subsequences per administrative unit (Appendix 2). This allows checking the similarity between subsequences for the same area and for different epidemics to determine the robustness of the CCR. All data were smoothed and z-normalized per zone and epidemic.

The process for delineating the CCR consists of the following steps (see Figure 3-1):

1. In the first step we determine the most suitable distance measure and clustering algorithm for the analysis of time series of disease data by using a synthetic dataset with known partitioning (blue elements in Figure 3-1);
2. In the second step, we identify zones that have a high probability of belonging to the CCR (green elements in Figure 3-1) using the percolation method ;
3. In the third step we apply multi-outbreak data mining to empirical time series of pertussis data for the percolation zones (orange elements in Figure 3-1).

The next three sections will provide detailed information about these steps.

Selection of distance and clustering method

Depending on the objective of the research, time series clustering can focus on different characteristics of the times series like phase (timing of the infection) or amplitude (areas with more or less disease cases). In this research we want to group together areas for with disease time series show a similar shape. By grouping similarly shaped curves, we identify areas where the shape of the time series indicates a stuttering chain of infection and fade-outs, versus another group where the infection maintains itself.

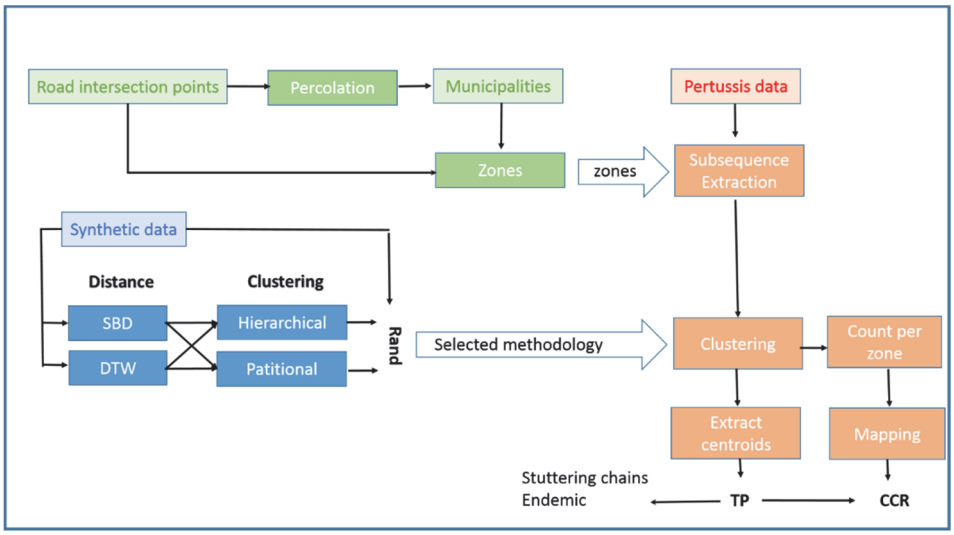

Figure 3-1 Overview of the methods used in this study ,blue elements (lower left side of the figure) represent the evaluation of the clustering methodology based on a synthetic dataset (2-2), the green elements (top left) consist of the delineation of the percolation zones (2-3), and the orange elements (right hand side of the figure) show the CCR region extraction part of the methodology applied to the case study of pertussis in the Netherlands (2-4)

In order to group time series, the similarity in shape between these time series is calculated by measuring the distance between its nodes (x_n and y_n). When working with subsequences of time series for two regions $\vec{x} = (x_1, \ldots\ldots, x_m)$ and $\vec{y} = (y, \ldots\ldots, y_m)$ or epidemics $\vec{x}_{t=1} = (x_1, x_2, \ldots\ldots x_m)$ and $\vec{x}_{t=2} = (x_1, x_2, \ldots\ldots x_m)$, these time series might show a shift invariance. This happens, for instance, when the infection peak occurs at different time steps. Such a shift needs to be accounted for before calculating the similarity of the two subsequences otherwise the actual distance will be over-estimated. Various time series alignment techniques exist. For instance, time series can be aligned by sliding one of them (global alignment - moving) or by aligning nodes (local alignment

- warping). Here we test time series clustering methods that apply both the moving (sliding technique) and the warping of time series.

Table 3-1 An overview of the synthetic dataset. The R_0 is the basic reproduction number indicating the number of secondary cases an initial case would produce in a fully susceptible population. The TP is the two peak metric of equation 3.6.

Group	Numbers	Number of time series	Mean R_0	Standard Dev R_0	TP	Description
S1	1-9	9	1.64	0.034	4.26	An initial time series was created using an index case, after incubation time a peak followed by a second peak (S 1)
S2	10-18	9	1.31	0.21	4.35	In order to generate a more endemic situation a basic number of disease cases was added to S1
S3	19-21	3	2.42	0.96	4.03	based on type S1 yet with re-occurring infections leading to multiple irregularly spaced peaks
S4	22-33	12	4.47	0.43	1.72	A double peak set was generated with first a major peak followed by a minor peak
S5	34-42	9	1.33	0.18	1.88	endemic cases were added to S4
S6	43-57	15	1.96	0.08	1.83	The opposite of the shape of S4 (first a small peak followed by a higher peak)
S7	58-72	15	1.65	0.63	1.83	Opposite shape of S5
S8	73-76	4	1.69	1.87	0.46	Partial curves with only the tail of the peak or the beginning of the peak

The choice of clustering technique is domain dependent and needs to be determined for every new application (Shekhar et al., 2015). Hence we use a synthetic dataset to decide on the best distance measure and clustering

algorithm. This synthetic dataset includes time series with the following characteristics: stuttering chains (primary infection followed by fade-out followed by a secondary infection), endemic versus epidemic time series, subsequences with displacement along the horizontal axis, differences in shape of the peak and partial peaks. An overview is provided in Table 3-1. This resulted in a dataset of 76 time series. All data were row-wise (per time series) z-normalized before the clustering (for more information about the synthetic patterns see Appendix 1).

Different alignment methods handle different types of variation like differences in amplitude and offset in different ways. Here we compare two shape-based distance measures: Dynamic Time Warping (DTW) and Shape-Based distance (SBD). DTW was chosen because of its frequent use for time series clustering and SBD was selected because it uses a normalized version of the cross-correlation measure which is more often used for epidemic time series (Chen, 2015). Both methods are performed using the R package DTWclust (Sarda-Espinosa, 2017).

When comparing two time-series $\vec{x} = (x_1, \ldots\ldots, x_m)$ and $\vec{y} = (y, \ldots\ldots, y_m)$ DTW computes a local distance matrix (M) between all the nodes of the two time series. DTW is defined as (Paparrizos and Gravano, 2015, Keogh and Ratanamahatana, 2005):

$$DTW(\vec{x}, \vec{y}) = min \sqrt{\sum_{i=1}^{k} w_i} \tag{3.1}$$

In which k is a set of matrix elements and W is the warping path, the minimum path over the local distance matrix $W = \{w_1, w_2, \ldots\ldots w_k\}$ between $\vec{x}$ and $\vec{y}$. In this way a node can not only be aligned with the corresponding node in the other time series, but also via a local none linear alignment with another node in the second time series. DTW can be used with different distance measures but here we use the most frequently used Euclidean distance measure.

Shape-based distance (SBD), is a cross-correlation statistical measure which determines the similarity between $\vec{x}$ and $\vec{y}$ by keeping $\vec{y}$ static and sliding $\vec{x}$ over $\vec{y}$ to determine the optimal alignment between the two time series. The goal of this method is to determine the optimal position at which the Cross-correlation ($CC_w(\vec{x}, \vec{y})$) is maximized.

The distance measure of SBD can be defined as (Paparrizos and Gravano, 2015):

$$SBD(\vec{x},\vec{y}) = 1 - max_w\left(\frac{CC_w(\vec{x},\vec{y})}{\sqrt{T_0(\vec{x},\vec{x}).T_0(\vec{y},\vec{y})}}\right) \quad (3.2)$$

In which $\sqrt{T_0(\vec{x},\vec{x}).T_0(\vec{y},\vec{y})}$ is the geometric mean of autocorrelations of the individual sequences.

Two types of clustering are tested: hierarchical clustering and partitional clustering that corresponds to the K-shape clustering discussed by Paparrizos and Gravano (2015).

Hierarchical clustering is frequently used to group time series because it can easily be visualized using dendrograms and because a prior identification of the number of clusters is not necessary. Hierarchical clustering first calculates a distance matrix for all the subsequences. Then, it identifies the two subsequences with the smallest distance (most similar) to form the first cluster. After that, it updates the distance matrix by calculating the distance between the newly formed cluster and the remaining subsequences. This process is repeated until all subsequences are assigned to a cluster.

K-Shape clustering is a partitional clustering method based on an iterative refinement procedure in which the algorithm minimizes the sum of the squared distances between the subsequences and the centroid of the cluster to which they are assigned. In every iteration the algorithm performs two steps: assignment step in which cluster members are updated and the refinement step in which the cluster centroids are updated.

We evaluate the match between the clustering outcomes and the "real" clusters using the Rand Index (RI) (Rand, 1971). This index performs a pairwise comparison by taking two elements from one clustering e.g. [2,3] and the same pair from the other clustering. It compares all possible pairs in the following way:

$$RI = \frac{a+b}{a+b+c+d} \quad (3.3)$$

In which *a* is the number of pairs that were assigned to the same cluster, *b* is the number of pairs that were in different clusters, *c* is the number of pairs that were in the same cluster in the first partitioning but in different clusters in the second partitioning and *d* is the number of pairs that were in different clusters in the first but different clusters in the second partitioning. The RI is a value between 0 and 1 with 1 representing a perfect match between the two partitionings.

Zone identification

Before clustering the disease time series, we would to identify the spatial units to use as input for the clustering. We are looking for areas large enough to maintain the disease, yet, want to restrict the areas to highly connected and urbanized areas. Road data is used to extract the input areas, as this type of data captures both connectedness and urbanization in one dataset. Our target areas are those areas with a high road density and short road segments.

The percolation method (Arcaute et al., 2015) is used to aggregate the available road data into larger regions that have tight commuting connections. Road networks have multiple hierarchical levels, which can be identified by phase transitions. We can imagine that regions with tight commuting links have short connecting road edges leading to many road intersection points. Contrarily, long and sparse edges represent areas without frequent commuting. When we remove the longest edges our network will fall apart into multiple clusters of more tightly connected areas. By using different distance threshold values for dropping line segments, starting with a large value and reducing this value gradually we will observe the transitions in the network.

We try to identify the threshold values at which urban agglomerations are split off and will use these clusters as our urban zones. Transition points can be identified by plotting the distance threshold against the average cluster size. A giant component is a connected component of a network from which all other points can be reached. . When there is a giant component, this means that the network has not fallen apart into separate clusters yet, but there is a large cluster in which almost all places are connected. When using this method on the network of the Netherlands, initially small islands split off, yet, the mainland remains one big cluster. This cluster is the giant component. When gradually the threshold distance becomes smaller the giant component will fall apart into individual clusters. In this study we perform percolation for a range of distance threshold value on street intersection points by applying the following steps:

1. Extract road intersection points.
2. Calculate a Euclidean distance raster based on these intersection points
3. Drop the line segments that are longer than the threshold distance, to let the network falls apart in separate (tightly connected) regions.
4. Remove clusters below the minimum cluster size (small fragments) and for larger distances the giant component.
5. Average cluster size per distance threshold was calculated and plotted to identify the appropriate distance value for the Netherlands.

When evaluating the average cluster size plot, it shows strong dips at certain distances. At these distances the network shows a transition point. Transition

points can be triggered by natural boundaries (islands, large natural areas), or by structural organization of urbanized areas. In this research we try to identify the latter.

After selection of the threshold distance most suitable for the Netherlands, clusters were extracted and compared to the municipality boundaries. Zones were constructed of municipalities that were completely covered by a percolation cluster or covered for more than 50% of their area. This last step is necessary to aggregate the disease data. The extracted regions will now be referred to as "zones". Disease incidence data is aggregated to disease time series per zone on which further clustering will be applied.

Delimiting CCR for pertussis in the Netherlands

The best distance and clustering method (according to the synthetic dataset) will be used to identify the CCR for pertussis in the Netherlands based on time series subsequences for the zones extracted via the percolation method. Because the optimal number of clusters is unknown, we evaluated the stability of the results for various number of clusters. In particular, we evaluated whether the cluster centroids represent stuttering chains or an endemic situation (i.e. reservoir).

After clustering the subsequences of the pertussis data for the identified zones, we extract the centroid of the clusters to evaluate if they represent stuttering chains or endemic situations. The extraction of the cluster centroids allows for calculation of indices to identify the cluster characteristics. Cluster centroids in this research are computed according to Paparrizos and Gravano (2015) where the centroid is calculated using an optimization problem in which the sum of squared distances to all other time series (μ_k)is optimized

$$\overrightarrow{\mu_k} = \underset{\overrightarrow{\mu_k}}{argmax} \sum_{\overrightarrow{x_i} \in P_k} NCC_c(\overrightarrow{x_i}, \overrightarrow{\mu_k})^2 \tag{3.4}$$

$$= \underset{\overrightarrow{\mu_k}}{argmax} \sum_{\overrightarrow{x_i} \in P_k} \left(\frac{max_w CC_w(\overrightarrow{x_i}, \overrightarrow{\mu_k})}{\sqrt{T_0(\overrightarrow{x_i}, \overrightarrow{x_i}).\, T_0(\overrightarrow{\mu_k}, \overrightarrow{\mu_k})}}\right)^2 \tag{3.5}$$

Centroids of the clusters are inspected in three ways: visually, by means of the basic reproduction number R_0, that represents the average number of secondary cases generated by an original infection and by calculating the two peak index (TP). The TP indicates the depth of the deepest valley in epidemic time series and was developed for the detection of multiple waves. The TP as introduced by Hoen et al. (2015) is calculated as:

$$TP(x) = max_{1\leq i<j\leq n}(\sqrt{(x_i - v_{ij}) \times (x_j - v_{ij})}) \quad (3.6)$$

Where x is the time series (x1,x2,…..xn) and vi,j=mini≤k≤j(xk) is the minimum value occurring in the time series between xi and xj. The distribution of the output values generates a value close to zero for single peak time series, values between 1.5 – 2.0 represent a single multi peak wave, and values above 2 represent the stuttering chain effect we are looking for (infection, fade-out, infection). Subsequences that exhibit the stuttering effect are excluded from the CCR. The endemic cluster is the cluster with a single peak and non-disappearing infection.

To spatially delineate the CCR area, we use the robustness of the pattern. For all zones we will count the number of times the zone was assigned to a particular cluster. and only zones that are assigned to the selected cluster for the majority of the subsequences (epidemics) belong to the CCR.

3.3 Results

Optimal distance and clustering methods

The results for the clustering of the synthetic data (Table 3-1) to identify the optimal clustering method are shown in Table 3-2. Highest scoring combination (based on Rand index) is the distance measure SBD in combination with hierarchical clustering (RI of 0.85).

Table 3-2 Comparison of different methods

	Hierarchical	**Partitional**
	RI	**RI**
SBD	0.85	0.80
DTW	0.79	0.82

The cluster members for this combination are shown in Figure 3-2 and the partitioning is shown in Figure 3-3.

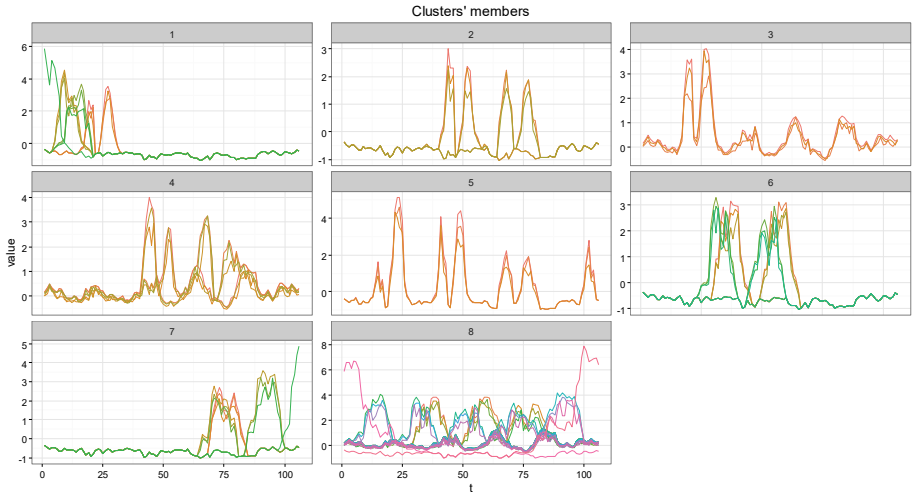

Figure 3-2Cluster' members of the synthetic dataset F and the SBD combined with hierarchical clustering. The numbers refer to the cluster numbers.

The group of synthetic time series S1 (identical but with a shift along the time axis) was split into two different clusters: 1 (time series 7,1,4) and 2 (time series 2,3,5,6,8,9). In cluster one we find the time series with an early peak and in cluster 2 the time series with a peak in the middle of the time axis. Apparently when we slide the peak along the time axis, this can lead to a split into multiple clusters despite the fact that the algorithm should perform a global alignment. Alignment is conducted for smaller displacements but that larger moves can lead to multiple cluster.

The next group (S2) is identical to S1, but with an endemic component added. These time series are assigned to clusters 3 and 4, again with the same split into early and later peaks we observed in series S1.

S3 is the set of multiple peak example which are the stuttering chains we are interested in. These make up a cluster on their own (cluster 5). The distance method can clearly distinguish based on the number of peaks.

S4 and S6 are situations with a double peak that differ in shape. S4 starts with a large peak followed by a smaller peak directly afterwards (without a fade-out of infection) and S6 has the opposite shape (first small followed by a larger peak). S4 and S6 are split over the clusters 1, 6 and 7 corresponding to the time at which the peaks occur (early for 1, middle for 6 and late for 7). Although this is a shape based method, it cannot detect such minor shape differences.

S5 and S7 are the endemic versions of S4 and S6 also with opposite peaks. All time series in S5 and S7 were assigned to cluster 8.

Group S8 consists of partial peaks (incomplete, only start or the end of the peak). Time series in S8 were mixed with the other groups. Two partial peaks are in cluster 8 and the others are in cluster 1 and 7.

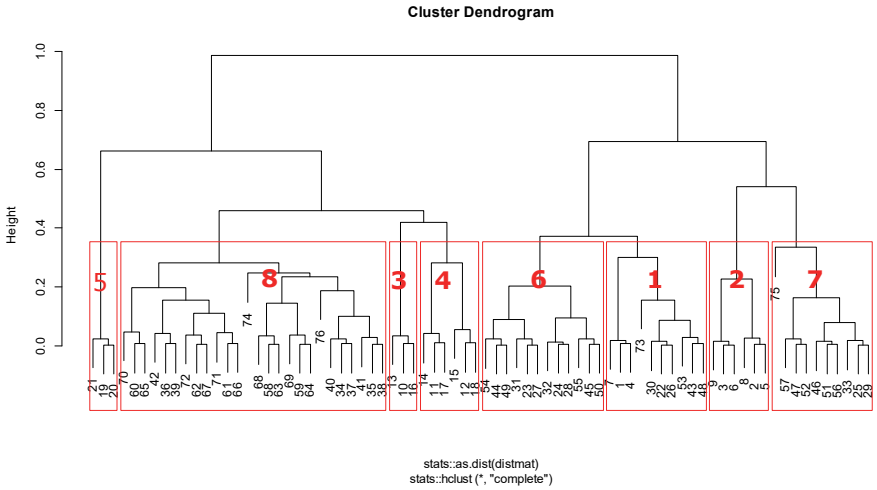

Figure 3-3 Dendrogram showing the combination of SBD with hierarchical clustering. Red boxes indicate the clusters with cluster numbers indicated in red. The black numbers refer to the time series numbers of the synthetic dataset.

It is clear that the level of detail of first small followed by a larger peak or first large followed by a smaller peak cannot be picked up by the algorithm. However, the algorithm does not mix series with and without endemic components. The algorithm separates out multiple peaks for multi-peak and single peak examples and puts partial peaks in the correct cluster. For our purpose this is enough to continue.

Zone identification

The step wise results of the percolation method are shown if Figures 3-4 and 3-5. At 5000 meters the first islands in the north of the country are already splitting off. At 800 meters one of the islands in Zeeland (south west) is split off but others are not yet. At a distance of 500 meters, islands and polders are splitting (all natural boundaries). At 220 meters, the regional separation is starting to appear. This is the split we are looking for. At 170 meters, we reach the level of the city. Only the Randstad region remains un-split. When we compare our findings with the distances in (Arcaute et al., 2015)we note that our distance of 170 meters is in good correspondence with the 160 meter split for cities as listed in Arcaute et al.(2015).

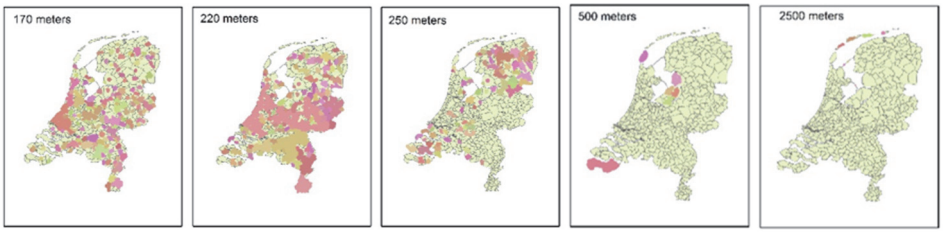

Figure 3-4The transitions of the cluster identification for different distance thresholds

Figure 3-5 shows the identified clusters for commuter traffic when applying a threshold of 5000 intersections. A total of 16 urban zones were found and these zones were confirmed using the commuting data obtained from CBS for 2013. Zones show a good match with the commuter flows and seem to form the heart of these flows.

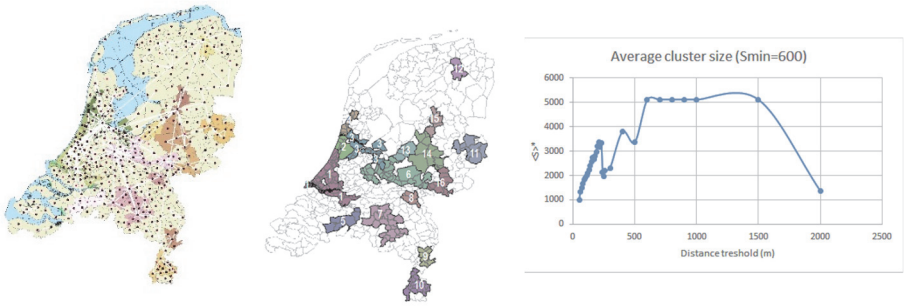

Figure 3-5Threshold identification. Left the identified zones with commuting flows larger than 1000 commuters as spider lines (white) Middle the zones showing the zone numbers, Right the average cluster size plotted against the percolation distance threshold. Dips in the figure show the transition to a next hierarchical level.

Table 3-3 gives an overview of the number of commuters per zone. C_{out} is the number of commuters living outside the zone and commuting to this zone. M is the number of municipalities in the zone. Total number of commuters including people both living in and outside the zone is expressed as $C_{all.}$ Population per zones varies between 71400 for the smallest zone to 1008800 for the largest zone.

Delineation of CCR for pertussis in the Netherlands

After extracting the number of pertussis cases per zone for the different subsequences (6 epidemics) we determine the number of clusters we will use by comparing every partitioning against all other numbers of clusters and calculating the mean Rand index per partitioning. The result is shown in Figure 3-6. Both 5 and 6 clusters have a RI of 0.798. Based on these results we will use 6 clusters for the extraction of the CCR.

Table 3-3. Zone details, in which M is the number of municipalities, C_{out} are the number of commuters to the zone, C_{in} are the number of people living in the zone that commute to another municipality inside the zone, C_{all} are the total number of commuters.

Zone	Name	M	C_{out}	C_{in}	C_{all}	population
1	Randstad Zuid	54	283300	1248000	1531300	1008800
2	Regio Haarlem	9	328100	408400	736500	333500
3	Amsterdam	33	157900	326100	484000	728000
4	Regio Alkmaar	11	24300	108900	133200	227600
5	Brabandstad West	17	58800	243300	302100	171700
6	Midden Nederland	39	232700	576800	809500	552500
7	Brabandstad Oost	41	115800	610600	726400	400300
8	Nijmegen	12	35000	115300	150300	143400
9	Noord Limburg	8	24400	104300	128700	71400
10	Zuid Limburg	17	120000	101500	221500	139900
11	Twente	14	22600	227300	249900	130400
12	Assen-Groningen	13	52300	165300	217600	101800
13	Regio Ermelo	6	20500	45500	66000	150400
14	Zutphen-Deventer	11	77100	108300	185400	260700
15	Zwolle	9	18800	63100	81900	112900
16	Regio Doetinchem	6	19000	66900	85900	110600

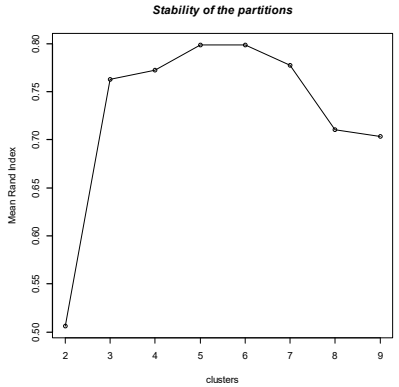

Figure 3-6 Stability plot showing the Rand index for different numbers of clusters

The final results are shown in Figure 3-7 and Table 3-4. Figure 3-7 shows the centroids of the 6 extracted clusters. We immediately notice the multi-peak stuttering chain patterns in clusters 2 and 4. Visually it is also clear that clusters 1 and 3 seem to represent single peak examples. Notice that both clusters 1 and 6 have a shorter time series. When some of the subsequences that are

mapped to a particular cluster are shorter, the centroid will be cut to the shortest length. Table 3-4 shows the TP values of the extracted centroids. Values for centroids 1 and 3 are again low, indicating a single peak (no re-infection). However, values for clusters 2, 5 and 6 are high, indicating a multi-peak.

When we evaluate the mapped clusters we see that all of our zones are represented in cluster 1. It is by far the most identified pattern in our time series followed by cluster 3. Further inspection of cluster 1 reveals that there are some zones that are mapped to this cluster less frequently. Examples are zones 5, 8, 11 and 13. These zones are more present in cluster 3. This is the pattern with only one peak yet, where the infection disappears (see tail of the centroid). The CCR can be identified as the area of cluster 1 minus the zones that are more frequent in cluster 3.

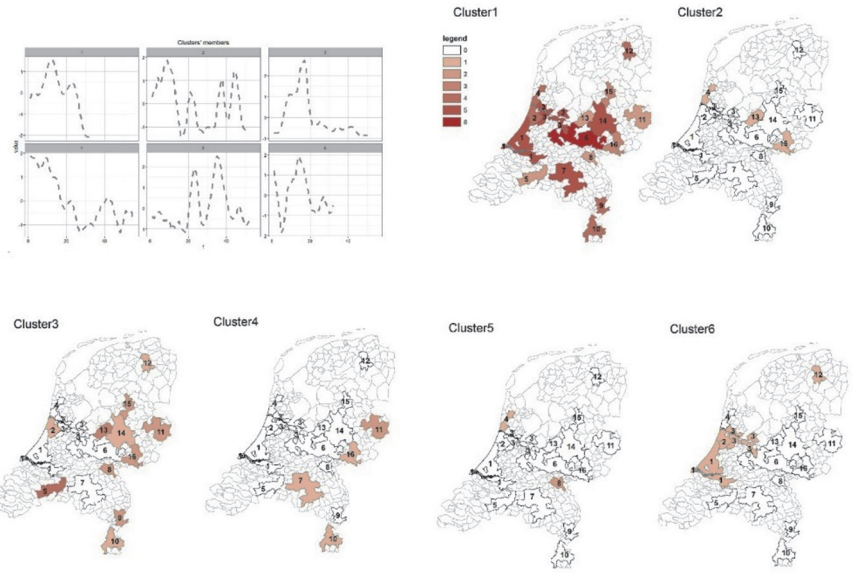

Figure 3-7 Final clustering results for the pertussis data showing the extracted centroids and the mapping of the 6 clusters.

Final CCR is show in Figure 3-8. This area consists of more than half of the original zones and has a distribution of the complete country. It is robust as all zones included were mapped >= 4 out of 6 epidemics to this cluster.

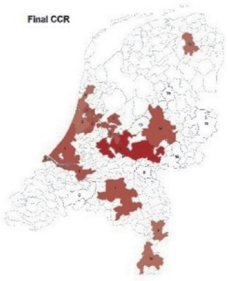

Figure 3-8 Final CCR

Table 3-3 Unequal length TP scores per centroid of cluster

Centroid of Cluster	TP	Mapping Count
1	1.007	60
2	3.293	3
3	1.088	21
4	2.113	5
5	3.025	3
6	3.417	4

From the observed distribution of mapping count (See: Appendix 3) we see that some of our clusters have very small counts. Discords or outliers are defined as time sequences that are relatively far from the closest match. However, all of our clusters do contain several subsequences. In time series clustering, self-matches are often the closest match but are considered to be trivial matches (Keogh et al., 2005). We can define discords in two ways, clusters that only contain subsequences of one zone, and clusters that only contain subsequences from the same epidemic as being discords. Under this definition, cluster 6 is a discord as it only contains subsequences of epidemic 2007-2008.

3.4 Conclusion and discussion

In this study we identified a CCR using clustering of time series. We tested this method on a synthetic dataset to determine the best performing distance measure and clustering techniques. The combination of SBD and hierarchical clustering outperformed the other tested combinations although the score were close.

SBD in combination with hierarchical clustering could differentiate between time sequences with single peaks and double peaks and endemic versus epidemic versions of the test data. Minor differences in the shape of the curve could not be detected Different combinations of distance measure and clustering technique had Rand values within a small range (079-0.85). Testing all methods on other datasets can reveal more information on suitability of these methods for CCR extraction.

The percolation method was used to extract geographic areas that are highly interconnected via roads. The extracted areas were compared to the 2013 commuting data and showed a good resemblance. This is encouraging as detailed commuting data may not be available for all countries and this provides an easy way of extracting zones based on more information than only administrative boundaries.

The CCR was extracted for pertussis in the Netherlands. It revealed that the CCR consists of a relatively large number of zones distributed over different parts of the Netherlands. This is not surprising as the country is small, daily commuting to other parts of the country is common and larger cities are located relatively close together. The delineated CCR can be validated using simulation techniques. Simulation studies can also be used to further validate the results of the percolation method.

In this research we assume that all our epidemics are valid samples. This may not necessarily be true. The fact that we found a cluster that only has incidences for 2007 indicates that this might be an outlier. A way to evaluate this aspect more closely could be to use a time series analysis technique where the whole time series is evaluated for discords.

We manually extracted the period for the subsequences at a national level. An automated extraction technique would be a good addition.
Our case study is relatively small yet, the methodology is generic and extendable to large databases covering much longer time periods and word wide data. The percolation method used to extract the zones could be applied on airline routes instead of roads by using airports as the centroids.

Chapter 4 Simulating informal settlement growth in Dar es Salaam, Tanzania: An agent-based housing model[2]

4.1 Introduction

Rapid growth of informal settlements (IS) is one of the largest problems of cities in developing countries. Informal settlements are urban areas that develop and grow without planning, and in which basic facilities are mostly lacking. The term 'informal' refers to the fact that these settlements are usually built without legal tenure and do not follow established building and planning regulations (Abbott, 2002). Important reasons for informal settlement growth are the weakness of statutory planning and a strong rural-urban migration, leading to substantial urban population growth. Although informal settlement growth is predominantly spontaneous, houses in these squatter areas usually exhibit particular growth patterns. Sliuzas (1988, p. 27) states that "the growth of a settlement is clearly not a random process and is likely to be influenced by a number of physical, cultural, and economic factors." Sobreira and Gomes (2001) argue that the geometry of informal settlements does not consist of irregularly distributed dwellings, but in fact displays a complex structure that can be defined by spatial patterns.

The Millennium Development Goals (MDGs) (United Nations, 2009) emphasize the need to "ensure environmental sustainability" and to achieve "a significant improvement in the lives of at least 100 million slum dwellers worldwide." For achieving this goal, the simulation of future informal settlements growth is an essential tool to manage the urbanization processes in the developing world in a more structured and sustainable way. Interventions like setting up of basic facilities for water distribution or health clinics can be better designed and implemented based on a proper understanding of growth processes taking place and driving forces influencing this growth.

So far, very few urban growth models have explicitly studied the growth of informal settlements. Sietchiping (2008) developed a cellular automaton (CA) model for Yaoundé, Cameroon that considers physical and socioeconomic influences. Barros (2004) studied the formation of low-income residential areas in the urban fringe of Latin American cities using an agent-based model. Sobreira (2005) developed an agent-based model for studying the contribution

[2] *Publishes as: Augustijn-Beckers, E.-W., J. Flacke, et al. (2011). "Simulating informal settlement growth in Dar es Salaam, Tanzania: An agent-based housing model." Computers, Environment and Urban Systems 35(2): 93-103.*

of global processes of urban growth to the formation of ISs and their changes in shape and size over time.

In this paper, we propose a spatially-explicit, vector-based, micro-scale simulation model for the growth of informal settlements that reveals the changing morphology of urban settlements. Although agent-based models can be implemented in a vector environment, modellers seem to be reluctant to do so (Crooks, 2010). The objective of this paper is to demonstrate a vector implementation of an agent-based model and to explore the added value of the vector-based implementation for simulating settlement growth.

Section 4.2 reviews recent approaches to urban growth modelling with a focus on informal settlements. Section 4.3 provides details about the case study area, the city of Dar es Salaam, Tanzania, and the growth processes observed there. Section 4.4 describes the conceptual framework for a vector-based housing model. Section 4.5 discusses analytical methods and the analysis results of the available empirical data. Section 4.6 describes the results of the simulation runs, while section 4.7 provides concluding remarks.

4.2 Urban growth modelling for simulating informal settlement growth

Growth of informal settlements differs from planned urban growth on the following points:

1. Essential is the concept that the individual house owner decides where to settle, based on personal preferences. This contrasts with formal settlements where usually planners decide on the location of plots/houses.
2. Diversity exists in the type of settlers (owners and tenants of different income classes and various socio-cultural backgrounds) as well as in the type and size of house constructed. (Sheuya, 2009).

Although there is no formal planning process, governmental involvement might take place, e.g. in the form of development of new infrastructure and upgrading of informal settlements to formal settlements.

Current techniques of urban growth modelling

Nowadays, two techniques of modelling urban growth are used predominantly, cellular automata modelling (CA) and agent-based modelling (ABM) (Batty, 2005, Benenson and Torrens, 2004). Urban CA models consist of a cellular representation of the environment (usually indicating land use), in which each cell has a state, a definition of its neighbourhood, and transition rules that determine its state may change depending on the state of neighbour cells.

Relevant examples for CA-based urban growth models can be found in Liu (2009) Rafiee et al. (2008) and Al-Ahmadi et al. (2009).

Agent-based models are simulation models in which decision-makers are represented as goal-oriented entities (agents), capable of responding to their environment and of taking autonomous action. Specific advantages of agent-based models include their ability to represent individual decision-makers and their interactions, to incorporate social processes, non-monetary influences on decision-making, and to dynamically link social and environmental processes (Matthews et al., 2007). Relevant examples of agent-based urban growth models are a model of an urban system of the South eastern parts of Michigan, USA (Brown et al., 2008), and an ABM for simulating emergent urban form in China (Xie et al., 2005), among others.

For the purpose of modelling the growth of informal settlements, raster-based CA models seem to be less suitable. As raster cells have no intrinsic semantic value (no ability to simulate shape), it is difficult to represent the shape and size of houses within a CA model. Furthermore, in CA models no explicit relationship exists between objects, like between a house and its owner. Another limitation of these models is the sensitivity to cell size and neighbourhood configuration (Moreno et al., 2009). Several lines of study exist to overcome these limitations as in vector-based CA models (Moreno et al., 2009) and incorporate real world geometry and dynamic or scalable neighbourhoods, to allow the representation of both local and regional dynamics (Vliet et al., 2009).

An ABM's ability to simulate individual behaviour and decision-making, in contrast, seems to be quite suitable to capture informal settlement processes and their effects on urban land-use change (Young, 2010). The decision to alter, extend, or erect a construction results from varying levels of dissatisfaction with present housing conditions, but is also highly dependent on the household's financial capabilities (Seek, 1983). In an informal setting such as ours, understanding individual behaviour is an intricate part of understanding settlement dynamics. Such understanding is expressible in an ABM.

Following Xie and Batty's (2003) observation "that pixel-based cellular dynamics seldom match spatial phenomena," Stanilow (2009) postulated to use parcel-based units instead of a raster-based tessellation in advanced urban growth models. This especially applies to the study of housing processes of informal settlement, where the exact size, location and orientation of a house in relation to other spatial entities is of importance, for instance, to assess infrastructure provision for single houses or to determine concrete measures of settlement upgrading. In this way, explicit spatial and geometric

relationships are incorporated into the model that allow to use GIS operations like buffering or point-in-polygon for defining neighbourhoods or analyzing the surrounding urban environment (Crooks, 2010).

Validation of ABM urban growth models

To demonstrate that a constructed model is a valid representation of the settlement processes that take place, a proper validation needs to be performed. Agent-based modelling is based on the principle that complex results will develop based on simple behaviour implemented at the level of the individual entity (constituent elements). To verify that the settlement rules implemented in the model are representative for the true settlement process, operational validation on the output should proof similarity between patterns that exist in reality and those observed in the simulated data. For vector-based ABM models, operational validation techniques for the comparison of simulated and empirical data are scarce, though various approaches to validate spatial simulation models exist.

Regarding the validation of raster-based urban growth models, Kocabas and Dragicevic (2009) distinguish between raster by raster map comparison approaches that compare the simulated output, pixel by pixel, against empirical data, and pattern comparisons that study the spatial overlap of certain geographic features and patterns with empirical data. To the first group belong techniques like relative operating characteristics (ROC) (Pontius Jr. and Schneider, 2001), chi-square and kappa statistics (Straatman et al., 2004) and multi-scale approaches (Kok et al., 2001). Pattern comparison methods mainly use spatial metrics (Herold et al., 2002) and goodness-of-fit methods (GOF)(Hargrove et al., 2006). The main criticism on raster by raster comparison is that disagreement between layers can be caused by data errors, instead of by actual land use change, and that the spatial distribution of errors in simulated data cannot be assessed (Hargrove et al., 2006).

Kocabas and Dragicevic (2009) developed an approach to validate agent-based land use change models using vector GIS and Bayesian networks for a pilot test area in Vancouver. Hargrove et al. (2006) developed a quantitative method for comparing categorical maps called map curves that graphically represent goodness-of-fit between a model output and real data. Both approaches mainly look at the degree of spatial overlap of land use polygons between empirical and simulated data. However, for the comparison of simulated growth maps of informal settlement with empirical data these approaches do not appear to be appropriate as the question at hand is not so much whether simulated new houses overlap more or less completely with existing houses, but whether the simulated spatial configuration and pattern of old and new houses matches with the real situation. In line with Parker et

al. (2001), we argue that these spatial patterns of the model are the emergent property of the ABM that has to be validated to evaluate the quality of the ABM. For this approach, new spatial metrics must be developed that quantify the housing pattern in terms of increasing compactness of a settlement and/or the ongoing outspread/extension of a settlement.

4.3 Case study

Case study area

Dar es Salaam, being the largest Tanzanian city with an estimated population of 3.5 million in 2000 is growing at an average annual rate of 8 % (2005). In 2003, there were 150 informal settlements (WorldBank, 2002), characterized by high housing densities, unstructured road infrastructure and inadequate water, electricity, and sewerage services. It is estimated that about 70 % of the city's population lives in informal settlements (Ramadhani, 2007), with an expected increase to over 80 % in the coming years.

Informal settlements in Dar es Salaam are unplanned but the ownership of the land is normally obtained via legal means. The early settlers had *de facto* ownership over the land through the customary land tenure, while later, as the land was included in the urban administration, an informal land market developed, through which the settlers most commonly purchased a plot (Kyessi, 1993).

Our area of study is Manzese settlement, an informal settlement of approx. 100,000 inhabitants (in 1980) located 6 km northwest of the city centre. The squatter settlement is inhabited by relocation from other (rural) areas. Initially, the area was located at the city boundary (periphery) but it changed into an inner-city densification area. A study by Kironde and Rugaiganisa (2002) showed that in the squatter settlement of Manzese 75 % of the land owners purchased the land.

Housing processes in Manzese squatter settlement 1967–1993

Detailed studies of Manzese that monitored the growth of the settlement, were carried out by Kajagi (1982), Sliuzas (2002) and Kyessi (1988). Kajagi (1982) identified two general growth patterns that were characteristic of the growth of Manzese settlement between 1967 and 1882:

1) **Extension**: houses were built on previously vacant land and hence increased the built up area, and

2) **Infill** (or densification): houses were built on the remaining land within the already built up area.

Infilling appeared in two forms, either *freestanding* (location of a new building relative to adjacent buildings indicates an independent function) or *grouped*, when the location of a new building relative to adjacent buildings indicates that there is a functional dependency with one or more buildings. De Bruijn (1987) observed that over time a settlement consolidation process took place that led to an enlargement of certain houses. House enlargement was seen as a process towards the typical Swahili house, that consists of two buildings, a small building (3 rooms) and a larger building (6 rooms) (Kyessi, 1993), whereas either the small building was constructed first and later the large building was added, or this happened in reverse.

Several factors influenced the growth process described above in Manzese. Sliuzas (1988, Sliuzas, 2004) identified that informal settlements are attracted by environmental hazardous areas (poor land quality, high slope, flooding zones) and that development in the area was influenced by the existence of roads and footpaths. This is in line with findings of Sietchiping (2005), who explains the increase in slum dwellers in developing countries in general as being triggered by marginal and less valuable lands, such as riverbanks and slopes, along transportation networks and near places of economic activity.

The area of Dar es Salaam is subject to flooding, although floods do not occur very frequently. A 1991 survey revealed that 53 % of the occupants in Manzese area do not see flooding as a problem (Kyessi, 1993). However, before 1982 the lower areas of Manzese were left almost empty and only during the period 1982–1987 did the growth rate on worse (lower) lands increase due to ongoing densification on the better (higher) land.

Economic factors that influence informal settlement growth are land tenure, housing tenure, value of the land, socio-economic characteristics of the settlers including their status of employment. Kyessi (1993) found that room rental is a common phenomenon in Manzese settlement and that it has an important role in the settlement process. The 1991 survey showed that 36.3 % of the buildings were owner occupied, 39.7 % were occupied by owners and tenants and 12.8 % were occupied by tenants only. Intake of tenants is linked to the consolidation of existing houses by house enlargement. Taking tenants generates extra income for the house owner and this enables the house owner to enlarge her/his house. About 50 % of the land owners had the intention to enlarge their houses.

4.4 Simulation

Informal settlement is a combination of attraction to favourable conditions and avoidance of non-favourable ones. Different actors play a role, ranging from the individual settler to the government, and all influence the settlement process. Where a person settles, depends on personal preferences that fall into one of three categories: *economic factors*, *social factors* and *physical factors*. The settler must be able to afford the location, is attracted to locations close to centres of economic activity, is likely to settle close to people with a similar background, and will choose the best landscape conditions, trying to avoid wet areas, steep slopes and other unattractive places.

General framework of the ABM

Agent-based simulation models are built upon the concept of agents that are located within an environment, maintain relationships with this environment, and interact with each other as well as with the environment. The agent is goal oriented in that it tries to accomplish a certain goal and displays behaviour that may lead to achieve this goal. An environment can consist of a spatially explicit representation of the real world; it can be either static or dynamic. Agents can change their environment and the environment can influence the behaviour of the agent.

In our model, two different types of agents exist: the home owner and the renter. They both represent a household of people living in one house. One important, shared goal of both renter and house owner is to live in a suitable home. An additional goal of a house owner is to find a place to settle in an attractive location and construct a new house that matches her/his income characteristics. The tenant strives for finding an already existing house to rent a room. Agents have different characteristics, for example, they have different income levels.

The environments in our model consist of spatially explicit, vector-based layers of houses (polygons), roads and footpaths (lines), elevation (zones based on contour lines) and swamp areas (polygons). The settlers are represented as agents and their location during the simulation is stored in an output layer. Houses are captured in a dynamic layer: it changes during the simulation, based on actions of house owner agents. The owner agent is able to create new houses and to extend existing house structures. All other layers are static during the simulation. Environments are created based on empirical data of the situation in Manzese settlement (see section Case study area).

Agents can have different types of relationships with their environment. In our simulation, agents live in a house, meaning that the agent has an ownership

or renter relationship with the house for the duration of the simulation. An assumption of the simulation is that after settling down, the owner or renter agent will not move again. A house is not equal to a building as a large Swahili house consists of multiple buildings that together make up the house. Agents have the ability to sense their environment to determine the suitability of the location before they settle. Agents are able to sense all aspects of the environment that influence their settlement decisions (see section Movement of agents).

The aim is to build a model with a level of complexity that matches complexity of building structures seen in real informal settlements. This complexity is introduced by means of the complexity in the attributes and behaviour of the agents and their interactions with the environment.

The following paragraphs explain how this heterogeneity is captured via different house construction rules, agent movements and settlement behaviour.

House construction rules

The house construction observed in Manzese settlement (described in Section 4.3) was translated to the model by distinguishing three types of building, small 3-room buildings, medium-sized 6-room buildings, and large Swahili houses. The typical Swahili house consists of two buildings; the medium-sized building facing the road, with the small building located behind it. The simulation does not contain an entity type "Parcel" to model plots that houses are located on. We assume that a large Swahili house covers the total area of land owned by the agent.

Three change mechanisms represent the construction of houses:
1. *Extension*: house constructed on any suitable vacant land. This mechanism ensures that already existing built-up areas can be extended.
2. *Infilling*: a house is constructed in the direct vicinity of (or is attached to) an already existing house.
3. *Enlargement*: Already existing small or medium-sized houses are enlarged to a full Swahili house.

Extension as implemented in this model allows for the simulation of non-attached houses, within an already populated neighbourhood, but not chained in a contiguous block of attached buildings. As indicated by Sobreira and Gomes (2001) informal settlements consist of a combination of large building blocks (many attached buildings) and small building blocks (few attached houses). The extension rule ensures that detached houses are created that can develop into small building blocks. Extension locations can be on the opposite

site of the road from already existing houses, or on undeveloped streets within the informal settlements.

The process of *infilling* is based on findings of several authors, who studied the morphology of informal settlements (Sobreira and Gomes, 2001, Sudhira et al., 2005). They state that informal settlements are comprised of an irregular distribution of groups of houses that are chained together (basic units).The infilling rule confirms that houses will be "chained together" in contiguous dwellings. It is a way to ensure that a required compactness of the settlement will be reached and no small, irregularly shaped empty spaces will be left over between buildings.

Enlargement of existing houses is based on the concept that house owners try to create a house fitting their income. In our model, both agents and tenants can belong to a low, medium or high income group. These income groups are relative notions, only indicative of financial means available for house construction. The model assumes that eventually all agents build a house with the same size, but they will not do so immediately. Rich people build a complete large house, but poorer settlers will start with a small or medium-sized house, and will later try to extend it to reach the full size. To make this possible, the model reserves the area of a complete house at settling time, independent of the size of the house that the agent is building. Enlargement of existing buildings will lead to an increase in compactness of the settlement.

Temporal resolution of the model is one day. The number of agents created depends on the target growth percentage for the simulated time period. Agents determine in each time step whether they target an infilling or an extension location. Their behaviour and movements change accordingly. Time allowed for agents to find a suitable location to build a house is limited and after the elapse of the search period, a number of agents will remain unsettled. These agents represent prospecting home owners that decide to settle elsewhere in the city.

Movement of agents

As prospective house owners may decide to settle in any vacant plot, the model assigns new house owner agents to random free locations in the modelled area. Movement behaviour differs between infilling and extension agents. The movement process of the extension agent remains random. For the densification process, agents are actively searching for existing houses. Initially, the agent is created at a random location but will relocate to the centroid of the nearest already existing house. This is done to ensure that agents actively look for locations close to already existing buildings. Movement of the agent is defined as movement from centroid to centroid of existing

houses. For each move, the agent selects a new house within a search radius around the current location and relocates to this house. Figure 4-1a shows the movement of an infilling agent during a number of time steps.

Implementation of agent behaviour

Settlers in general have a preference to find locations for their new homes that are close to roads (or footpaths) and high enough to avoid flooding areas. The flood zone was defined as the area below the 18 m elevation line. The flood zone is not the same as the swamp area. Only the lowest area is actually a swamp area, most land below the 18 m line only floods occasionally. The avoidance of swamp areas and lower suitability of flood zones is implemented in the general preference (GP) of all agents in the model. Other settlement behaviour differs per agent type.

Extension
For each random location, a dummy building with the position of the agent as its centroid is created and this building is re-aligned to the nearest road or footpath. We verify that the building does not contain, or intersect with, any other building, road or footpath. A buffer is created around the building to check for other buildings in the direct neighbourhood. If other buildings are close, the location is not suitable for extension (it is considered an infilling location). When a location is potentially suitable, a check is conducted to evaluate the general preference conditions. When the GP is sufficient, a permanent building is created. (See Figure 4-1b)

Infilling
Amongst settlers it is common to search for a place close to existing buildings. In the model, infilling agents show this behaviour. Before they settle, they will check for vacant locations around existing buildings. When sufficient space is available, the agent will check the attractiveness of the location by creating a dummy building and calculating its suitability using the general preference conditions. When
sufficient criteria are met, the agent will settle. Houses will always be aligned to transport infrastructure and existing houses. The agent will identify the corner of the existing house that is closest to the road or footpath and identify the vertices of the edge it belongs. This information is used to calculate the coordinates of two new vertices to create the new house. Figure 4-2 shows the alignment process and the creation of a dummy building.

Building enlargement
When a house owner, currently occupying a smaller house, acquires more financial means, s/he may extend the existing house. Analysis of settlement

patterns for Manzese settlement has revealed that the enlargement of small houses, with a medium-sized building, is always situated between the existing house and the road, and that extensions of medium-sized houses (with a small house) will always take place behind an existing building.

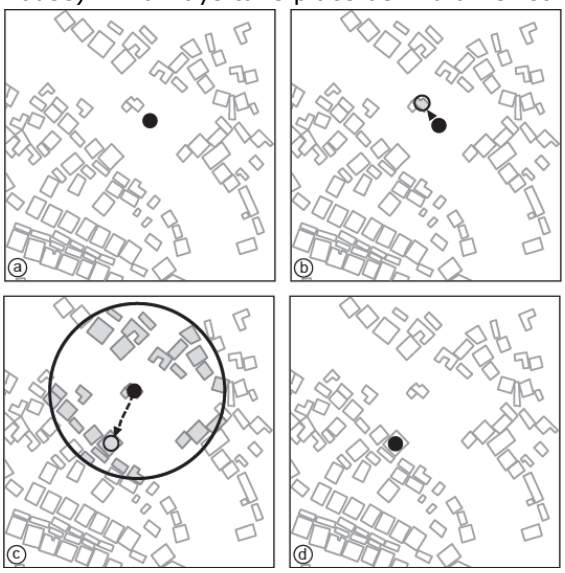

Figure 4-1 Movement of the infilling agent. Step a: agent will be created at a random location. Step b: agent moves to the centroid of the nearest existing house. Step c: agent will select all houses in its vicinity and randomly select the next house within the search area. Step d: agent will move to the new location.

Figure 4-2 Movement of the extension agent. Step a: agent is created at a random location. Step b. check for other houses in the direct neighborhood. Step C. move to new location. Step d: house construction.

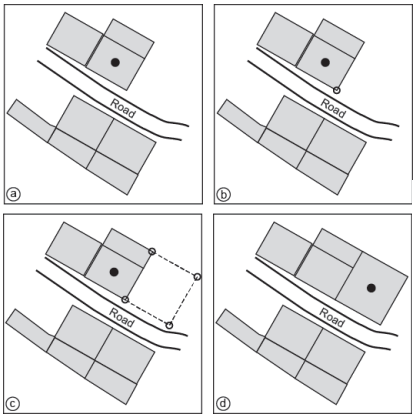

Figure 4-4 House construction process

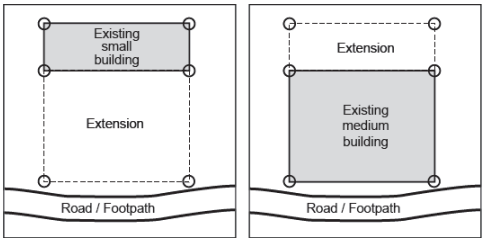

Figure 4-3 Extension of a small building (left) toward the road, and extension of a medium building (right) behind the existing building.

Enlargement rules are similar to rules for the construction of a new house. The simulation will check to find the corner closest to the road. It will identify the edge parallel to the road and will calculate two new vertices in the direction of the road, or will find the opposite edge of the building to extend backwards. The process of building enlargement for a small- and medium-sized house is shown in Figure 4-3.

4.5 Methods and analysis of the empirical data

Analyses on historical growth of Manzese settlement were conducted to allow for a comparison with the simulated results (section 4.6). The first group of analyses (section *Roads, footpaths, flood zones*) relate to the general settlement preferences. Analyses are performed to determine the alignment of houses in relation to infrastructure and the number of houses located in flood zones. The second group of analyses focuses on the processes of settlement extension (section *Extension of the settlement area*) and infilling.

Five different datasets are available representing the settlement situation in the years 1967, 1975, 1980, 1982 and 1987. Vector based data were partly obtained from the municipality of Dar es Salaam. Other data originate from satellite images, verified by field work conducted by ITC over a large number of years. These data include spatial layers showing the location of buildings, type of buildings, date of construction and number of inhabitants. Analyses results are provided in Table 4-1.

Table 4-1 Analysis results empirical data

	road %	Foot path %	Flood zone %	MD (m)	SD (m)	MMD (m)	SI (m)	mean S	max S
1967	86	14	25	228.7	184.9	14.9	608.2	1.1	3
1975	64	36	58	270.6	213.8	10.9	2077.5	1.2	6
1980	41	59	80	287.6	227.0	10.6	2382.5	1.2	7
1982	38	62	79		232.9	10.4	2445.5	1.2	7
1987	40	60	82		243.6	9.9	2825.5	1.3	7

Roads, footpaths, flood zones

In 1967, the settlement contained 186 buildings, which increased to 1280 buildings in 1987. Data were analysed to determine whether the orientation of new houses was towards a road or a footpath. When

$$\frac{distance\_road}{2} \leq distance\_footpath \tag{4.1}$$

the orientation of the new house is assumed to be towards the road, otherwise orientation is towards the footpath. In 1975, 577 new buildings were constructed, of which 517 were new houses. Of these houses 64 % were aligned towards a road. In 1980, another 244 new buildings were created (227 new houses) of which 41 % were aligned towards a road. In the period 1982 to 1987, the settlement increased by 246 new buildings (222 new houses) of which 40 % were built along a road and 60 % along a footpath. During the years, the preference to build close to a road seemed to decrease, as the majority of the houses was initially along a road, whereas in later years, the majority was oriented towards footpaths.

In 1967, 46 houses (25 %) were located in the flood zone. During the period 1967 to 1975, 301 new houses were built inside the flood zone. This is 58 % of the new houses built during this period. In the period 1975–1980, 80 % of the new houses (182) were built below the 18 m level, this even increased to 82 % in the time frame from 1982–1987. As less land was available in the higher regions, almost all new development took place in the flood zone.

Extension of the settlement area

Our assumption is that extension of the built-up area primarily took place during the first time period (1967–1975) and that in the two subsequent time periods (1975–1980 and 1982–1987) infilling was the predominant mechanism.

Extension is measured by the mean distance (*MD*) between building centroid and other building centroids of the settlement. Mean distance is defined as

$$MD = \sum_{i=1}^{n} \sum_{j=1}^{n-1} \frac{\sqrt{(x_i - x_j)^2 + (y_1 - y_j)^2}}{n(n-1)} \tag{4.2}$$

in which x_i and y_i are the coordinates of the centroid of building *i*, and *n* is the number of houses. The empirical data show an *MD* for 1967 of 228.7 m with a maximum of 693.6 m, and a minimum of 4.7 m. In 1975, the *MD* increased to 270.6 m (max: 747.9 m, min 0.5 m) and by 1980, *MD* was 287.6 me, which is a small increase compared to the first years.

The standard distance (*SD*) was also used as a measure for the growth of the settlement. The standard distance is calculated as:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2 + (y_i - \bar{Y})^2}{n}} \tag{4.3}$$

in which again x_i and y_i are the coordinates of the centroid of the building *i*, $\overline{X}$ and $\overline{Y}$ represent the mean center feature, and *n* equals the number of buildings. The *SD* increased from 184.9 m in 1967 to 213.8 m in 1975, 227.0 m in 1980, 232.9 m in 1982 and 243.6 m in 1987.

Although the most significant increase in *SD* and *MD* occurred during the first years, the increase continued in later years. This indicates that some growth did take place in later periods.

Infilling

Two processes take place that together characterize the infilling within the already built-up area: enlargement of existing houses, and construction of new houses between already existing houses. Due to the latter process, the mean distance to the nearest house will decrease.

In 1967, the mean minimum distance (*MMD*) between the centroids of the houses was 14.9 m, this decreased by 1975 to 10.9 m, to decrease slowly to 10.6 m in 1980 and 9.9 m in 1987. It is not possible to use the mean minimum distance to differentiate between extension of existing houses and infilling (new attached houses).

During the infilling process, building units are added to already existing building blocks creating larger islands of buildings. This is evaluated using the Shape Index (*SI*). The shape index is calculated as:

$$SI = \sum_{i} \frac{Perimeter_i}{\sqrt{Area_i}} \tag{4.4}$$

in which i denotes a polygon counter for connected houses. The less compact the area, the higher the shape index, as compactness is related to the shape and size of the islands of attached houses.

In the empirical data, the *SI* for 1967 was 608.2, this value increased for 1975 to 2077.5, it further increased for 1982 to 2445.5 and for 1987 it was 2825.5. There is a steady and gradual increase in values (decrease in compactness). This indicates that although new buildings are constructed, the block size (number of connecting buildings) is decreasing rather than increasing. This can be explained by the fact that small openings exist between the buildings (although the buildings are close, they are not attached) and the fact that in the initial settlement the number of medium sized houses was large compared to the number of large houses (existing of both a medium and small house) and the number of small houses.

Another measure for the compactness of the settlement is the number of dwellings per building block (*S*) and the diversity of building block size. *S=1* indicates an isolated house, *S=2* characterizes a pair of attached houses. Informal settlements are known to be characterized by a small number of large islands, and a larger number of smaller islands (Sobreira and Gomes, 2001).

The mean frequency is defined as the mean number of dwellings (mean *S*) per building island. The values of *S* vary from 1.1 with a maximum frequency of 3 for 1967, to 1.2 with a maximum of 6 for 1975, 1.2 and 7 for 1980 and 1.3 with a maximum of 7 in 1987. The relatively low value for the islands is in line with the results of the SI.

4.6 Simulation results

To test the functionality of the model, a total of 150 test runs was conducted, divided over three time periods, 1967–1975, 1975–1980 and 1982–1987. For each time period, the existing buildings were loaded into the model. The increase in population was calculated from the empirical datasets. Apart from the increase in population and the initial buildings, all settings were fixed except for preference for infilling or extension. No new roads or footpaths were created during the simulation. Input for footpaths was the situation of 1967.

Analysis of empirical data showed that roads and footpaths in the study area were already in existence before the development of the area.

Test runs are evaluated below to determine if the general preferences for site selection, attraction to roads and footpaths and avoidance of the flood zones, provide good results. Furthermore, the analysis of the empirical data show that infilling and extension take place during the complete period. In the simulation, it is possible to mix the percentage of infilling and extension. Tests are conducted to determine whether a different mix of rules indeed leads to a different result and if the trends observed in the empirical data can also be detected in the simulations. Results of these analyses are shown in Table 4-2 and in Figure 4-4 and 4-5.

Roads, footpaths, flood zones

For the three time periods, the alignment of new houses with roads and footpaths was calculated, as described in the Section Roads, footpaths, flood zones. For the period 1967–1975, the simulation created an average of 77 % of new houses along a road, compared to 64 % in the empirical data. For the

Table 4-2Analyses results simulated data

	Extension rule	Road %	Footpath %	Flood zone %	MD (m)	SD (m)	MMD (m)	SI (m)	Mean S	max S
1967-1975	0%	74.5	25.5	38.5	238.2	189.1	9.2	806.1	1.7	9
	25%	78.8	21.2	62.6	280.6	221.4	10.1	1675.4	1.6	9
	50%	79.0	21.0	65.0	282.8	222.9	10.4	1919.7	1.6	12
	75%	76.8	23.2	66.6	286.2	225.7	10.9	2084.9	1.5	10
	100%	77.3	22.7	90.7	291.2	229.5	14.4	2411.5	1.5	10
1975-1980	0%	70.6	29.4	56.9	272.1	214.8	9.2	2575.4	1.6	12
	25%	73.1	26.9	58.6	276.1	217.8	9.6	2684.9	1.6	9
	50%	72.7	27.3	61.2	278.5	219.7	9.8	2759.5	1.5	12
	75%	71.9	28.1	62.9	280.1	220.9	10.0	2842.1	1.5	10
	100%	67.7	32.3	69.9	283.3	223.3	10.3	2956.6	1.5	10
1982-1987	0%	65.8	34.2	62.8	292.7	230.9	9.1	3104.5	1.6	12
	25%	69.8	30.2	63.4	294.6	232.4	9.2	3153.3	1.5	10
	50%	72.5	27.5	62.8	295.2	232.9	9.3	3203.4	1.5	12
	75%	70.0	30.0	65.2	296.0	233.5	9.4	3250.6	1.5	10
	100%	73.0	27.0	69.3	297.0	234.3	9.5	3299.9	1.5	10

period 1975–1980, the average number of houses aligned to a road in the simulated data was 71 % compared to 41 % in reality. This was repeated in the period 1982–1987 with 70 % alignment along roads. The preference for a location along a road as implemented in the simulation is stronger than observed in the empirical data. Currently, this preference is not dependent on the availability of land, but is fixed to twice as preferable as along a footpath. Results indicate that the preference to build along a road should be made space

dependent. When less space to build new houses is available preference for roads becomes less important.

For the period 1967–1975, the simulation created 62 % of the simulated buildings in the flood zone, compared to 58 % in reality. For the period 1975–1980 and 1982–1987 periods, the simulated data showed 62% and 70% of the new houses in the flood zone area. In the real data, the percentage was higher (80 % and 82 %). Results indicate that settlers indeed avoid flood areas although this avoidance behaviour is not strong.

Infilling versus extension

Three housing rules were implemented that together should lead to both infilling and extension of the settlement. The enlargement of buildings is fixed and takes place during all time periods. This amount of infilling and extension was varied from 0 % extension (100 % infilling) to 100 % extension (0 % infilling) in steps of 25 %. The results were analysed and compared to the results for the empirical data. Results of the simulation runs are shown in Table 4-2 and Figure 4-4.

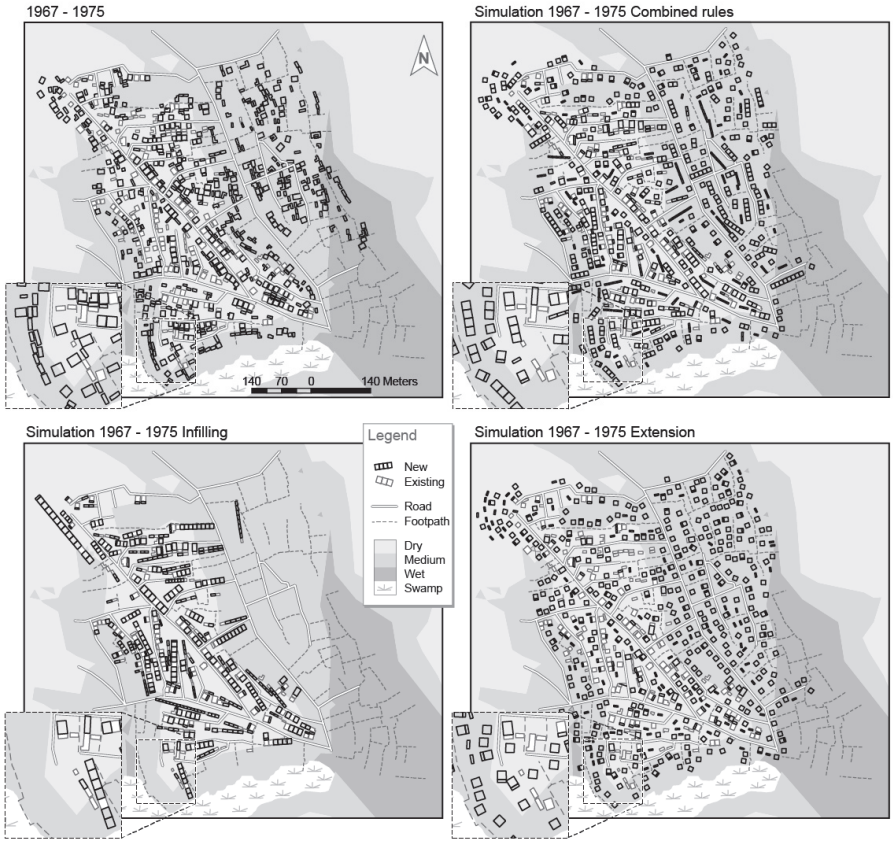

Figure 4-5 Comparison of the different house construction rules (top right 50% infilling/extension, lower left 0% extension, lower right 100% extension) with the empirical data (top left) for the 1967-1975 period

Extension

The mean distance (MD) between the centroids of the buildings (Table 4-2) shows for all time periods that the distance depends on the percentage of infilling/extension. In all cases, the 100 % infilling leads to the smallest mean distance, and 100 % extension leads to the largest. When comparing the different time periods, the mean distance increases over time, this is in line with the expectations and empirical data, the settlement is growing.

Extension of the settlement was discussed in the Section *Infilling*. The *SD* showed an increase from 1967 onward, with the first period showing the largest growth of the settlement. When looking at the simulation runs (Table 4-2), evaluating the different percentages of infilling/extension, for all time periods, there is an increase of extension with a higher percentage of the extension housing rule. For the 1967-1975 simulation period, the SD ranges between

189.1 – 229.5 (empirical value 227). The 75% and 100% extension leads to the best results. This is in line with our expectations. For the other two periods, the increase of SD for the simulated data is low compared to the empirical data.

Infilling
The mean distance to the closest house (MMD) varies with the housing rule during the first period, with 0 % extension leading to the smallest distance and 100 % extension leading to the largest distance. The initial 1967 value was 14.9 m. The value for the 100 % extension rules is too high in comparison to the real data; a mix of rules is therefore necessary to simulate the infilling process. For the two later time periods, the effect of the housing rules on the mean distance is relatively small but all of the results are in the same order as the value of the empirical data.

The shape index (*SI*) shows a similar result as the mean distance, in the sense that the mix of infilling/extension has most effect during the first time period and that the simulated results in general show the same pattern of increase over the years compared to the empirical data. The mix of the housing rules leads to better results than applying a single housing rule.

The mean frequency of the simulation runs is shown in Table 4-3. For all runs, the 100 % infilling rule leads to the highest frequency. The mean frequency Interesting is that the maximum frequency for the 50 % runs is higher than decreases with the percentage of extension, this is according to expectations. for the 25 % case. In general, there is a very small increase in mean frequency between the three time periods from 1.20 for 1967–1975 to 1.25 for 1982–1987.

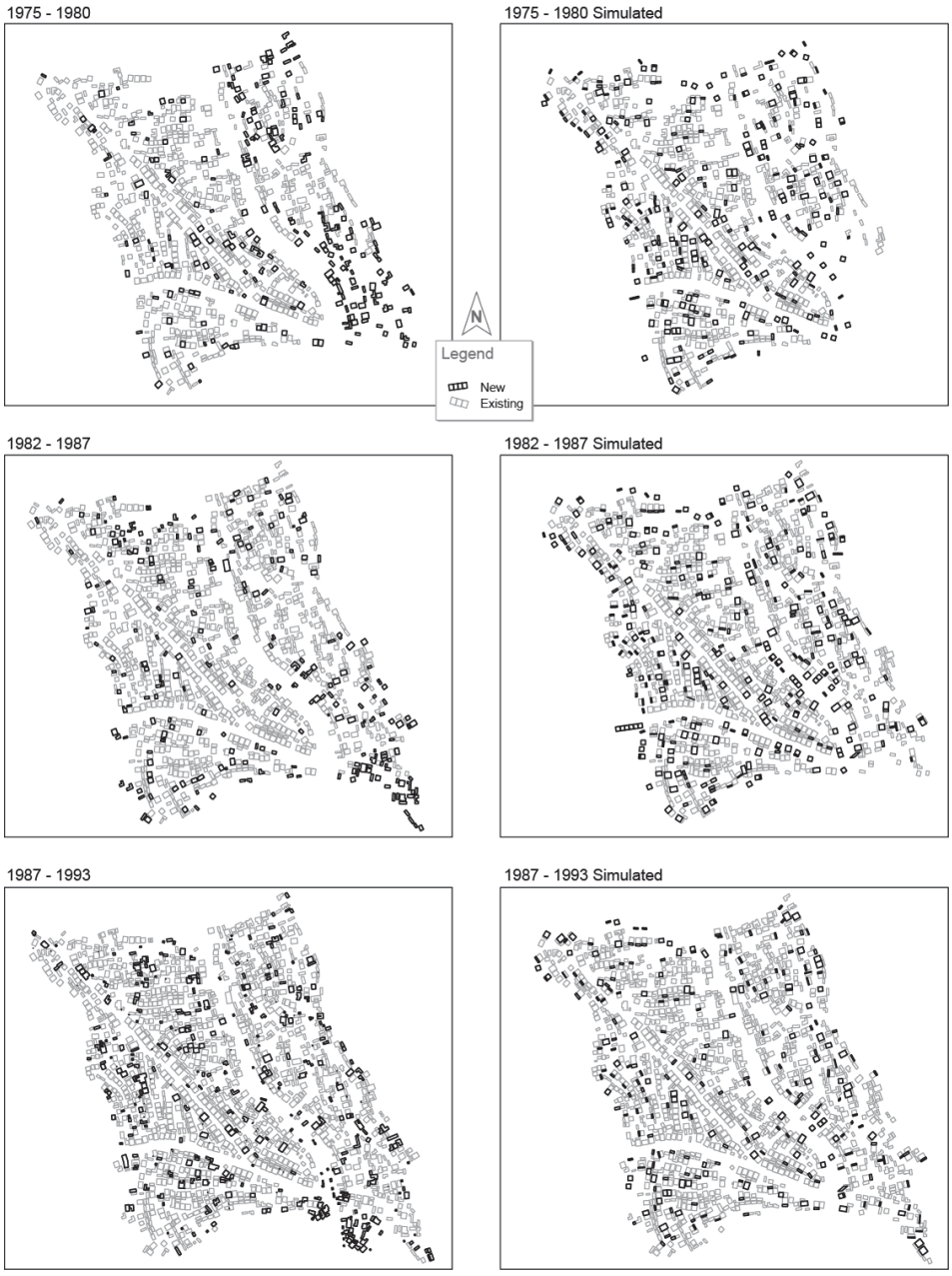

Figure 4-6 Results for different time periods, empirical data on the left, simulated data on the right.

4.7 Discussion

Rapid growth of informal settlements is a large problem in developing countries and an improved understanding of the process could help to indirectly steer it. Our work shows that a vector-based housing simulation model is a promising tool to improve this understanding. The settlement pattern resulting from the various simulation runs is consistent with patterns observed in empirical data for the Manzese IS over time periods of five years. Simulating the geometry of the settlement can help to determine how many houses can still be built in the settlement, where these houses will arise, and what the impact of roads, flood zones and other landscape factors will be on (further) development of the area.

The general preferences in our model only include the distance to roads, footpaths, flood zone and swamp as physical features. This can be extended to include slope, water distribution points, and possibly the access to further utility infrastructure. Especially more wealthy people are probably attracted to areas with access to electricity and water. The attraction to roads instead of footpaths should be made dependent on the density of the area. In later years, when little space is available, this criterion appears to become less important. The same applies to avoidance of the flood zone: the model has a stronger avoidance of the lower areas compared to the empirical data.

Three house construction rules were implemented, of which the impact of extension version infilling was varied during the simulation runs. The influence of the percentage infilling versus extension has a clear impact during the 1967–1975 runs. The combination of the two rules gives better results than the use of a single rule. It should be noted that the way in which the metrics are calculated influences the detectability of the impact of housing rules on the results, as the number of new buildings in relation to existing buildings gradually decreases with time.

The housing model can be further improved by changing the static footpaths to a dynamic footpath layer to simulate the development of paths (and roads) together with the building process. There is no mechanism of selling houses, so agents that settle in the area will not leave; this is an unrealistic limitation of the current model.

Results of the analysis of the empirical data showed that both expansion and infilling takes place in all time periods but extension decreases in the later periods and is more obvious in the time frame 1967–1975. During the later periods, the process of infilling dominates. In order for the model to run over longer time periods (more than 5 years) the mixing of extension and infilling should be made dynamic in such a way that it becomes dependent on the development state of the settlement.

Although this model was developed for informal settlements, it has wider applicability. The principles of building next to an existing house, building detached houses and enlargement of existing houses are generic and can be adjusted to different types of settlement. However, the housing model presented here is not usable for situations in which the settlement is steered by a planning process, as it assumes the freedom of the owner to settle anywhere as one of the guiding principles. Other than informal settlement processes exist where this principle is met. It is interesting to test the model on a historic city as the growth process of these cities is also based on spontaneous developments and inhabitant choices.

Chapter 5 Agent-based modelling of cholera diffusion[3]

5.1 Introduction

Cholera is a disease spread by *Vibrio cholerae*, causing diarrhea and severe dehydration in about one out of 20 patients. Cholera can be endemic, leading to seasonal outbreaks, or epidemic. According to the World Health Organization (WHO), cholera incidence has increased globally since 2005 with in 2012 48% of cholera cases occurring in Africa (WHO, 2014).

Cholera infection can be caused by ingestion of food or water contaminated by *V. cholerae* and has two distinct life-cycles, one in the environment and another in humans (Harris et al., 2012). The pathogen occurs naturally in coastal waters, preferring brackish water and can live in association with zooplankton and shellfish (Harris et al., 2012, Sedas, 2007). The intake and passage of *V. cholerae* through the human body results in conversion of the pathogen to a hyperinfectious state. When shed via faecal excretion of infected individuals, hyperinfectious bacteria can be re-introduced into the environment and pose a severe risk to other individuals as the infectious dose is 10 to 100 times lower compared to natural, non-human shed low-infectious organisms (Harris et al., 2012). When present in the environment either as a natural pathogen or in hyperinfectious state, bacteria can be transported via rivers, leading to further propagation of the disease to previously uninfected areas potentially causing new exposure.

Health risks also depends on human factors, like the cultural and socio-economic environment (Tamerius et al., 2007). When a community has access to safe water and does not use surface water in their daily routines (drinking water, sanitation, hand-washing and food preparation), chances of infection are minimal (Penrose et al., 2010). According to a study by Devas and Korboe from (2000), only 30% of Kumasi households had satisfactory sanitation arrangements in their own home, 40% depended on public toilets and 24% of the households were using buckets. With public toilets often having waiting queues, people (especially children) relieve themselves on open dumpsites. Besides problems with sanitation, Kumasi also struggled with water supplies. Although most households have access to tap water, these are unreliable (Devas and Korboe, 2000).

[3] *Published as: Augustijn, E.-W., T. Doldersum, et al. (2016). "Agent-based modelling of cholera diffusion." Stochastic Environmental Research and Risk Assessment: 1-17.*

In recent research on the 2005 cholera outbreak in Kumasi, it has been suggested that dumpsites played a role in the spread of hyperinfectious *V. cholerae*. Due to fast urbanisation and growing population, Kumasi Waste Management Department (WMD) in 1999 collected only 40% of the total waste (Post, 1999) leading to many open refuse dumps. Spatial dependency of cholera infections on the proximity to and density of refuse dumps was shown by Osei et al. (2010). This implies that runoff from dump sites could carry faecal materials to local rivers, creating a pathway for faecal contamination of surface water. A strong increase in the number of communities reporting cholera cases during/after rainfall periods underlines this hypothesis. A study from Obiri-Danso et al. (2005) revealed high bacteria counts (faecal coliforms) in the Subin river running through Kumasi in 2000-2001 during the rainy season. During periods of rainfall, Kumasi suffers from sporadic water shortage, leading to a higher use of river water. The booster effect of rainfall on cholera outbreaks has been reported for a number of other countries like Haiti and Guinea-Bissau (Gaudart et al., 2013, Luquero et al., 2011). Effect of extreme precipitation on waterborne diseases in the United States was reported by Curriero et al. (2001)

The link between disease and inproper solid waste disposal has gained more attention lately (Ayomoh et al., 2008, Kwasi Owusu Boadi and Kuitunen, 2005, McMichael, 2000, Obiri-Danso et al., 2005). Worldwide, the problem of proper solid waste disposal increased because of growing urbanization, increase in the amount of solid waste produced per household, and lack of proper waste collection, dumpsite planning and treatment methods. In Ghana, Asomanin Anaman and Bernice Nyadzi (2015) and Bagah et al. (2015) studied the improper disposal of solid waste showing the magnititude of this problem in Accra, and linking this to a cholera outbreak. Rego et al. (2005) found a link between diarrhoea and garbage disposal in Brazil. Abdul (2010) related cholera to solid waste disposal in Swaziland for homes located within a 200 meter distance of dumpsites.

Modelling provides a good means of gaining deeper understanding of the process that caused the 2005 cholera outbreak, as it can provide more insight in disease diffusion patterns (Meade and Emch, 2010, Mikler et al., 2007). Most cholera models fall back on a model developed by Codeço (2001) which was extended to include a transient hyperinfectious state of the pathogen by Hartley et al. (2006). Several geographically explicit cholera models have been developed including both natural and hyperinfectious cholera transmission including transportation via river networks (Mari et al., 2012, Bertuzzo et al., 2009, Bertuzzo et al., 2008). Most of these models focus on spread of cholera between villages and cities linked by a river, coupling a local disease model with a transport model in order to model cholera diffusion. These models provide excellent tools to study cholera diffusion patterns at a larger scale, but

are less suitable for micro-scale modelling. Although these models simulate the transport of the pathogen via the river pathway, they do not include the route of *V. cholerae* to the river. This is not easily incorporated into the network structure as there is no permanent flow of water, but flow only occurs after heavy rainfall.

Investigating the potential transport route of *V. cholerae* from dumpsite to river is interesting, because interventions can easily be executed (e.g. relocation of dumpsites) when the contribution of runoff as a transport mechanism can be proven. Agent-based models (ABMs) are particularly suitable to perform geographically explicit micro-simulations, which include both human behaviour as well as environmental aspects. They have proven to be particularly useful when assessing control measures (Dommar et al., 2014). A strategy called pattern-oriented modelling (POM) which attempts to replicate multiple spatial and non-spatial patterns observed in the real system is often applied for ABMs (Grimm et al., 2005).

We therefore propose an agent-based model including a disease sub-model and a hydrological sub-model to perform micro-scale simulations to determine if open refuse dumps could have played a role in cholera diffusion during the 2005 cholera outbreak in Kumasi, Ghana. This article is structured as follows: section 2 contains the conceptual design of the agent-based model, section 3 describes the experiments conducted to test the hypothesis of runoff from dumpsites causing cholera diffusion, followed by the discussion (section 4) and the conclusions and recommendations (section 5).

5.2 Conceptual model

The agent-based model (ABM) will be described using the standard ODD (Overview, Design Concepts and Details) protocol for ABMs (Grimm et al., 2010, Polhill et al., 2008). In this structure, the "Overview" part (2.1) will discuss the purpose of the model (2.1.1), the most important components (2.1.2) and the main processes and their scheduling (2.1.3). The "Design Concepts" part (2.2) reflects on the design concepts underlying the ABM. In the "Details" part of the protocol (2.3), the three steps leading towards the implementation are discussed, starting with the initialization (2.3.1), input data needed for the model (2.3.2) and the sub-models (2.3.3).

Overview

Purpose
In order to test the hypothesis of cholera diffusion via runoff from dumpsites for the urban area of Kumasi we developed a model that allows us to study diffusion and persistence by means of two mechanisms: non-hyperinfectious

V. cholerae transmission (environment to human, EH) and hyperinfectious *V. cholerae* transmission via runoff from dumpsites (human to environment to human, HEH). We assume that non-hyperinfectious bacteria are already present in the environment at the beginning of the simulation. When faecal waste from infected individuals is dumped on open refuse dumps, rain can carry these bacteria from the dumpsites to the river, temporarily infecting this river with hyperinfectious bacteria. When exposed to this water, this can lead to new (HEH) infections (see Figure 5-1). When we refer to HEH transmission henceforth, this automatically includes EH transmission.

The model builds onto existing non-spatial mathematical models for cholera dynamics that model environmental reservoirs of *V. cholerae* (Codeço, 2001, Capasso and Paveri-Fontana, 1979). However, where they model a concentration of *V. cholerae* in water based on the number of infected individuals and the time that *V. cholerae* remains infectious, we model the infectiousness of the water based on the number of disposals of faecal waste on dumpsites and the time the dumpsite remains infectious. Modelling the exact concentration would require knowledge about the exact number of bacteria and volume of water, at any location and every moment in time for the study area. With the available data this is not possible.

The aim of this research is not to fully describe the 2005 cholera outbreak in Kumasi but to determine if, based on spatial-temporal patterns, the importance of the runoff from the dumpsites in the diffusion of cholera can be confirmed. The model is fully geographically explicit, using actual locations of buildings, dumpsites etc. and recorded rainfall.

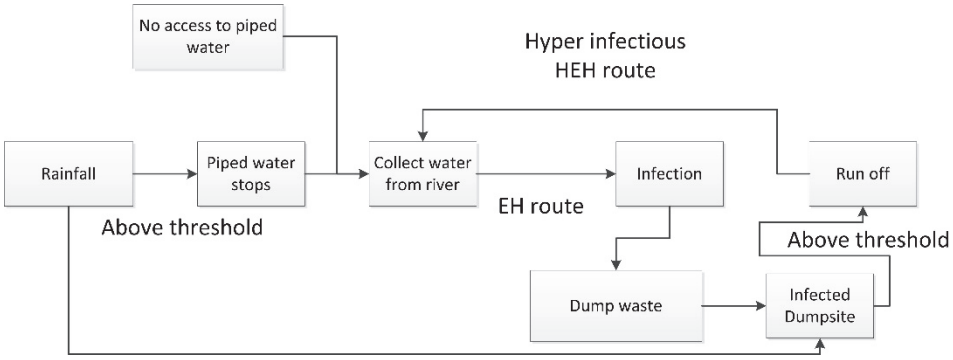

Figure 5-1 Overview of the processes included in the model

Entities, state variables and scales

Agent-based simulation models are built upon the concept of agents (entities) that are located within an environment, and interact with each other as well as with the environment. Agents can represent both human beings as well as

more abstract phenomena like organisations and natural features. By definition, agents are heterogeneous, differentiated by their variables.

The cholera model contains three types of agents: *households*, *individuals* and *rain particles* (Table 5-1). Households are collections of individuals, however, households also have specific variables and behaviour. Examples of variables that belong to a household are income level, access to tap water and house location.

Individuals are members of a household and inherit the properties from the households, but also have a number of personal characteristics like age and gender. Important state variable is the health status of the individual – susceptible, infected or recovered.

Rain particles are considered as agents, as they have the behaviour to move over the surface to the surface water. They can become infected (carry *V. Cholerae*) when their pattern of flow runs via an infected dumpsite. Examples of rain particle characteristics are the volume and travel time. The state variable of the rain particle is the infection level.

In the simulation, agents interact with a number of environments. The rain particle movement is based on a digital elevation model (DEM) for which a flow direction and flow accumulation environment are determined. These environments are static for the duration of the simulation, and are all raster layers.

Table 5-1 Overview of entities included in the model

Household agent: Community scale	
Variable name	**Brief description**
Coordinates	X and Y coordinates of the house ,
WaterFetching Point	Location on river where water is collected
Income Level	Influences type of water
Hygiene Level	Influences chance of infection
Access to tap water	Determines type of water used
Individual agent: Household scale	
Variable name	**Brief description**
HouseholdID	Determines household of individual
Gender	Used for composition of realistic households
Bloodtype	Influences chance of infection
Age	Used for composition of realistic households
HealthStatus	State variable: susceptible, infected, recovered
Rainparticle agent: Catchment scale	
Variable name	**Brief description**
Coordinates	X, Y coordinates of rainparticle
InfectionLevel	State variable, indicating the infection level
Volume	Used to control the amount of rain
Travel time	Travel time less than total time needed to travel a cell is stored and added to the travel time during the next time step
Environments: Catchment Scale	
Variable name	**Brief description**
Digital Elevation Model (DEM)	Rain particles flow over this surface
Flow direction Layer	Used to determine direction of flow
Flow accumulation Layer	Used to calculate accumulated flow
Dumpsites	State Variable is the dumpsite-infection-level. Infection can decay over time. Used by household to dump waste and by rain particle to get infected.
River	Households use the river water collection points. Rain particles flow until they reach the river and transfer infection.
Communities	Used to aggregate the disease cases
Houses	Represent locations of households
IncomeLevel	Used to determine type of water used

Besides the environments used in the hydrological sub-model, there are two environments related to the household and individual agents: dumpsites and houses. Dumpsites represent actual locations of open refuse dumps and have a dumpsite-infection-level as their state variable. This variable is dynamic during the simulation, as dumpsites can become infected after dumping of infected waste. Dumpsite infection will disappear over time due to loss of

hyperinfectiousness of *V. cholera* when no new re-infection occurs. Houses are used to locate the households.

All spatial layers have a resolution of 30 by 30 meters. The model runs for 90 days (duration of the cholera outbreak in Kumasi), divided in time steps of one hour. This temporal resolution is adequate to represent the flow of water over the DEM and allows for scheduling several human activities in a sequential order.

Process overview and scheduling
The model contains the following processes: *flow of rain particles* (over the DEM to the river), *households fetching water* and the *households dumping faecal materials* on the dumpsites.

The flow of rain particles over the surface will be described in detail in the section on the hydrological sub-model (2.3.3). Rainfall triggers the flow process. Rain particles can get infected by running over an infected dumpsite. Non-infected rain particles are removed from further simulation to save computation time. When reaching the river this will lead to hyperinfectious bacteria present in the river which after exposure (households fetch contaminated water) can lead to infection.

Use of water is determined by two household variables: access to tap water and income level. Households with access to tap water will use this water for their daily activities and this water is assumed to be safe for consumption. In case a household has no access to tap water, income level will determine if the household will buy safe water or use river water. This specific group of households will *fetch water* at the beginning of every new day. River water is collected by the household at the river point closest to the house and all individuals belonging to this household will consume this water. River water can be either infected via HEH or EH. Depending on the hygiene level of the household, the infected water can be treated (cooked) or not. Consumption of infected water can lead to infection.

All households that have infected individuals will *dump waste* on the nearest dumpsite. Dumpsites have an infection level that is determined by the number of infected households dumping on this site. Dumpsites become infected, and able to infect runoff, when the dumpsite infection level exceeds a threshold value.

Activities are scheduled in the following order: flow of rain particles, fetch water, dump waste. Flow of water occurs in every time step, fetching water and dumping waste only once a day. On the days it is raining, the rain falls at the beginning of the day.

Design concepts

Emergence
The route of cholera diffusion via runoff from dumpsites into the river can lead to a fast spread of infection to new communities downstream. This should be clearly visible in the epidemic curve, showing an early peak approximately two weeks after onset (Hartley et al., 2006). As model runoff is linked to heavy rainfall, this means that the HEH route will only exist when it is raining. Even when new cases of infection occur, and infected waste is dumped on the dumpsites, when it does not rain and the transport mechanism toward the river is missing, river water will remain free of hyperinfectious bacteria.

HEH infected water disappears from the area relatively fast (the water leaves the area within 6.5 to 13 hours). When combining both EH and HEH transmission, we expect that EH transmission determines the onset but HEH soon takes over. Both transmission mechanisms will decline over time due to less susceptible individuals in later phases (Hartley et al., 2006).

The epidemic curve for non-hyperinfectious transmission (without HEH component) builds up infection slowly and maintain itself longer at a more constant rate. According to literature the epidemic curve is flat, and can peak as late as 25 weeks after the first individual is infected (Hartley et al., 2006).

Spatial variations in infections are expected to occur due to differences in number and locations of dumpsites, upstream versus downstream location of communities in comparison to the source of infection and distribution of income classes per community. The location of dumpsites (in relation to the location of the river) can potentially make a large difference in the diffusion process. Also characteristics of the community itself may influence the diffusion pattern. When a community has a high number of higher income households, the river water will not be used. Consequently, the community will remain "safe" and the infection level of the dumpsite will not increase.

Adaptation
In this model the rainparticle agent adapts its direction and speed of movement to the steepness of the terrain and updates its infection level when the path of movement intersects with an infected dumpsite. The individual and household agents do not adapt their behaviour to changing conditions, however, this could be included. Based on awareness of the infection risk, households could change the location where they collect water, increase the hygiene level or buy bottled water.

Objectives
We assume all individuals and households have the objective to stay healthy (not to get infected by cholera). Assuming that water is necessary for every day live, this may lead to risk avoidance in the type of water a household uses. In this model we make the assumption that households that can afford to buy safe water will do so. The choice to use river water is driven by income level.

Learning and Prediction
This model does not include any prediction and does not use any Artificial Intelligence Learning algorithms.

Sensing
Household agents are aware of the location of the river (water fetching points) and the nearest dumpsites. Agents do not sense the level of infection in their community. This means that higher levels of infection do not trigger behaviour change.

Interaction
Rainparticles interact with the dumpsites (can get infected) and households interact with the river (water fetching points) and the dumpsites (dump waste). Indirectly there is interaction between the individuals belonging to a household as they for example share the same water.

Stochasticity
The largest stochastic element in this model is the synthetic population. The process of generating the synthetic population is explained in detail in section 2.3.1. Every time a new population is generated, this will lead to a different collection of households and individuals, living in different spatial configurations. During testing and running of the model, results are always based on a number of iterations that include several different synthetic populations.

Besides the randomness included in the distribution of the population, there are also other random elements in the model that influence the results. As infection is based on a probability, a different number of individuals can get infected even when the population remains the same. In each run of the model, we conduct 10 replicate runs before creating a new synthetic population to account for this variability.

Collectives
A household in this model can be regarded as an intermediate level of organisation. Households are collections of individuals living in the same house. All individuals within a household have their individual variables but are also

assumed to share water. This does not mean that all household members will become ill at the same time, as infection is dependent on individual characteristics.

Observations
At a global level we observe the number of susceptible, infected and recovered individuals differentiated by the source of infection (EH or HEH) over simulation time. This allows us to create an epidemic curve for every model run. Information includes the community the infected individual belongs to allowing us to aggregate the number of infections and recovered individuals per community.

Details

Initialisation
The initialisation of this model can be split into two parts: the loading of the input data (environments and input variables) that will be discussed in 2.3.2 and the generation of the synthetic population (discussed below).

Synthetic population of agents

The study area consists of a water catchment area and does not coincide with administrative boundaries. The area contains 21 communities, of which some communities are completely inside the study area and others only partially (see Figure 5-2). The population per community was obtained from Osei (2010). The average number of individuals per household used for the model was 3.9 (GSS, 2008). The number of households (67,000) in the study area was calculated based on the community population and average household size. The conducted experiments are based on 8,500 households (12.7% of the actual number) amounting to approximately 33,800 individuals. Because the rain particles are also agents, the combined number of agents in this model will otherwise exceed the capacity of the modelling software (Netlogo).

The synthetic population is generated based on statistical information (aggregated to the community level) obtained from Ghana Statistical Service (GSS, 2012). Composition of the synthetic population is based on the Monte-Carlo sampling method proposed by Moeckel et al. (2003). In this method, first older people (head of households) are assigned to a household, followed by additional household members. This leads to a natural order of sampling, in which the features of the individuals and households are sampled in the order in which they influence each other. Generating the synthetic population is done by the following steps: 1. Creation of the households with household attributes; 2. Creation of a set of individuals with assigned characteristics; 3. Selection of the head of household (an individual created in step 2) and assigning this

individual to one of the households generated in step 1; and 4. Randomly adding additional household members (from the individuals not selected as head of household).

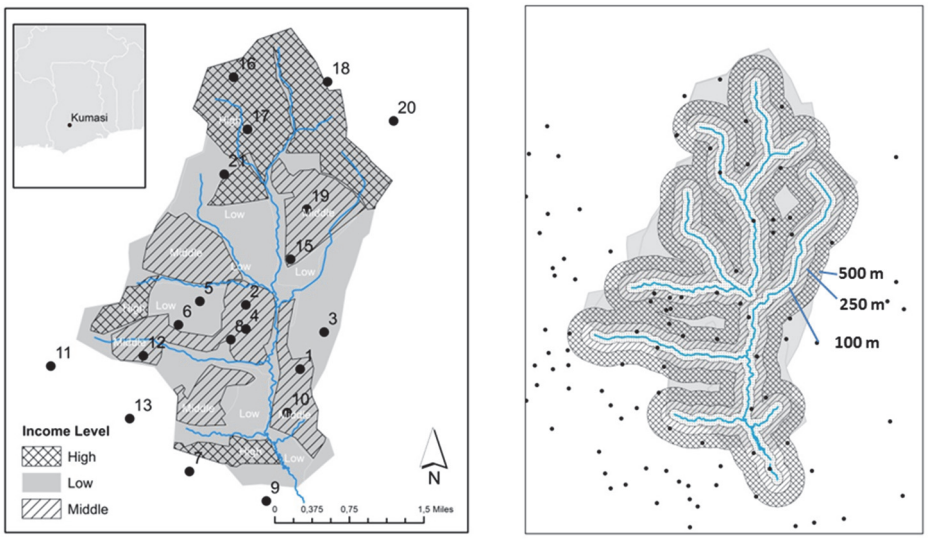

Figure 5-2 Study Area.Left Income levels for the catchment area included in the simulation model. The dots indicate the centre of the communities with labels indicating the community numbers. Right Locations of the dumpsites shown as black dots, and distance zones of 100, 250 and 500 m from the river.

Household variables are *hygiene level*, *income level*, a*ccess to tap water* and h*ousehold location*. Hygiene level consists of three classes (low, medium, high). The size of each of these groups was determined during the calibration phase (section 3.2), leading to a distribution of 19% (low), 52% (medium) and 29% (high), respectively. This is in line with other sources indicating a poverty line of 28% for urbanised areas outside Accra in 1992 (Devas and Korboe, 2000) and 19% of urban poverty in 1998-1999 (DSDC, 2005). Data on access to tap water was derived from national statistical information from the Ghana Statistical Service (GSS, 2012). As 86% of the households have access to safe water, this leads to the following division of access to safe tap water over income groups: 100% for high incomes, 88% for middle incomes and 78% for low incomes. Creation of the household variables also includes the selection of a house location.

The spatial distribution of the households is performed via a two-step procedure. First, the correct number of households are assigned to each community. Next, the household is assigned to a particular house, ensuring a match between the income level of the household and the income level of the house. In order to do this, a polygon income level layer (See Figure 5-2) is

used to assign an income level to each house. This procedure leads to densely and sparsely populated areas.

As the temporal duration of this simulation is only 90 days, the population is static during the simulation process.

Input data
Besides the synthetic population (2.3.1) and the input variables listed in Table 5-2, we use a DEM which was downloaded from CGIAR website (2010) as a Geotiff image. Flow direction and flow accumulation layers have been calculated based on this DEM using ArcGIS. Houses were digitised based on a Google Earth image of the area of 2006 and refuse dump locations have been collected using GPS. Water fetching points were selected based on flow accumulation greater than the threshold value (T_{wp}). The rainfall data was obtained from the Tutiempo Network SL, providing daily recorded rainfall data from September 30 to November 30, 2005. As no hourly rainfall information is available, the assumption was made that the duration of rainfall is two hours per day. The income levels (Figure 5-2) are polygons digitized based on expert knowledge in combination with building characteristics.

Sub-models
Disease sub-model

The probability of a household using contaminated water is influenced by the likelihood the household has to fetch water from the river. The likelihood that the household has to get water is influenced by two factors: income level (higher incomes are supposed to buy water) and rainfall. During heavy rainfall there may be no tap water due to power shortage in Kumasi. In the model this is accounted for by introducing a threshold rainfall (P_{max}) beyond which tap water stops working and more households will be forced to use river water.

Table 5-2Values of variables related to the synthetic population, hydrological sub-model and the disease sub-model used for calibration of the complete model (Min, Max) and the selected value (Value).

Parameters	Symbol	Min	Max	Value	Source
Switch point low to middle Hygiene level (%)	H_{lm}	0	50	45	calibration
Switch point middle to high Hygiene level (%)	H_{mh}	50	100	71	calibration
Number of rain participles on dumpsite	#R	1	8	5	calibration
Threshold dumpsite infection level **no decay**	D_{max}	1	500	22	calibration
Threshold dumpsite infection level **decay**	D_{max}			3	assumed
Probability fetching water (%)	P_{fw}	0	20	7.2	calibration
Probability fetching EH contaminated water (%)	P_{feh}	1	6	3	See Figure 5-4
Dumpsite decay function ($hour^{-1}$)	D_{decay}			20	Estimated from Codeço (2001)
Probability of infection drinking water (%)	P_{ew}	0	30	5.4	calibration
Threshold value for water points	T_{wp}	517	5000	585	calibration
Threshold heavy rain (mm/day)	P_{max}	1	19	2.3	calibration
Duration illness (days)				10	(Grad et al., 2012) list 2.9 to 14 days

There are two types of contaminated water (EH and HEH). The probability of fetching EH contaminated water, P_{feh}, is set to 3% (see section 3.2).

The probability of fetching HEH contaminated water depends on the runoff from infected dump sites and the travel time and varies in space and time. It is assumed that after dumping infected waste, the infection level, D, of the dumpsite will increase by 1. Dumpsites will start to induce runoff particles with hyperinfectious bacteria when the dumpsite infection level is above the threshold, D_{max}. Literature indicates that *V. cholerae* populations have an extinction rate ranging from 0.02 days to > 3 days (Feachem et al., 1983, Codeço, 2001). Conditions on dumpsites may vary in relation to temperature and humidity. Two options are built into the model: dumpsite infection level without and with a decay function. In case of no decay, once infected, the dumpsite will remain infected until the end of the simulation. For the option with dumpsite infection decay:

$$D_{t+1} = D_t - (\frac{D_t}{D_{decay}}) \quad (5.1)$$

in which D_{decay} is a decay constant (>1), and t is the time step (hour). Infection can take place when individuals that are susceptible drink water in a household that has fetched contaminated water. The probability for an individual to get infected by drinking contaminated water (P) is calculated as:

$$P = (P_{ch} + P_{ew}) \quad (5.2)$$

In which P_{ch} is the probability of infection based on household and individual characteristics and P_{ew} is the probability of getting infected due to drinking contaminated water independent of the characteristics. One of the individual characteristics that determines the probability of infection is blood type, assuming a higher risk for individuals with blood type O (Holmner et al., 2010). Relevant household characteristics are hygiene level and income level (see Table 5-3).

When simulating a river system covering a larger area, river water will remain in the study area long enough for the transition of a hyperinfectious state to lower infectiousness. However, this is not the case for areas the size of our case study area making the transition mechanism redundant. We found that water leaves the area within a time frame of 6.5 – 13 hours. This is faster than the hyperinfectious state persistence of up to 24 hours documented by Harris et al. (2012).

Hydrological sub-model

The hydrological sub-model simulates the downstream transport of hyperinfectious bacteria from dumpsites. In order to simulate surface water flow, a DEM, flow direction layer and a flow accumulation layer are loaded into the model. Water is simulated as rain particles (agents) flowing over the DEM surface. The amount of surface runoff is based on actual rainfall data for the case study area. The volume of water per rain particle is calculated by dividing the total volume of rain by the number of rain particles. The flow direction layer is used to determine to which neighbouring cell a rain particle will flow using the steepest downhill slope.

Table 5-3 Relationship between the household and individual characteristics (hygiene level, income level and blood type) and the probability of infection (P_{ch}).

Hygiene Level	Income Level	Blood type	P_{ch}
Low	Low/middle	O	30
Low	Low/middle	Other	15
Average	Low	O	10
Low	High	O	10
Low	High	Other	5

In this model the travel time of rain particles, T, is calculated using the general Manning formulas. Three different types of flow are used: *sheet flow (sf)*, *gully flow (gf)* and *river flow (rf)***,** corresponding to different travel times. We determine which type of flow applies by using the flow accumulation. For sheet flow (patches with flow accumulation < S_{gf}) travel time is calculated by (SCS, 1986):

$$T = 184 \frac{(F_L m_{sf})^{0.8}}{P^{0.5} S^{0.4}} \quad (5.3)$$

where m_{sf} is the Manning coefficient for sheet flow ($m^{-0.375} \cdot hour^{1.25}$), F_L is the flow length (m), P is the rainfall (m) and S is the slope in (m/m). For gully flow (patches with flow accumulation ≥ S_{gf} and <S_{rf}) and river flow (flow accumulation ≥ S_{rf}) the Manning formula for channel flow will be used (Shaw et al., 2011):

$$T = \frac{F_L m_c}{R^{2/3} S^{1/2}} \quad (5.4)$$

in which R is the hydraulic radius (m), i.e. the cross sectional flow area divided by the wetted perimeter and m_c is the Manning coefficient for open river channel flow ($m^{-⅓} \cdot hour$). F_L, m_{sf}, ($R^{2/3}/m_c$) for gully and river, S_{gf} and S_{rf} and S are determined in the calibration process (Table 5-4).

In case rain particles travel only part of a cell (during a particular time step) this is accounted for by storing the remaining travel distance in memory and adding this travel distance to the distance travelled during the next time step.

Model output

In many models the basic reproduction number (R_0) is taken as measure to quantify the epidemic. However, R_0 is known to be very sensitive to input parameters of a model (Grad et al., 2012) and some of the parameters used are also not obvious in our model. We also think that the spatial patterns should be considered as our model is spatially heterogeneous.

The model provides output on the level of the individual (e.g. time step in which individual was infected and recovered, including the type of infection, EH or HEH), of the community (epidemic curve per community), of the dumpsite (infection level at a certain time step) and of the total population (epidemic curve). The data available to compare the simulation results to are the number of cholera patients per community registered by the Disease Control Unit (DCU).

The model performance will be based on the relative diagnosed disease cases per community in percentage:

$$GD = \frac{Number\ of\ cases\ in\ community}{Total\ number\ of\ cases} * 100 \tag{5.5}$$

Model performance will be determined by comparing the simulated percentage of disease cases in each community (GD_s) to the percentage of diagnosed disease cases in the 2005 outbreak (GD_d). The accuracy of the simulations will be expressed as r^2 (with $r^2 = 1$ a perfect simulation), where r is the Pearson's correlation coefficient defined as:

$$r = \frac{\sum_{i=1}^{n}\left((GD_{s,i} - \overline{GD_s})(GD_{d,i} - \overline{GD_d})\right)}{\sqrt{\sum_{i=1}^{n}(GD_{s,i} - \overline{GD_s})^2 \sum_{i=1}^{n}(GD_{d,i} - \overline{GD_d})^2}} \tag{5.6}$$

where i refers to community i and n is the number of communities. The hydrological model requires the study area to be a catchment area. As a consequence, delineation of the case study area does not coincide with administrative boundaries. In this area we have 21 communities of which 11 are completely within the study area and 10 only partially. The value of r^2 is calculated for the 11 communities completely inside and for the total 21 communities (r^{2*}).

The spatial element is contained in the fact that for every community the simulated percentage of cases is compared to the observed number of cases in that community. By doing so, the overall spatial pattern can be evaluated.

In addition we evaluate the time of first infection and duration of infection per community.

5.3 Model implementation

Case study

The model is implemented in Netlogo version 5.05 and initialized with a subarea of the city of Kumasi, Ghana. Kumasi is a metropolis located in the

Ashanti Region. It is a fast growing regional capital with a population in 2013 of 2,069,350 people (World-gazetteer.com). Kumasi consists of a number of communities with no exact geographic boundaries. Disease cases are registered per community.

Living and housing conditions in many of the communities are overcrowded (Whittington et al., 1993). This is mainly because of rapid urban population growth. Poor housing conditions reflect high probability of low incomes. The city of Kumasi experiences different raining seasons, a longer raining season from March through July and a shorter raining season from September to November. The cholera epidemic of 2005 started during the short raining season and covers a period from September to December. During this period, water shortages occurred due to the low power voltage (DNA, 2010). During these water shortages, most households used surface water from rivers and streams for drinking, cooking, and other household activities (Osei et al., 2010).

The study area does not include the complete Kumasi area but was restricted to one catchment in the centre of the city.

Model parameterisation and calibration

Calibration has been conducted in two steps: first the hydrological sub-model was calibrated, followed by a calibration on the complete model. After the calibration a stability check was performed.

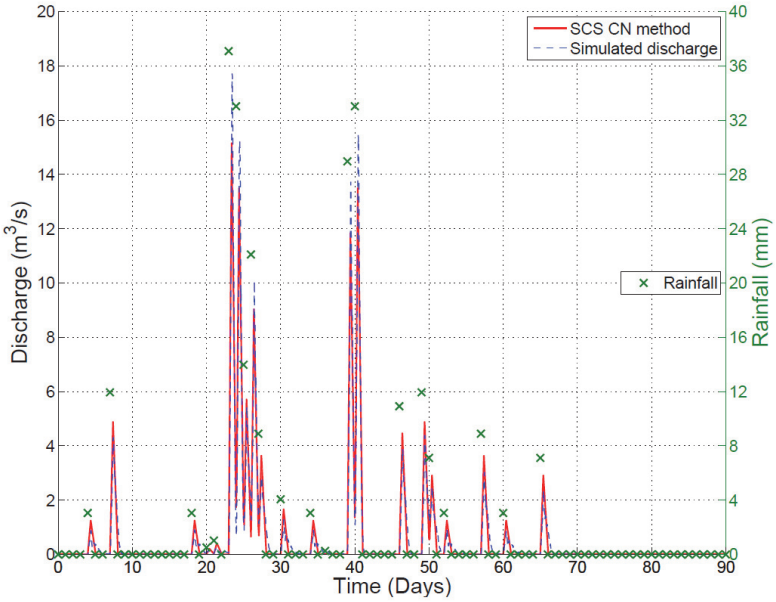

Figure 5-3 Discharge of the simulation model (simulated discharge) compared to the discharge calculated by the CN method for the duration of the simulation. Rainfall is included as green crosses.

For the calibration of the hydrological sub-model the simulated discharge was compared to the discharge calculated by the Curve Number method (CN) (Boonstra, 1994). The CN method is widely used (Arnold et al., 1998, Boonstra, 1994) to simulate surface runoff. The range of parameters used in this calibration are shown in Table 5-4. After calibration of the parameters in the manning equations, the agent-based hydrological model showed good agreement with the discharge generated by the Curve Number method (see Figure 5-3). The simulated average velocity for river flow was 0.65 m/s and for gully flow 0.35 m/s, which are in the correct order of magnitude (Riscassi and Schaffranek, 2003). A detailed description of the calibration goes beyond the objective of the paper. For more information see Doldersum (2013).

The dataset used for the calibration of the complete model consisted of cholera cases for the 2005 epidemic. All cases were confirmed by bacteriological tests and were reported to the Disease Control Unit (DCU). Not all communities have a hospital but all have clinics or community volunteers.

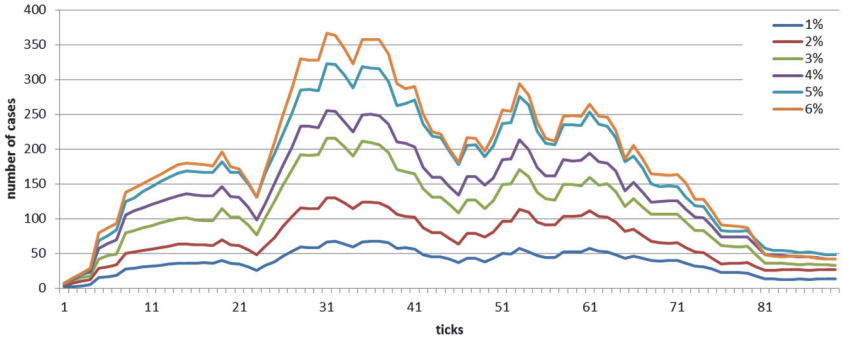

Figure 5-4 Calibration of the probability of fetching water contaminated with EH (P_{fch}) by varying the value between 1 and 6%

Calibration of the complete model was done globally using a Monte Carlo simulation after selecting ranges for all parameters. Every parameter was varied randomly within the range, after which r^2 was calculated. The range of values was narrowed until the r^2 showed no further improvement. The final parameter values after calibration, including the ranges of values used during the calibration are shown in Table 5-2.

The value for the probability of fetching EH contaminated water (P_{feh}) was determined by running the model with a range of input values (1– 6%) and selecting the lowest value at which spontaneous infection occurred and rainfall influence due to lack of tap water was visible. Contrary to other models, this value is constant during the simulation. Results are shown in Figure 5-4.

The stability check was conducted by running the model 250 times, creating a new population every 10 runs, and checking when values became stable. The check was conducted for EH only, for HEH (including EH) without dumpsite decay and with dumpsite decay. The model became stable after approximately 100 runs (see Figure 5-5).

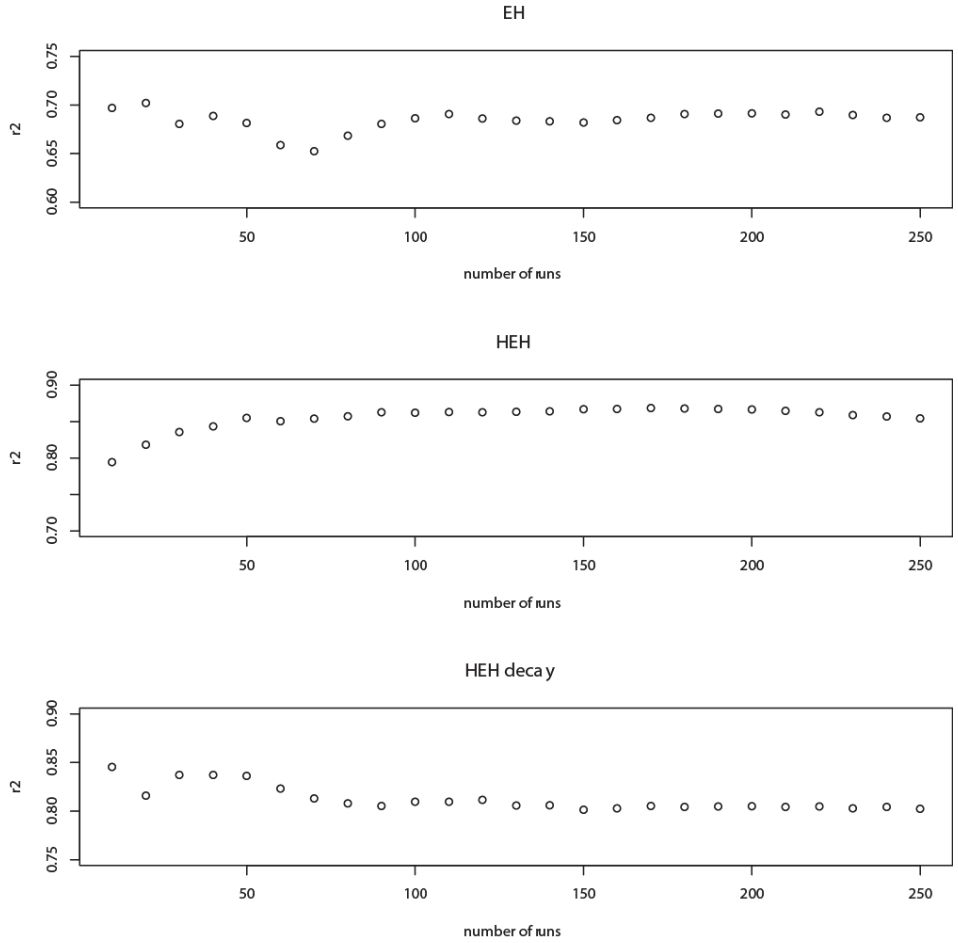

Figure 5-5 Stability check for EH, HEH (without decay) and HEH with decay, showing the r^2 calculated over an increasing number of simulation runs (10-250).

5.4 Results

Two sets of experiments are conducted to test the hypothesis of dumpsites playing a role in the diffusion of cholera during the 2005 outbreak in Kumasi. In the first experiment the model is tested under three conditions: EH transmission (section 4.1), HEH transmission without decay and HEH transmission with decay of the infection level of the dumpsites (4.2).

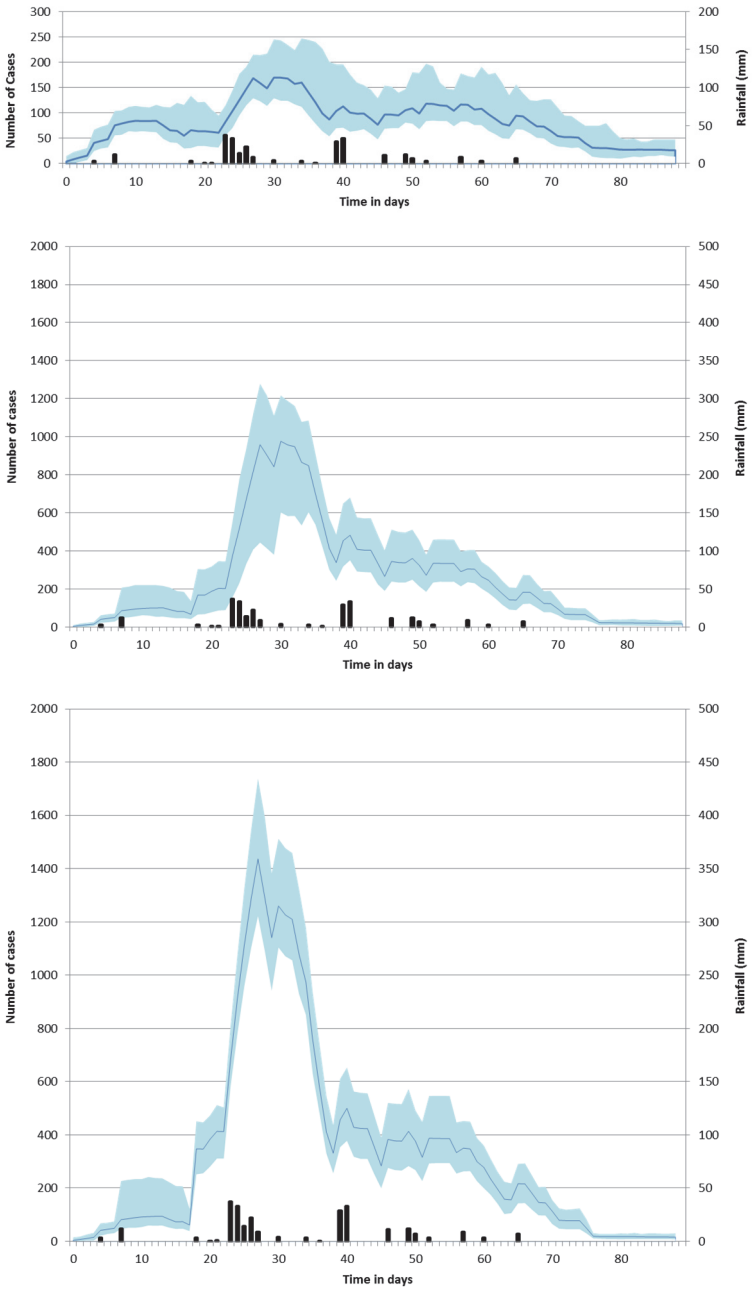

Figure 5-6 Epidemic curves for EH (top), for HEH (middle) and for HEH with decay of dumpsite infection level (bottom). Blue line showing the mean epidemic curve (100 runs) with, in light blue, the variation between the runs. Black bars indicate the days and intensity of the rainfall. Vertical scale of EH differs from scale of HEH and HEH decay.

In the second experiment, we test the assumption that it is possible that only part of the dumpsites actively contributed to the diffusion process (4.3). We do this by selecting only dumpsites that are within a certain distance (100, 250 and 500 meters) from the river for the HEH runs. Distances were selected based on findings of Osei and Duker (2008), indicating that 500 meters is the range within which a spatial dependency exists between cholera and refuse dumps.

All experiments are influenced by the synthetic population; therefore each experiment is repeated 100 times, generating a new synthetic population for every 10 runs.

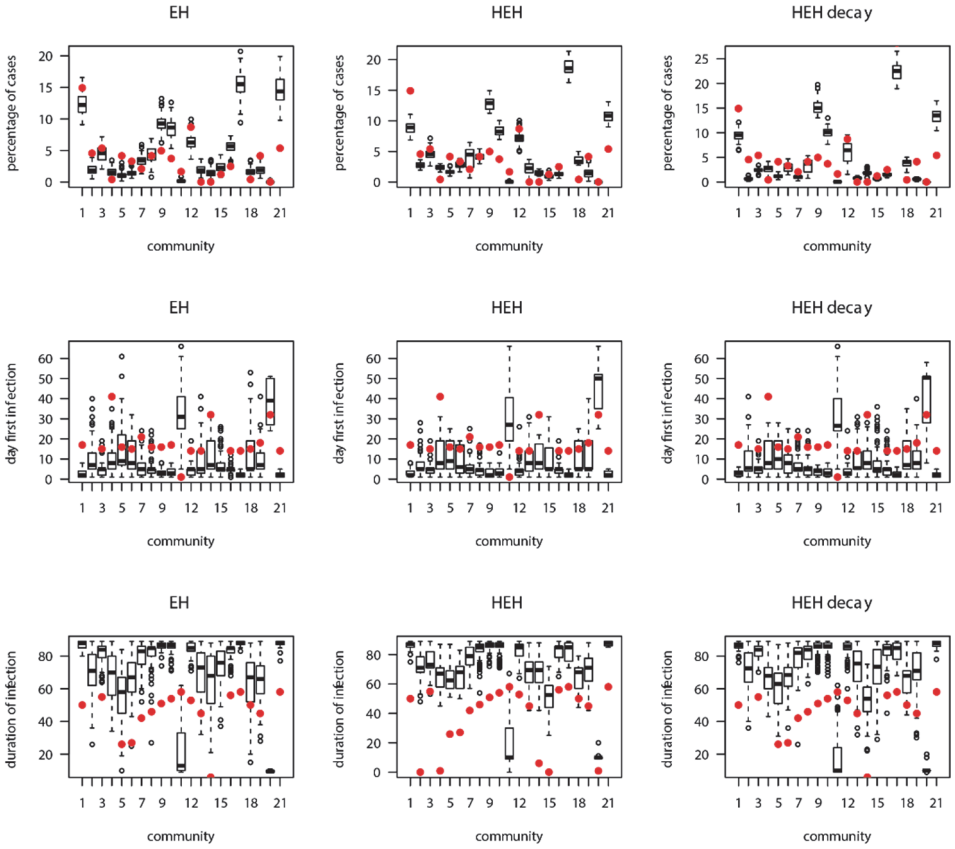

Figure 5-7 Boxplots for EH (left), HEEH (middle) and HEH with decay (right) showing the percentage of cases (top), the day of first infection (middle) and the duration of infection (bottom) per community. Red dots indicate the values during the 2005 epidemic.

EH transmission

The EH transmission experiment benchmarks the situation without hyperinfectious *V. cholerae* diffusion. In this case, infection only takes place via natural *V. cholerae* and is independent of the number of infected individuals. However, the number of people that are exposed to contaminated water will vary, as the number of people dependent on river water varies as a result of tap water availability during the simulation.

Results from a global perspective are included in Table 5-5 and the epidemic curve is shown in figure 5-6. The r^2 for this experiment is 0.69 indicating some resemblance to the spatial pattern in the 2005 dataset. As the probability of getting infected is the same throughout the study area, this resemblance can only be explained by the distribution of the population in space.

The epidemic curve (Figure5-6) shows a weak peak after 33 days (approximately 5 weeks) and infection persists throughout the simulated period with a decline to a minimal number of infections after 80 days (approximately 11 weeks), resembling a more endemic situation, similar to results obtained by Hartley et al. (2006). The epidemic curve shows a weak response to rainfall which can be attributed to a higher number of households being exposed, due to malfunctioning of tap water.

When evaluating the boxplots of the individual communities (Figure 5-7) we see that for a number of communities (1, 2, 5, 12, 19) the relative number of cases is underestimated. These communities are located both in the upstream and downstream parts of the study area and are all completely within the study area. For these communities, hyperinfectiousness may have played a role. Overestimation takes place in communities 9, 10, 16 and 21. Where communities 9 and 10 are located downstream, communities 16 and 21 are located in the upstream area. Communities 9, 10 and 16 are only partially within the study area.

The timing of the first infection and duration of infection in the simulated results show an earlier onset of the epidemic (faster infection) compared to the empirical data. The simulated duration of infection is longer compared to the empirical data for the communities of the study area.

HEH transmission

In the HEH experiment (without decay) the r^2 increases to 0.86 (Table 5-5), showing a closer resemblance of the spatial pattern found in the empirical data compared to EH. The epidemic curve shows a clear response to the rainfall (peak with onset on day 23) as can be seen from Figure 5-6.

The pattern at the level of the communities (Figure 5-7) shows that we have an underestimation in the communities 1, 2 and 5 and an overestimation in 9, 10 and 21. These are the same communities showing over/underestimation in the EH experiment. Patterns for onset of infection and duration of infection show earlier first infection and longer duration compared to the empirical data.

For the experiment with a decay function, the r^2 is 0.81 which is slightly lower than in the experiment without decay, with a small decline in the number of infected individuals from 2773 to 2208. The relative contribution of the EH transmission route is 22% in the experiment with decay compared to 15% in the experiment without decay. The epidemic curve of the simulation with decay differs from the experiment without decay in that the variability between the individual runs is much larger (see Figure 56). When we compare the boxplots of the communities (Figure 5-7) the first infection and duration of infection are similar except for a few small deviations.

Distance

In the previous experiments it was assumed that all dumpsites played a role in the diffusion process. However, it is likely that runoff from dumpsites that are located closer to the river will reach this river more often, compared to runoff from more distant dumpsites due to stagnation and infiltration of water. For both HEH transmission experiments (with and without decay) we have conducted runs in which we only included dumpsites within a certain distance from the river (100, 250 and 500 meters). For the results see Figure 5-8 and Table 5-5.

When including only dumpsites within 500 meters from the river, this leads to an r^2 value for HEH with and without decay of 0.80 and 0.87, respectively. These results are similar to including all dumpsites. This implies that dumpsites further than 500 meter from the river did not play a role in the runoff. However, as the study area is small, only three dumpsites are located outside this buffer distance (see figure 5-2). The observed effect can also be the result of a mismatch in time it takes runoff to reach the rivers and water fetch time. When this runoff time is long and the water fetching has already been completed for the particular day, this can result in a similar effect. When restricting the distance further to 250 and 100 meters, the r^2 value becomes lower. This is probably due to the fact that the number of dumpsites included drops considerably.

When we compare the r^2 (in which only the 11 communities that are completely within the study area are included to the results of r^{2*} in which all 21

communities are included, we see that the pattern remains the same although the values are consistently lower.

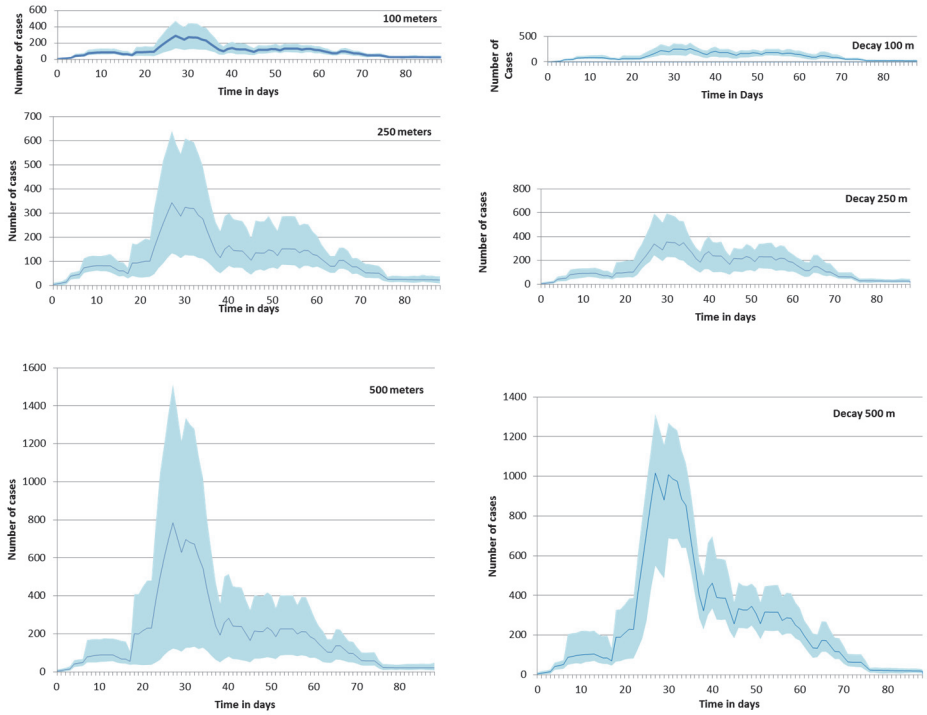

Figure 5-8 Epidemic curves for the distance experiments for HEH and on the right the experiments for HEH with decay. On top the results with only dumpsites within 100 m from the river, in the middle dumpsites within 250 m, and bottom dumpsites within 500 m.

5.5 Discussion

Even in the experiments with only EH, we see a reasonable r^2 of 0.69. This is unexpected given the simple structure of the EH model, with a constant infection probability, the omission of hyperinfectiousness and the fact that we minimized the P_{feh} value (determined via calibration). We believe this is caused by the structure of the population and in particular by the distribution of the population in space. While generating the synthetic population, we controlled the number of individuals in each income group and by doing so, we indirectly influenced the number of potentially exposed individuals, as higher incomes are assumed to have access to clean water and lower income groups resort to using river water more often. This was reflected in the spatial distribution of the households, meaning that we generate communities with a large percentage of higher income households, and communities with a large number of lower income households at the correct location. This confirms the

dependency of cholera transmission on the socio-economic structure of the area but also underlines the importance of spatially explicit models.

In the results we see higher r^2 values and more realistic epidemic curves for the situation with hyperinfectiousness compared to the EH experiments. This confirms that hyperinfectiousness played a role in the 2005 Kumasi outbreak. In the model we were able to model the hyperinfectiousness via runoff from infected dumpsites. The results show that dumpsites within 500 m from the river contributed to the diffusion (as we see no significant difference in the r^2 between the experiments including dumpsites within 500 m and all dumpsites), but also show that this must have been a fairly common diffusion route in which most dumpsites play a role (when we reduce the distance, and therefore the number of contributing dumpsites, the r^2 decreases).

The model we used in this research has a hydrological sub-model that requires the case study area to coincide with a hydrological catchment. In our situation, a part of the communities extended beyond the catchment boundary. As a result the comparison between the empirical data and simulated cholera incidences was more difficult, forcing us to calculate the r^2 including only the communities completely located within the catchment, and the r^{2*} including all communities. For the communities that are partially outside the study area, the recorded diagnostic data may refer to individuals living outside the catchment. This is reflected in lower values of the r^{2*} compared to the r^2 values. Because no exact administrative boundaries of the communities are known, correction for this issue was not possible.

In general, parameterization of cholera models is believed to be difficult because many important factors are unknown, like the exposure to contaminated water, the concentration of *V. cholerae* in the river water, the decay of the infectious vibrios and the infectious dose (Grad et al., 2012). For the 2005 Kumasi outbreak, no micro-biological data is available to make reliable estimates. We have therefore chosen to create a simplified model solely meant to test the runoff transmission route hypothesis and to analyse the results based on general spatial and temporal patterns.

The model itself contains some clear limitations which are mainly linked to the limited behaviour of the agents. Agents collect water only once a day at a fixed time step creating a sensitivity to the order in which the processes are scheduled (rain versus fetching of water).Movement of agents to other locations within the study area is currently not modelled. Most important limitation of the model is probably the fact that agents that collect river water, do so at a fixed water point (closest to their home) without evaluating the risk of using this water. Evaluation of risk can be based on media attention, knowledge about disease cases in their neighbourhood, or even spatial

intelligence (water upstream from dumpsite is cleaner). Extending this model with an artificial intelligence component that enables agents to evaluate their risk and change their behaviour could address these limitations.

5.6 Conclusions and recommendations

This work proposes an agent-based model (ABM) for micro-simulation of cholera diffusion, that incorporates an environmental reservoir of natural *V. cholerae* (EH mechanism) and diffusion of hyperinfectious *V. cholerae* via runoff from dumpsites. The proposed model is simple in its setup and can be extended by adding new elements like human movement and change of behaviour of individuals based on disease awareness (avoid using river water).

The model output showed a good correlation to the spatial-temporal pattern of the 2005 outbreak in Kumasi, Ghana. The correlation of the EH transmission depended mainly on the quality of the synthetic population and more precisely on the spatial distribution of the different income groups. Model output was improved by adding the HEH transmission, indicating that hyperinfectiousness played a role. This may have been caused by runoff of dumpsites within 500 meters from the river, although in this case, a large number of the dumpsites contributed to the infection.

ABMs seem to be a good choice for simulating the spread of a disease like cholera where both environmental processes and human behaviour are allies in the diffusion mechanism. ABMs are especially suitable for micro-simulation and this type of simulation can lead to new understanding of underlying diffusion mechanisms. A complete validation of an agent-based cholera model will be difficult to achieve, as this would mean that not only *V. cholerae* ecology and epidemiology but also human behaviour will have to be validated, and it is arguable if this will ever be possible. However, if specific spatial and spatial-temporal patterns can be found that are reproducible via pattern-oriented modelling, good results might be achieved.

This research shows the potential health effect of open dumpsites in Kumasi, but the problem of improper solid waste disposal is widespread in many other cities in developing countries. Besides relating to cholera, open dumpsites can also have other health effects. Results stress the importance of proper solid waste disposal, which could lead to relocating open dumpsites, conversion to covered landfills or improved waste collection systems. Solving the cholera problem requires equal attention to proper sanitation and access to clean water.

Chapter 6 Comparing Simulated and Empirical Pertussis Patterns using Self-Organizing Maps

6.1 Introduction

Infectious diseases form an important threat to human health and pose an economic and social burden on our society. Infectious diseases that were almost extinguished (e.g. measles and pertussis) are re-emerging because of resistance to vaccines and genetic changes in the pathogen reservoirs (de Melker et al., 2006). This re-emergence makes constant monitoring of disease outbreaks necessary. Fortunately, the use of better surveillance methods and the movement towards open science allow accessing long-term disease data at suitable resolutions for spatio-temporal analysis. This data can then be used to support various outbreak monitoring efforts.

In addition to more and better disease data, more simulation models are nowadays available to, for instance, test out interventions. To increase the validity and representativeness of these tests, we should ensure that simulation models are structurally realistic – i.e. that they capture all essential elements of the disease diffusion processes. The diffusion process of infectious diseases can be regarded as a complex system, influenced by spatial factors such as the distribution of the population and its mobility, and by non-spatial factors such as the replenishment of susceptible hosts after epidemics (Grenfell et al., 2004).

Several authors have recently studied the link between disease diffusion and mobility (Balcan et al., 2009, Tizzoni et al., 2014, Colizza et al., 2007, Belik et al., 2011), and a number of global scale models have been developed (Colizza et al., 2007, Balcan et al., 2010, Epstein et al., 2007). Global mobility is important as most pandemics are triggered by long distance human movement. Disease models that work at a regional scale and that consider work and school related commuting are much scarcer than global models. Yet, these regional models could guide local interventions and, for instance, support geographically targeted vaccination strategies.

Although mobility is now included in most disease diffusion models, very little attention is paid to age-specific mobility. Most modellers assume that mobility is the same for all age groups or that only adult commuting has a significant impact on the disease diffusion pattern. The only example found of the inclusion of age-specific mobility in a disease model was by Apolloni et al.

(2013). However, their results only indicate that when the expected number of secondary cases caused by one infected individual (R_0) is large enough, the effect of childhood mobility is negligible.

These days many disease models are spatial or geographically explicit (Belik et al., 2011, Dalziel et al., 2013, Colizza et al., 2007). Yet, very few studies validate the spatio-temporal diffusion of their disease models, and when they do so, they use simple means like the infection period per continent (Colizza et al., 2007). Other models simply link empirical data with the model output via the R_0. However, this metric has no spatial meaning and it only measures the explosiveness of the epidemic but not the order in which different spatial locations are infected or the fact that the disease maintains itself in some areas but not in others. Other validation metrics such as the global invasion threshold or the front velocity (Belik et al., 2011) are found in literature. However, it is hard to observe fronts at regional scales because this concept is only applicable to global diffusion processes.

The validation of simulated spatio-temporal diffusion patterns requires being able to find and characterize diffusion patterns in empirical data. This characterization can be done via a diffusion trajectory that describes the sequence of typical spatial states during the movement of a disease over a given geographical region. In time, disease diffusion shows tipping points, where the diffusion moves from an initial infection to an epidemic. Diffusion trajectories can then help to characterize these tipping points in space, e.g. to determine if infection in a certain area triggers diffusion to other regions. Little is known about similarity in the sequence of infection of geographic regions during different epidemics. Does the same diffusion pattern repeat itself during multiple epidemics? Infectious diseases are known to show both expansion and relocation diffusions often combined in a hierarchical diffusion pattern (Haggett, 1976), indicating the importance of large cities in the diffusion process.

This research using diffusion trajectories identified via a combination of Self-Organizing Maps (SOMs; Kohonen, 2001) and Sammon's projection to explore spatio-temporal diffusion patterns. SOMs are a type of unsupervised neural networks that can be used on large datasets, allow for missing observations, and can be applied in both space and time (Zurita-Milla, 2013, Kohonen, 2001, Augustijn and Zurita-Milla, 2013). In spatial-epidemiology, SOMs have been used to study multivariate patterns (Wang et al., 2011, Basara and Yuan, 2008, Koua and Kraak, 2004) but they also enable an integrated analysis of different spatial-temporal diffusion patterns. A method for this was developed by Augustijn and Zurita-Milla (2013), who compared spatio-temporal diffusion of measles in Iceland using empirical data. Here we further test and develop this method for the comparison of empirical and simulated diffusion patterns. More

precisely, in this research SOMs are used to detect similarities and differences in spatio-temporal diffusion patterns between epidemics and between empirical and simulated data while exploring the impact of different (age-specific) mobility patterns.

6.2 Data, model and methods

This section describes the empirical data used in this study, the model with which the simulated data was generated and the methods used to compare the empirical and simulated data.

Empirical pertussis data

Pertussis is a re-emerging disease in the Netherlands, despite high levels of vaccination. Since its re-emergence in 1996, the disease has become endemic showing periodic epidemics every 2 to 4 years. An empirical pertussis dataset for the Netherlands was obtained from the Netherlands National Institute for Public Health and the Environment (RIVM). This dataset consists of 18 years of daily surveillance data at the spatial aggregation level of the 4 digit postal code. The number of postal codes changes per year with an average number of around 4000 and an average size of 8.7 km^2. The area of postal codes varies considerably between urban and rural areas with smaller postal codes in cities and larger postal codes in rural areas.

For our analysis, the data was smoothed, spatially aggregated to percolation zones, z-normalized and epidemic periods were extracted. Percolation zones are areas with a high probability of being occupied, they are the densely populated areas. The percolation zones were generated based on a method described by Arcaute (2015). This method uses road network nodes, determines the distance between these nodes, and iteratively drops the nodes that are spaced furthest apart. In this way, areas with a high level of urbanization and connectivity can be identified.

For the Netherlands, the percolation method leads to 16 spatial areas (see Figure 6-1), each with a population of over 100,000 inhabitants and with high numbers of commuters (see Figure 6-3).

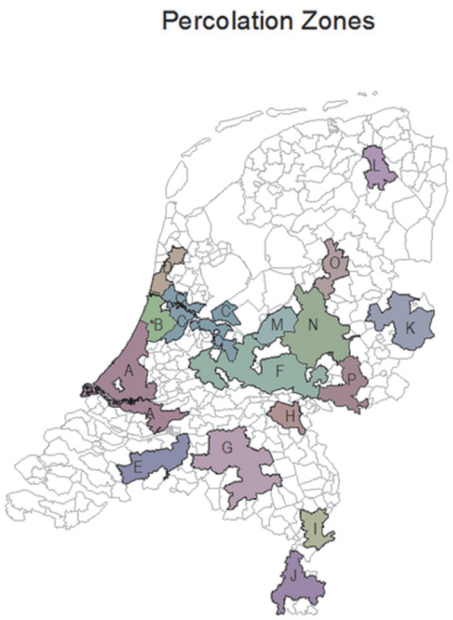

Figure 6-1 Percolation zones with Letters for the Netherlands

After aggregating the pertussis data to the zones, 6 epidemic periods with a duration of 55 weeks each were extracted to map their spatio-temporal diffusion patterns (Figure 6-2). The chosen duration correspond to the maximum duration found in the dataset but, logically, some epidemics diffuse faster and other diffuse more gradually. Hence the length of the actual epidemic varies. The dataset was organized in such a way that the rows represent time stamps (subsequence of 55 weeks) and the columns represent percolation zones (spatial locations).

The Model

Dutch pertussis simulated data was generated with a compartmental disease model combined with a population model and with a spatial matrix commuting model (see Tjalma (2016) and Figure 6-3). This disease model is also a SEIR model where the S, E, I, and R define the fraction of the population in each age-related group that are Susceptible, Exposed, Infected and Recovered. Individuals can move from S to E, from E to I, from I to R and from R back to S. The rate at which this happens is the rate of disease transmission (β).

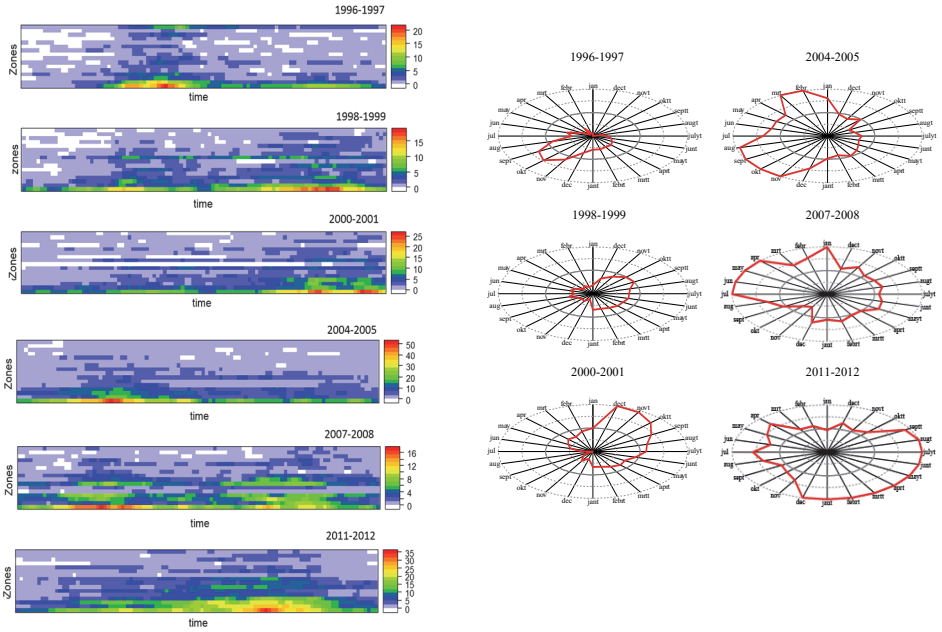

Figure 6-2 Disease frequency for time series subsequence ordered by population per zone from high (bottom) to lower (top). Disease incidence indicated as red line, time counterclockwise

The model is compartmental because it is age specific, and the number of contacts varies per age group. Variation in age specific contact rates are implemented via a WAIFW (Who Acquires Infection From Whom) matrix, based on the estimated contact pattern of the different age groups in the Netherlands as estimated by Mossong et al. (2008).

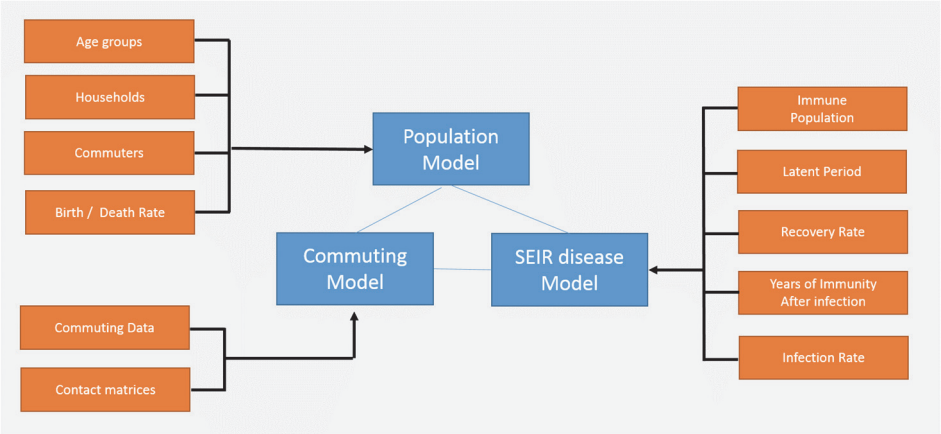

Figure 6-3 Overview model elements (adjusted from Tjalma (2016))

The rate of disease transmission (βij) between a susceptible individual in age group i with individuals of age group j can be defined as the product of the average number of contacts (γ_{ij}), the susceptibility (α_i), the infectivity (ε_{jk}), and the probability of disease transmission (P) (Del Valle et al., 2007). This can be formalized as follows:

$$\beta_{ij} = \gamma_{ij} * \alpha_i * \varepsilon_{jk} * P_{ij} \quad (6.1)$$

Where P is a function of the duration of the contact. By making the transmission rate time dependent (i.e. seasonal), we can generate complex dynamics (Earn, 2008). For the simulations, the population is positioned at the centroids of each Dutch municipality, giving a total of 396 population nodes. The population is divided into nine age groups, that coincide with vaccination regimes, school types and retirement age (Table 6-1). Besides age, population is also split into commuters and non-commuters. All commuters in the age group 12 to 17 years are school commuters and all the adult commuters are assumed to be work commuters. The age groups 0-5 months, 5 months to 5 years, 5 to 12 years, 17 to 25 years and 65 and older are assumed to be non-commuters. Commuting data at municipality level was obtained from the Dutch Central bureau of Statistics (CBS, 2013[4]). In this model, households are also explicitly modelled: children always belong to a household and adults are split over household and non-household groups (Table 6-1).

Table 6-1 Setup Population

Group	**Age**	**Units**	**Commuters**	**Household**
1	0-5	months	-	household
2	5-60	months	-	household
3	5-12	years	-	household
4	12-17	years	school	household
5	17-15	years	-	(non)household
6	25-35	years	work	(non)household
7	35-50	years	work	(non)household
8	50-65	years	work	(non)household
9	Older than 65	years	-	(non)household

The population model controls two processes, the population dynamics (birth, death) and the aging process (distribution over the age groups). Over time, agents change from younger to older age groups (e.g. $A_1 \rightarrow A_2$). The initial population per population node is based on actual population data per municipality for the year 2013. This data was acquired via the Dutch Central

[4] *https://www.cbs.nl/en-gb/news/2013/23/more-than-half-of-employees-commute-to-work*

Bureau of Statistics (CBS, 2014[5]). In the Netherlands, there are on average 470 births each day (CBS, 2014). We assume that this birth rate is valid for all the municipalities. We also assume that birth and death rate are equal. Municipalities with less than 7500 inhabitants were merged to avoid having many small and potentially less informative (more noisy) spatial units.

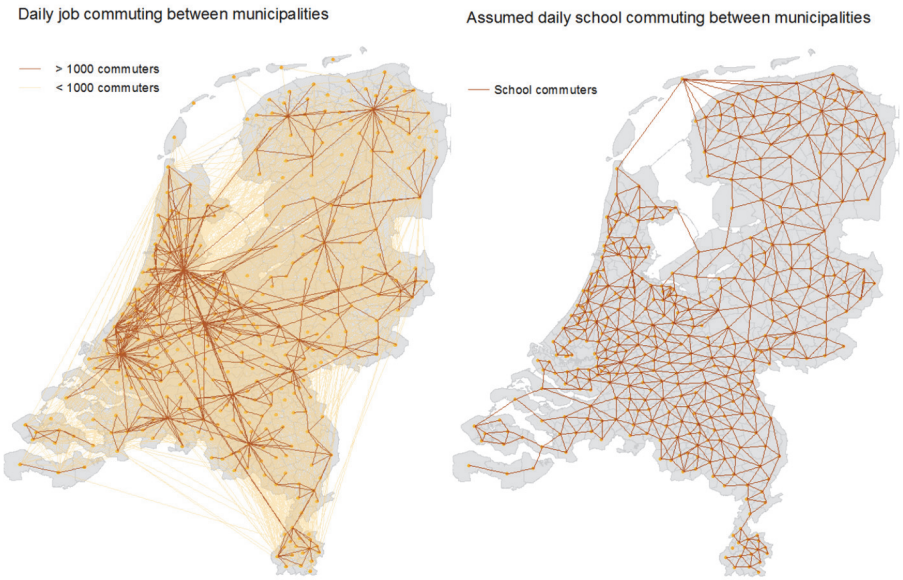

Figure 6-4 Commuting as implemented in the model (Tjalma 2016)

Another core aspect of our disease model is the commuting sub-model. Matrices were established to model commuting in the Netherlands. Infected work-commuters are able to infect all non-commuting susceptible people in the age groups 25-35 year, 35-50 year and 50-65 year. Infected school-commuters are only able to infect non-commuting school children in the destination municipality. The complete model with the integrated commuting sub-model results in the following situation for the home municipality x:

$$CI_{xy} = \beta S_x I_y \tag{6.2a}$$

$$CI_{xz} = \beta S_x I_z \tag{6.2b}$$

$$\frac{dS_{3x}}{dt} = S_{3x} - \beta_{3x} + v_{3x} + (A_{S_{2x}} - A_{1S_{3x}} - A_{2S_{3x}}) - CI_{xy} - CI_{xz} \tag{6.3}$$

$$\frac{dSE_{3x}}{dt} = E_{3x} - \beta_{3x} - \varepsilon_{3x} + (A_{E_{2x}} - A_{1E_{3x}} - A_{2E_{3x}}) + C_{xy} + C_{xz} \tag{6.4}$$

[5] *https://www.cbs.nl/nl-nl/publicatie/2014/50/demografische-kerncijfers-per-gemeente-2014*

Where S is the susceptible population, E is the exposed population, I is the infected population, C are children. The destination municipality is indicated with the variable x, and the home communities of commuters with y and z. The force of infection (CI) is calculated by multiplying the susceptible population in the destination community with the infected population in the home community times the rate of transmission (β).

Age groups (A) are indicated with numbers. Thus, equation 6.3 calculates the change in susceptible population in age group 3 of municipality x over time, based on four components: the susceptible population in the community multiplied by the infection rate, inflow of susceptible population due to waning of immunity (v), the movement of people between the age groups (inflow from group 2) and commuting to the municipality from a number of other municipalities (in this case x and y). The variable v indicates the natural loss of immunity over time due to waning. Equation 6.4 calculates the change in exposed population over time for age group 3 and community x in a similar way. In this expression ε represents the infectivity, the capability of a pathogen to cause an infection.

Adult commuting is based on daily commuting patterns at the municipality level (CBS 2013[6]). No commuting data was available for adolescents so we assumed that 5% of the adolescents will commute to one of the neighbouring municipalities. Commuting takes place once per day during 5 days a week with the option to apply holidays. The resulting patterns are shown in Figure 6-4. School commuting is more homogeneously spread over the country with similar distances, whereas work-commuting shows smaller and larger distances and concentrates in larger cities.

The model starts with the introduction of an initial number of cases at an origin location. Both the number of initial cases and the origin location can be set by the user of the model. Spatial heterogeneity in risk is achieved by applying different childhood immunity levels in different municipalities. These levels are based on childhood vaccination data from the Dutch health atlas ("RIVM-Zorgatlas" (2013). Adult immunity is assumed to be negligible at the start of the simulation (1996). Immunity after infection is set to 12 years, with an infection period of 14 to 21 days (see table 6-2). The time step of the model is 1 day and duration of the simulation is in principle indefinite.

To facilitate the comparison of the empirical and simulated data, the simulated data were again aggregated spatially to the percolation zones (Figure 6-1),

[6] *https://www.cbs.nl/en-gb/news/2013/23/more-than-half-of-employees-commute-to-work*

and epidemic periods with the same duration (55 week time interval) were extracted.

Evaluating spatial-temporal diffusion patterns

Trajectories of disease diffusion were mapped using a combination of Self-Organizing Maps (SOMs) and Sammon's projection, as proposed by Augustijn and Zurita-Milla (2013). SOMs are a type of unsupervised artificial neural networks that were introduced by Kohonen (2001) to cluster high dimensional data by projecting it onto a low-dimensional lattice that consists of neurons that are iteratively trained to extract patterns from the input data. These patterns are generalizations of the input data and are referred to as codebook vectors.

The following steps were followed to use SOMs in this study: (1) train a SOM with the available input data, (2) map the data onto the trained SOM. (3) Create a Sammon's Projection of the codebook vectors, and (4) connect the neurons in a chronological order to obtain the disease diffusion trajectory.

The Kohonen R package (2001, Wehrens and Buydens, 2007) was used to create and train the SOMs. A 3x4 lattice was used because of the relatively low number of training vectors (for the empirical data 55 * 6) and the fact that no secondary clustering will be applied. The same lattice size was used throughout the experiments.

The Sammon's Projection allows to further reduce the dimensionality of the trained SOM. In fact, it allows the representation of the spatio-temporal diffusion pattern of each epidemic as one vector that connects different characteristic states (SOM neurons). Each of the states can be visualized in a map showing the spatial infection level that corresponds to this state. This projection is used in two different ways: to compare the trajectory of the different epidemics to find epidemics with similar diffusion patterns, and to determine critical states and neurons that are important for the disease diffusion process.

Setup of the experiments

In this study we performed the following experiments (see Figure 6-5):

Experiment 1: Check the similarity between the diffusion patterns in the empirical surveillance data. We do this to detect self-similarity between epidemics (a similar Sammon trajectory), but also similarity in starting place and similarity in neurons (e.g. the same neuron appears in several epidemics as one of the starting neurons). Via this experiment we gain information about

the level of heterogeneity of spatio-temporal diffusion patterns of pertussis in the Netherlands. In a follow-up experiment we check for a similar level of heterogeneity in the simulated data. For this experiment, the SOM is trained using the surveillance data and the same dataset is mapped back.

Experiment 2: Check if the model used to generate simulated data is structurally correct. We do this by checking how sensitive the model is to changes in starting place of the infection, and by checking if subsequent epidemics generated during the same simulation run show different/similar diffusion patterns:

- The origin of infection. Three starting places are used, Amsterdam, Rotterdam and Utrecht, three large cities in the Netherlands. This experiment allows us to check if the place of origin has an effect on the spatial diffusion pattern.
- Subsequent epidemics: We run simulations for a long period of time (up to approximately 40 years), in which multiple epidemics can take place within one simulation run. We evaluate the differences between initial and follow up epidemics.

Experiment 3: Map back the empirical data onto the SOM trained using the simulated data. This allows the following comparisons:

- Evaluation of the similarity of the trajectories generated via this experiment with the trajectories of experiment 1 (empirical data mapped back to a SOM trained with the same data). In this way the loss of information of mapping empirical data to a different SOM can be evaluated.
- Evaluation of the similarity of trajectories between experiment 2 (simulated diffusion patterns) and empirical patterns.

Experiment 4: Conduct simulation runs using only adolescent commuting, only adult commuting and combined adolescent and adult commuting. This allows the evaluation of the impact of commuting on the diffusion patterns. More particularly, we can check if adding adolescent commuting to a disease model has an impact on the spatio-temporal diffusion patterns of the model.

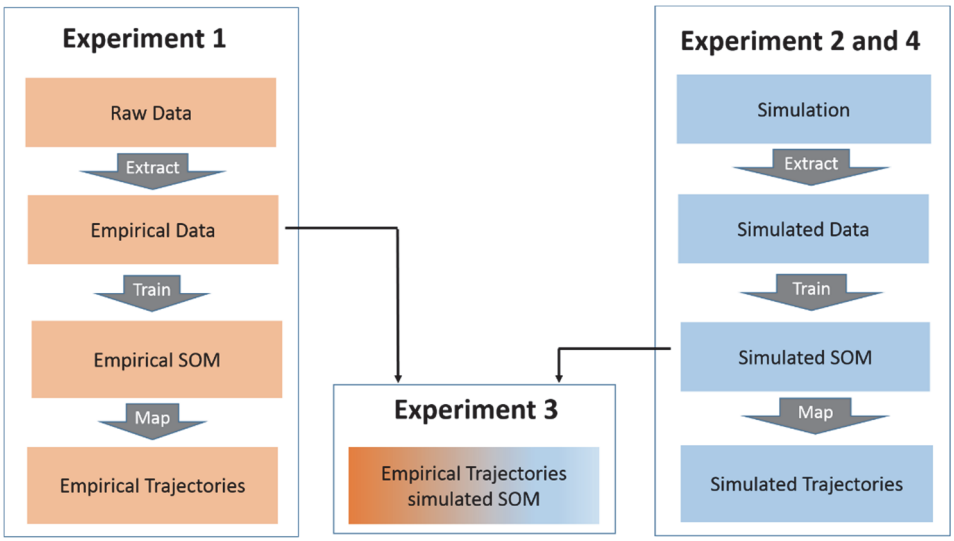

Figure 6-5 Overview of all experiments

6.3 Results

Experiment 1: Mapping diffusion of surveillance data

The 6 epidemic periods of 55 weeks each were used to train a 3x4 SOM lattice. Results are shown in Figure 6-6. The left side of the Figure shows the codebook vectors for the trained SOM lattice, the right hand side shows the same neurons as maps showing the intensity of infection. The codebook vector (red line) does not represent a time series but the number of infections per zone, starting with zone A on the left to zone P on the right.

We immediately notice that neurons 1-3 represent a situation where infection levels are low in all zones. Intuitively we might expect that epidemics will start from neurons 1, 2 or 3 and also end in these neurons. The vector of neuron 4 shows a high infection on the left side (zones with letters A, B, C are located in the west of the country) and might be one of the initial stages of the trajectories where diffusion is forced from the highly populated areas in the western part of the country. We see a similar but intensified pattern in neuron 7.
Neurons 8 and 10 seem to represent situations where all zones are infected (equally). Neurons 9 and 12 show the highest peaks on the right hand side (zones with letters M, N, O, P mainly in the east of the county) indicating that perhaps these neurons represent the "decline" of the diffusion pattern.

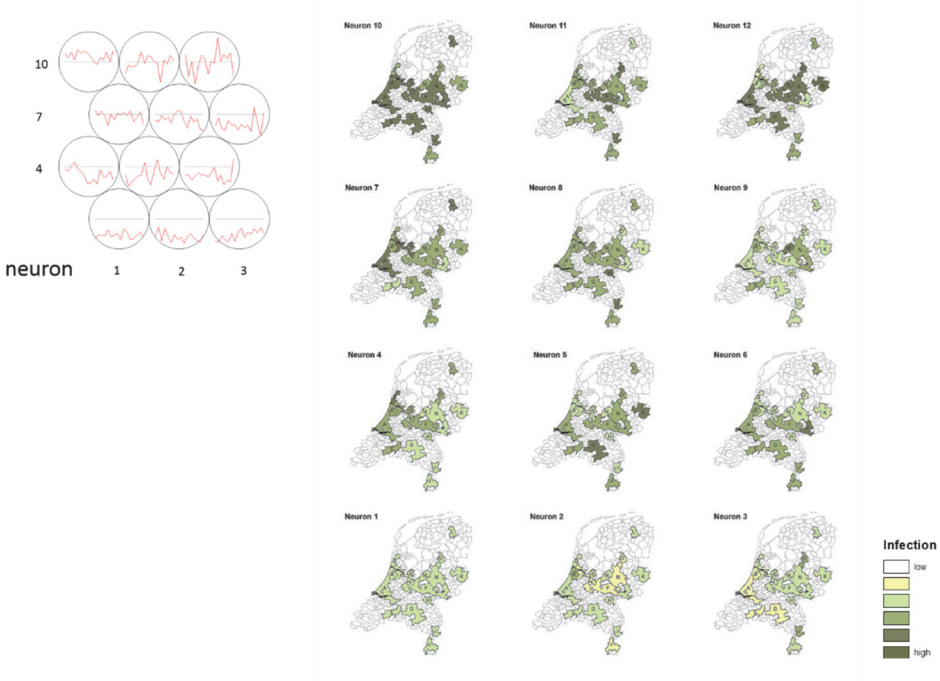

Figure 6-6 Trained 3x4 SOM lattice. Numbers indicate neuron numbers (left) Mapping of neurons 1-12 as maps, yellow represents low infection, middle green represents medium infection, dark green is high infection, letters indicate the different zones (right)

Figure 6-7 shows the results of the Sammon projection of the 6 epidemics to a two-dimensional space. This was done by first projecting each of the epidemics to the trained SOM and connecting the mapped values to form a trajectory, visualizing the spatial temporal diffusion pattern.

By comparing the trajectories of different epidemics, we can identify if similar spatio-temporal diffusion patterns exist. It is clear that there is some similarity between the patterns of 1996 and 2004, but that there are more striking dissimilarities. Different epidemics have different diffusion trajectories.

We can also evaluate the neurons (diffusion steps) that seem to be critical in the diffusion process. Especially at the start of the diffusion process as these neurons might represent tipping points, leading to the further diffusion of the disease.

All of the patterns move from the endemic state (neurons 1,2,3) to a combination of neurons 4 and 8. These steps seem to be crucial for the development of an epidemic. These neurons represent patterns with higher infection in the western part of the country but also in the south. The main difference between neurons 4 and 8 is that in neuron 8 Rotterdam (zone A) is

more dominant. When we make a thorough comparison we see the following similarities between the epidemics:

1996 and **2004** follow the same pattern. Diffusion starts in the western part of the country, proceeds to neuron 10, further develops to neuron 11, and in the later stages persists in the south of the country (neurons 5, 6 and 9). They visit almost all neurons, move from the left side of the projection, via the bottom to the right side to move up and back to neurons 1.

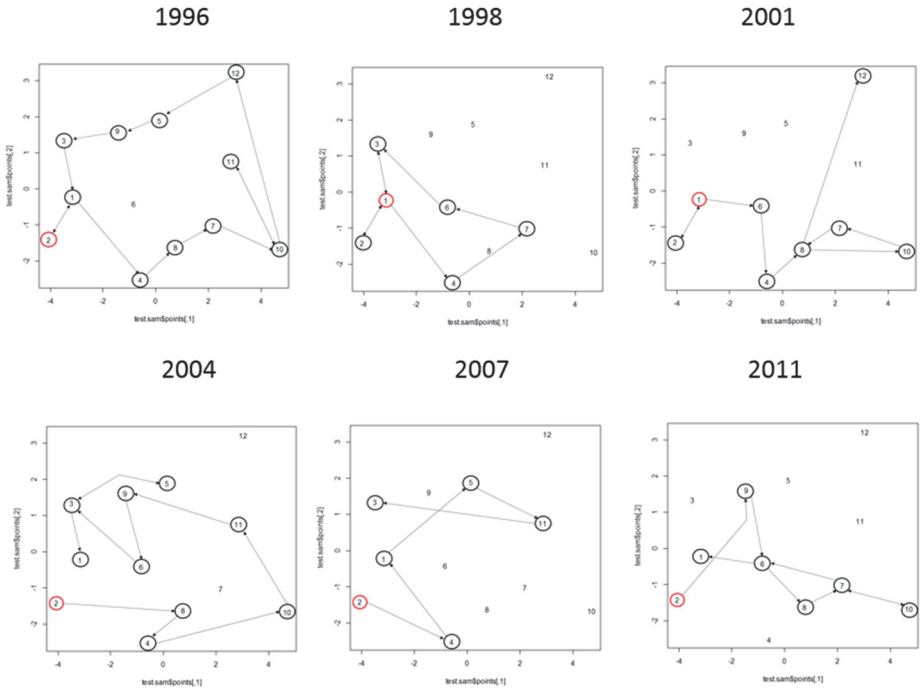

Figure 6-7 Sammon's projection with diffusion vectors for epidemic 1 (top left) to epidemic 6 (bottom right). Red circle indicates the start of the diffusion. Arrows indicate the direction between neurons.

1998 and **2007**: These two epidemics may not be grouped on a first examination. Yet both patterns are very small and many neurons are <u>not</u> visited. Here one might argue if these are full blown epidemics.

2001 and **2011**: At the first glance they do not display similar patterns, yet both have a striking element: there is a neuron connected to several different other neurons. In 2001 this is neuron 8 (with connections to 4, 7, 10 and 12) and in 2011 this is neuron 6 (with connections to 1, 7, 8 and 9). These seem to be "persistent" neurons. Neuron 6 covers the south-eastern part of the country and neuron 8 the zones 1 and 3 (west of the country).

Based on the diffusion trajectories of all 6 epidemics we conclude that subsequent epidemics have different spatio-temporal diffusion patterns. However, similarity between epidemics seems to happen between epidemics that are a larger number of years apart (around 9 years) and have several in-between epidemics. All trajectories seem to move via a critical state of high infection in the western part of the country.

Table 6-2 overview of simulation parameters

run	summer	spring/ autumn	winter	years	Infectiousness (days)	adult commuting	Adolescent commuting	start infection	number start infections
1	0.4	0.2	0.2	12	21	on	on	Utrecht	1
2	0.4	0.2	0.2	12	21	on	on	Utrecht	1
4	0.4	0.2	0.1	12	14	on	on	Amsterdam	10
5	0.4	0.2	0.1	12	14	off	on	Amsterdam	10
6	0.4	0.2	0.1	12	14	on	on	Amsterdam	10
7	0.4	0.2	0.1	12	14	off	on	Utrecht	10
8	0.4	0.2	0.1	12	14	on	on	Rotterdam	10
9	0.4	0.2	0.1	12	14	off	on	Rotterdam	10
11	0.4	0.2	0.1	12	21	on	on	Utrecht	10
12	0.4	0.2	0.1	12	14	on	on	Rotterdam	10
13	0.8	0.2	0.1	12	21	on	off	Utrecht	10
14	0.8	0.2	0.1	12	21	on	off	Amsterdam	10
15	0.8	0.2	0.1	12	21	on	off	Rotterdam	10

Experiment 2 –Patterns in simulated data

Simulated pertussis data was generated using the model parameters listed in Table 6-2. This data was then split into epidemic periods of 55 weeks. A 3 x 4 SOM lattice was trained and the diffusion trajectories were then displayed using the Sammon's projection. Results are shown in Figure 6-8. A total of 25 simulated epidemics were analysed. This includes some runs that led to multiple epidemics: 2 (2), 4 (2), 5 (3), 7 (6), 9 (3), 11 (2) and 12 (2).

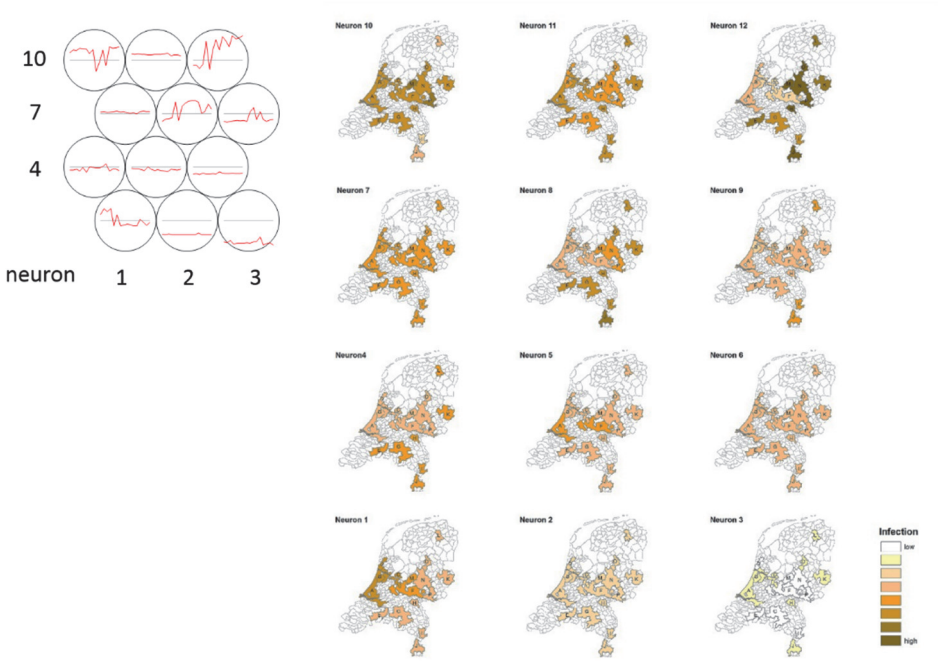

Figure 6-8 Trained lattice with numbers referring to the neuron number (left). Maps of the codebook vectors of the trained lattice with letters indicating the zone (right)

When we evaluate the neurons of Figure 6-8 there are several recognizable patterns: neuron 1 represents a situation with high infection in the western part of the country. There is also a high level of infection in neuron 10, yet this infection is not only in the west but in the complete middle of the country, only excluding the south and north. Neuron 12 also represents a situation with a lot of infection yet excluding the west and middle of the country. There are also many neurons that vary in the intensity of infection, show little difference between the geographic regions. Examples are neurons 2, 6 and 7, and to a lesser extent neurons 4, 5 and 11. When we compare this with the SOM trained based on the empirical data, we notice that there were also neurons showing little geographic variation (e.g. neuron 1) but in general the diversity between the regions was larger.

Comparison of multiple epidemics generated by the same run

The diffusion trajectories of the epidemics generated by the same simulation run differ, as shown in figure 6-9 for run 7 (adolescent commuting, starting in Utrecht). There seems to be similarity between the epidemics generated within the same run but not for consecutive epidemics. Epidemics 7a and 7e show almost identical diffusion patterns, and these epidemics are quite similar to 7c. All of these patterns start in neuron 12 (heaviest infection in the east) and end in 5. The other three epidemics (7, 7b and 7d) also show a certain similarity,

starting in neuron 9 and proceeding via neuron 5 to 10 and 11. The fact that we find different diffusion patterns within a single simulation run is encouraging as we found in the analysis of the empirical data that in reality also different diffusion patterns exist. It proofs that the model is able to generate different diffusion patterns (model shows a good level of realism).

The neurons representing a high infection in the west of the country are neurons 1 and 10. We have seen in the empirical data that all diffusion patterns had a high infection in the western part of the country at the start of the diffusion (neuron 4 of figure 6-6). A similar pattern can be observed in the simulated trajectories, in these patterns the third or fourth neuron are also either 1 or 10.

As can be seen in figure 6-9, the trajectories cover less neurons per epidemic compared figure 6-6. We see no large circular movement in the trajectories like we see in the result of experiment 1 for 1996 and 2004. Not all trajectories return to their starting neurons. Especially epidemic 7a, 7c and 7d end far from their original starting neuron.

Difference in starting place of the infection
Figure 6-10 shows the results of different starting places of the infection. When only adolescent commuting is turned on, there is no effect of the starting place on the diffusion pattern, all three trajectories start in neuron 9 and proceed to neuron 5. Also for only other commuting types, the patterns for the different starting places are very similar. The Netherlands is a small country, with high commuting levels. The infection spreads fast and the effect of the initial location of infection seems negligible.

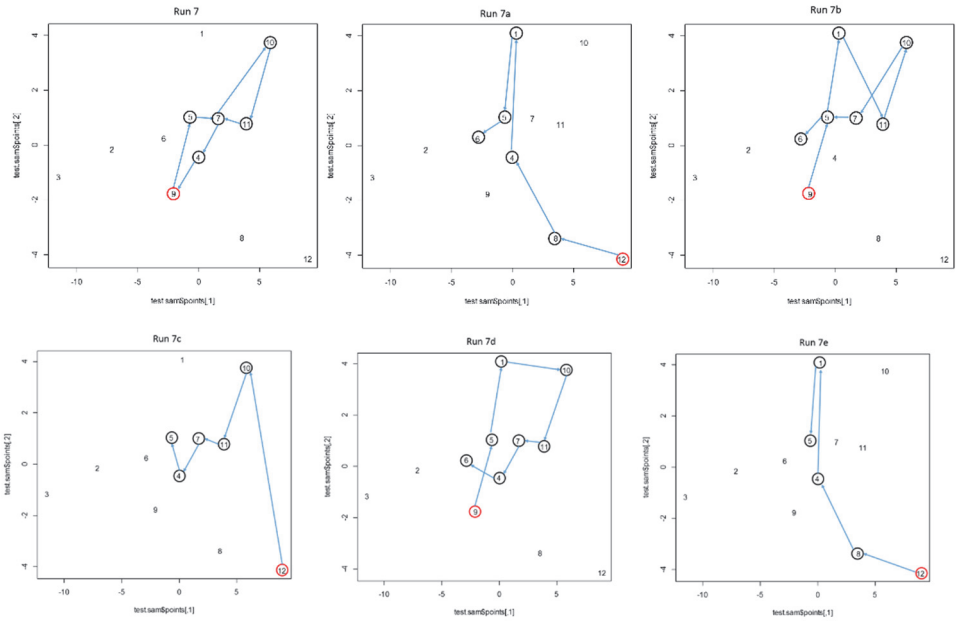

Figure 6-9 Comparison of outbreaks generated in the same simulation run. Numbers indicate the neuron number. Red circle is the starting neuron. Arrows indicate the direction of diffusion.

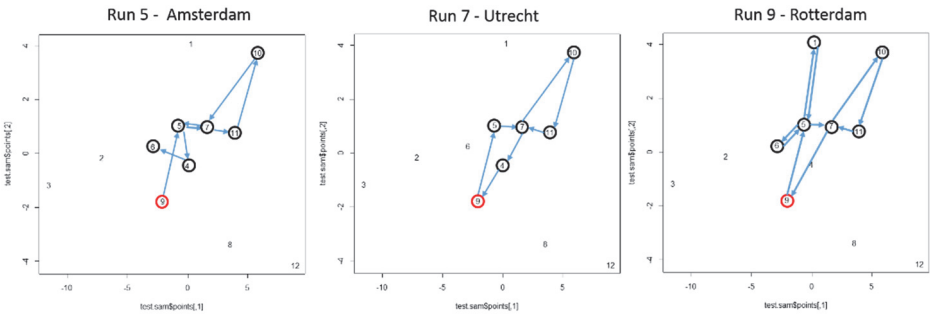

Figure 6-10 Comparison starting places adolescent commuting only. Numbers indicate the neuron number. Red circle is the starting neuron. Arrows indicate the direction of diffusion.

Experiment 3: Comparison of simulated and empirical patterns

We can compare the empirical and the simulated data like we did in the previous experiments, by visually comparing the two trained SOMs (Figures 6-6 and 6-8) and the corresponding trajectories. In this experiment we will compare the simulated and empirical diffusion patterns based on a single SOM.

For the comparison of simulated and empirical patterns we use the SOM trained with the simulated data. This choice was made because the simulated data has less "noise" compared to the empirical data and shows more generalized patterns making it easier to interpret.
First we mapped back the 6 empirical epidemics. Then we generated the diffusion trajectories using the Sammon's projection (Figure 6-11).

Comparison of the current mapping with experiment 1
We try to identify if the properties of the data were maintained while mapping it back to the SOM trained with simulated data. In experiment 1 we noticed that consecutive epidemics had different trajectories, this seems to be maintained when mapping the data to the SOM trained with simulated data (Figure 6-11)

In Figure 6-6 (SOM trained with empirical data) we observed similarity between the epidemics of 1996 and 2004, the epidemics of 1998 and 2007 and between 2001 and 2011. When mapping back the same data to the SOM trained with the simulated data (Figure 6-11) we again see some similarities but not between the same epidemics. The trajectory of 1996 completely overlaps the trajectory of 1998 and 2011 but it is not similar to 2004.

In experiment 1 the trajectories for 1996 and 2004 were very large, visiting many neurons. In this experiment, the trajectory of 2004 is again large but for 1996 we find a very small trajectory, containing only 4 neurons.

In experiment 1 we also observed that all epidemics moved through a "critical" neuron (neuron 4). Most trajectories (2004, 2007 and 2011) start in neuron 5 which represents an higher infection in the west of the country. The trajectories of 1996 and 1998 start in neuron 6 yet, with a double arrow movement to neuron 5. The trajectory of 2001 starts in neuron 9 represents an infection in the south, yet moves from 9 to 5.

Based on these findings we can conclude that a lot of information on spatio-temporal diffusion patterns is lost when mapping empirical data to a SOM trained with simulated data. Yet, the start of the trajectory also seems to follow a consistent order via the west of the country.

Comparison with the simulated runs (experiment 2)
When we compare the patterns of the empirical data (Figure 6-11) with the simulated data (Figure 6-9 to 6-10) we notice that in the empirical epidemics we see many double arrows indicating that there are back and forward changes between the neurons. This is less common in the trajectories of the simulated data. In the simulated patterns with both types of commuting turned on, the

pattern often started in neuron 9 (lower side of the lattice). For the empirical data, we see different starting points; all in the centre of the Sammon plots.

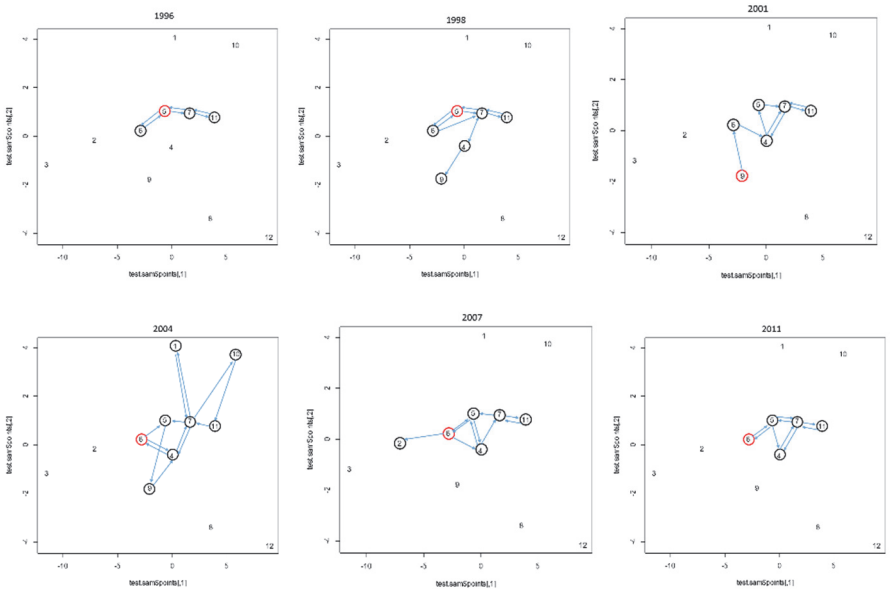

Figure 6-11 Diffusion patterns of the empirical data mapped onto the SOM trained with simulated data Numbers indicate the neuron number. Red circle is the starting neuron. Arrows indicate the direction of diffusion.

Experiment 4: Effect of commuting on the simulated diffusion patterns

Three types of runs were conducted: only adolescent commuting, only adult commuting and the combination. Results are shown in Figure 6-12. In this figure, the top row shows only adolescent commuting, the middle row only adult commuting, and the bottom row shows patterns generated using both types of commuting. For adolescent commuting, the trajectories start in neuron 3 (left side of the plot), the trajectories for adult commuting start in neuron 9 (bottom). Neuron 9 represents a situation of higher infection in the north and south of the country. The trajectories with combined commuting start in different places.

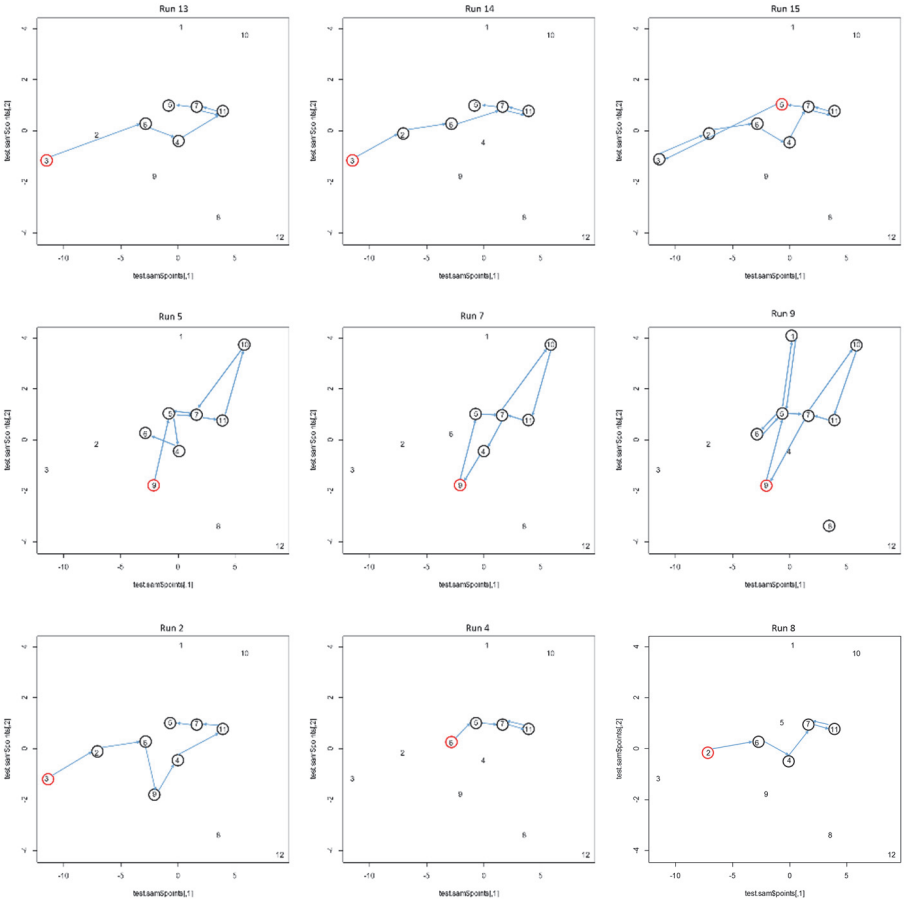

Figure 6-12 Comparison of epidemics with three types of commuting (top – only adult commuting, middle – only adolescent, bottom – both types). Numbers indicate the neuron number. Red circle is the starting neuron. Arrows indicate the direction of diffusion.

In general, the epidemics generated with both types of commuting seem to be more concentrated in the centre the Sammon's projection compared to the examples with only adolescent commuting or only adult commuting. Trajectories with both types of commuting show more similarity to only adolescent commuting then to only adult commuting. This is surprising as adolescent commuting is implemented for a smaller population group (5% of 12 – 17 years old) and adult commuting covers a larger age group (25-65 years old). However, it should be noted that adult commuting is over both small and large distances.

Runs performed with adolescent commuting lead to multiple epidemics. However, when only adult commuting or both adolescent and adult commuting are turned on, the disease becomes endemic and no second outbreak is

generated. This was expected as the empirical data shows that the disease becomes endemic in the Netherlands, however with periodic epidemics. None of our experiment reproduces this result.

Comparing the results of Figure 6-12 with the results of experiment 3, we notice that the situation of 1996 closely resembles the pattern of both types of commuting of run 4. The situation with only adult commuting is most similar to 2004 of the empirical data.

6.4 Conclusions and further work

In this chapter we present a method to compare spatial-temporal diffusion patterns of empirical and simulated data. SOMs were used in combination with a Sammon's projection to extract the disease diffusion trajectory in a low-dimensional space, which makes it easier to visualize and to understand the diffusion of a disease.

Our approach allows comparing the sequence of diffusion (order in which neurons are visited) as well as the complete diffusion pattern (shape of the trajectory). Especially the sequence of neurons in the trajectories turned out to be useful. Where the shape of the trajectory, and the number of neurons visited in each trajectory differ, the starting neurons of the trajectories turn out to be relatively stable and revealed useful information in both empirical and simulated patterns.

The similarity between spatio-temporal patterns of pertussis epidemics in the Netherlands is limited. In the empirical data we identified similarity between the epidemics of 1996 and 2004, between 1998 and 2007 and between 2001 and 2011 based on the shape of the trajectory. However, there were also clear differences between these pairs. The order of the neurons for all epidemics indicated a start of diffusion via neuron 4 (representing an increase of infection in the west of the country). This indicates that areas with a higher population are driving forces for epidemic diffusion.

Two SOMs were trained, one with the empirical data and one with the simulated data. The SOM trained with the empirical data show more differences in infection levels between zones compared to SOM trained with simulated data. The latter SOM shows a more homogeneous infection level of all zones. This reveals more variation between zones in the empirical data compared to the simulated data.

Empirical and simulated diffusion patterns were compared in two ways, by mapping them to two different SOMs and by mapping empirical and simulated data to the same SOM (in this case, a SOM trained based on the simulated

data). Mapping the empirical data back to the SOM trained with simulated data, allows for good visual comparison between empirical and simulated data yet, also shows a large change in the trajectories of the empirical data. The trajectories show many back and forward movements between the same neurons. Small changes in infection levels can trigger these changes.

The quality of the comparison of simulated and empirical diffusion patterns is affected by the level of completeness in the surveillance data, which is never complete due to disease characteristics (e.g. a-symptomatic and a-typical disease cases). Some authors state that the number of reported cases is in the order of 0.3% to 2% of the total number of infections in a population (McDonald et al., 2014; Van Der Maas, 2013). Moreover, it is hard to infer the initial time of infection because of the notification system.

A specific objective of this research was to identify if introducing age-specific mobility (of adolescents) in simulation models influences the spatio-temporal diffusion patterns. Results show differences in spatio-temporal diffusion patterns for different commuting types. This finding emphasises the importance of mixing different types of mobility for different population groups and it is important because the simulation of complex systems requires the inclusion of all the essential processes.

We also found that state transitions from an epidemic to an endemic state only take place for adult commuting. However, re-occurring epidemics with loss of infection (fade-outs) between epidemics can only be re-created when using adolescent commuting. Mixing both types of commuting generated transition to an endemic state. Yet, this did not re-create the periodic epidemics currently taking place.

There is more movement of population besides work and school related commuting. On a daily basis we all make many trips for either shopping, or social activities. This type of commuting is not yet included in the model, and has a different spatial-temporal pattern compared to our experiments. These trips are frequent but over a shorter distance and are conducted by all age-groups. It would be interesting to add this as a new commuting component to the model. This can lead to an endemic state with periodic epidemics.

Chapter 7 Synthesis, conclusions and future work

The core of this PhD thesis consists of three journal articles and two chapters produced as "stand alone" research elements but related to either the recognition of patterns in empirical data, or the reproduction of these patterns via simulation. In this chapter, I discuss the relationships among the previous chapters. The discussion is articulated as a synthesis of knowledge (i.e. a reflection) that focuses on the three sub-objectives listed in chapter 1:

a. To develop and evaluate pattern detection techniques that can identify (robust/self-repeating) spatio-temporal patterns that can be used to build and validate ABMs.
b. To develop and evaluate methods to use these patterns when building and validating geographically explicit ABMs.
c. To develop and evaluate methods for the comparison of simulated and empirical patterns.

To facilitate the comparison between the clustering methods discussed in chapters 2 and 3, here I apply SOMs to the pertussis dataset and check if the same Critical Community Region (CCR) is identified (*reflection 1; section 7.1*). In this thesis, three agent-based models are presented: a model for informal settlements in Dar es Salaam (chapter 4), a cholera model for Kumasi in Ghana (chapter 5) and a pertussis model for the Netherlands (chapter 6). Here I evaluate in what respect the models differ and the differences in the patterns they produce (reflection 2; section 7.2). In Chapters 4, 5 and 6 simulated and empirical patterns are compared. Here I reflect on this comparison (*Reflection 3; section 7.3*).

After the synthesis of knowledge, I answer the research questions and summarize the main research achievements (section 7.4), and provide an outlook to future research opportunities (Section 7.5). The setup of this chapter is illustrated in Figure 7.1.

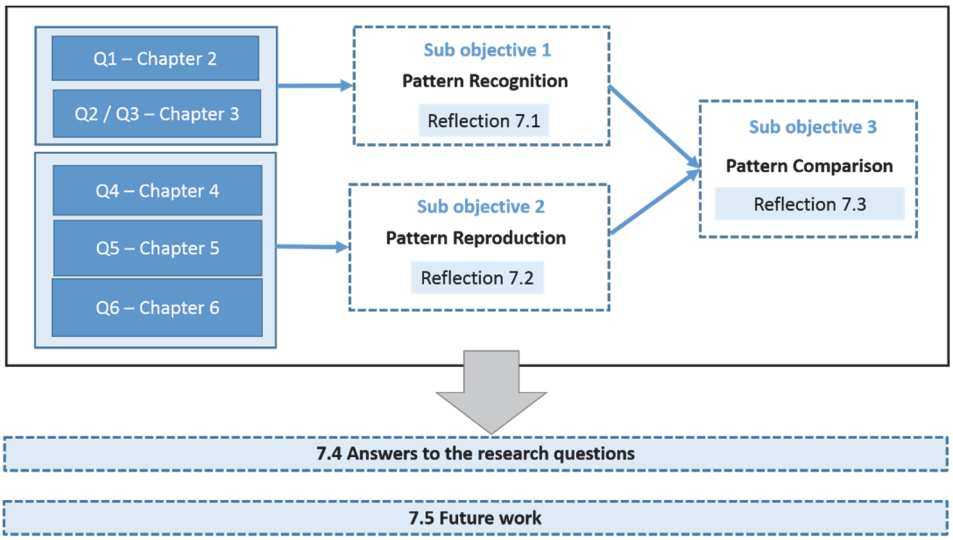

Figure 7-1 Overview of chapter 7

7.1 Reflection on pattern recognition

In chapters 2 and 3, SOM and SBD were respectively used to cluster empirical time series of disease data. The results of these methods cannot be directly compared because the SOM was used to cluster measles data for Iceland and the SBD to cluster Dutch pertussis data. Here I compare SOM and SBD as a primary clustering method by extracting the CCR for the Dutch pertussis dataset using SOM. This is followed by a comparison of the methods listing their advantages and disadvantages.

To train the SOM, equal length periods of 55 weeks, representing the 6 epidemics, were extracted from the input datasets. The SOM was then trained with this data using a lattice of 6 by 6 neurons and 100000 iterations. A secondary hierarchical clustering was applied to this SOM lattice, to make the output comparable with the 6 clusters obtained via the SBD clustering presented in chapter 3. After mapping back the training data, a map was created showing the zones that were mapped to each of the 6 clusters (Figure 7-2).

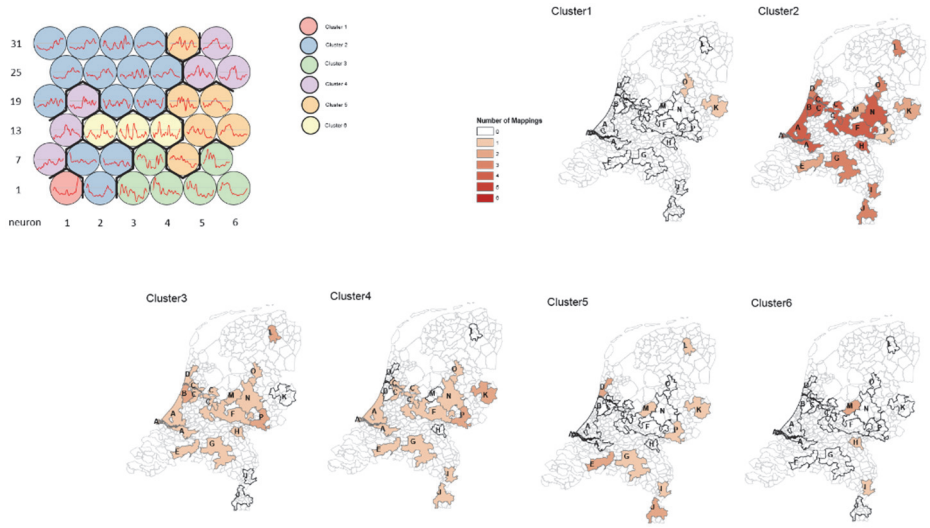

Figure 7-2 Trained SOM lattice. (organized from neuron 1 (bottom left) to neuron 36 (top right)) showing in dark lines and colors the hierarchical clustering revealing six clusters (top left) and the mapping of the 6 extracted clusters (right). Darker colors indicate a higher number of mappings. Letters refer to the zone numbers.

The interpretation of the six extracted clusters is as follows:

Cluster 1: Seems to be a late peak in the east of the country (zone 11 and 15). The cluster consists of only one neuron (neuron 1) and only two mappings were found to this cluster. This is related to a single year and should be regarded as an outlier.

Cluster 2: Has 49 mappings and is the cluster with the largest number of mappings. The cluster includes neuron 2, 8, 9, 19, 21, 22, 25, 26, 27 etc. of the training lattice. The codebook vectors vary considerably, yet his cluster contains some vectors that seem to have a steady infection throughout the time series. This cluster represents the CCR we are looking for.

Cluster 3: Has a total of 16 mappings. The cluster includes neurons 3, 4, 5, 6, 10, 12. It represents an early peak with a decline in infection later in the time series.

Cluster 4: Has 13 mappings. The cluster includes neurons 7, 13, 20, 29, 30, 36. This cluster represents a large peak somewhere in the middle of the time series.

Cluster 5: This cluster has 12 mappings. It includes neurons 17, 18, 23, 24, 35. It represents an early peak with decline afterwards.

Cluster 6: Cluster 6 has only 4 mappings. It relates to neurons 14,15, 16. These neurons represent the typical stuttering chain sequence of infection and fade-out.

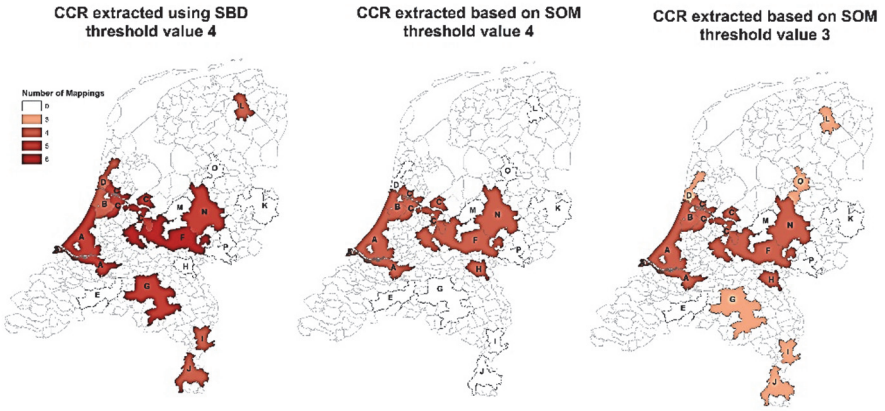

Figure 7-3 CCR region derived via the SBD method(left) the SOM with the same threshold value of 4 (middle) and SOM with a threshold value of 3 (right). Letters indicate the zones.

The CCR is extracted using the same threshold of 4 out of 6 epidemics that was used in chapter 3. Areas that are mapped 4 or more times to cluster 2 are shown as the CCR (Figure 7-3). When we compare the results of the two methods, the CCR delineated by the SOM (Figure 7-3 middle) is smaller than the one extracted by SBD (Figure 7-3 left). Yet, the zones included in the CCR derived via the SOM are also included in the SBD version, except for zone 8. In the south, the zones 7, 9 and 10 have dropped out of the CCR and in the north zone 12. Areas that drop out are mostly located on the fringe of the country (north and south). Areas located in the middle and the highly populated west part of the country are in both zones. If we would relax the threshold to 3 instead of 4 for the SOM based method (Figure 7-3 right), zones 7, 9, 10, and 12 would be included in the CCR and the identified zones would be "fairly" identical.

It is not surprising that so many of our zones belong to the CCR. These zones represent the most densely populated areas of the Netherlands and are the most likely candidates to show endemic behaviour. In case we would have included all of the country, we would most likely not have found any additional areas. By first extracting the percolation zones and then applying the clustering we already pre-selected the most likely candidate areas.

The Rand index between the SOM and SBD mappings is 0.56. This indicates clear differences between the two mappings. SOM results seem to be more topological (splitting later and earlier peaks) but the SOM mixes single and double peak time sequences in the same cluster because it is not a shape based method. Moreover, the SOM does not apply an alignment along the time axis, it groups based on when the peak occurs and how high the peak is.

Table 7-1 Comparison of the two clustering methods

	SBD	**SOM**
Input data	z-normalized	Can be log transformed as this maintains some of the amplitude variations or z-normalized.
Input data	Can have unequal length of time series	Must have equal length of time series
Method		Requires the selection of the size of the lattice, however, when secondary clustering is applied this becomes less relevant
Method	Iterative training in which centroids are updated and mapping is redone	Iterative training process in which codebook vectors are adjusted
Method	One step process of distance (similarity) measuring and clustering	Multi-step process of training the lattice, mapping data back to the trained lattice and secondary clustering.
Method	Global alignment method (aligns pairs of time series) Two identical time series with a time displacement will be fully aligned and mapped to the same cluster	Topological method, without alignment along the time axis, but with grouping of time series that peak around the same time. Consequently, time series with a horizontal displacement may end up in a different cluster.
Method	Suitable method when shape matters but phase is unimportant.	Suitable method when phase (timing) matters, and shape is of lesser importance
Output	Easy to extract the centroids of the clusters which provide an easy interpretation.	Visualization of the training lattices can be useful in the visual analytics depending on the complexity and length of the input time series.

Table 7-1 provides an overview of the two methods applied. The general conclusion is that when clustering a single epidemic in areas with similar time series, the SOM is the preferred method. However, when working with multiple epidemics, time shifts between the time series should be eliminated and an SBD based approach is therefore preferred.

The SBD method can "fix" small shifts in time so that areas with similar time series with a phase difference will be clustered together. This can be an advantage or disadvantage. As shown in chapter 1, hierarchical diffusion leads to populations of the same size being infected simultaneously, and smaller populations being infected later. Difference in timing is "meaningful" in this type of diffusion, and information can be extracted from it. Fixing the displacement removes useful information. However, when we compare the time series of different epidemics, small time shifts will appear, as alignment of the different epidemics will never be perfect. As we want to identify shape similarity, regardless of small shifts in time, SBD is a better alternative.

In general, both methods are easy to apply, and both assume scaling of the input data (z-normalized or log transform). Amplitude is therefore a less important factor. Yet, assuming that when the number of cases is high in two populations, the shape of the time series curve will also be similar this is not a large restriction.

The visualization of the output of these two methods differs. SBD visualizes the clusters showing the original inputs, whereas the codebook vector of the SOM is generalized data. Yet, in a trained SOM lattice we can detect the topological relationships between clusters and this is not visible in the SBD result. Both methods can lead to an extracted vector representing a cluster. In the SBD this is the cluster centroid, in the SOM the codebook vector.

The biggest advantages of the SOM are that it can be applied both in space and time and that it can be combined with the Sammon's projection to create a diffusion trajectory. Application in time is impossible with SBD as when comparing in time, the input vector is no longer a time series but a representation of spatial locations and the alignment of these spatial locations does not make any sense. Although the diffusion trajectory is a powerful representation of a complex spatio-temporal process it consists of a range of sources of information including: the general shape of the trajectory; number of neurons visited; back and forward transitions between neurons; critical transitions that can be linked to state transitions of the system etc.. Most likely, not all of these aspects are equally informative, but more information is needed to pick out the most important aspects. In principle SOMs allow for real-time streaming of surveillance data (mapping back data during the simulation) as incomplete time series can be mapped back. This might be an additional advantage.

The largest disadvantages of the SOM method are the alignment problems and the fact that a considerable size dataset is needed to train the SOM.

7.2 Reflection on pattern reproduction

The second sub-objective focusses on ways to use patterns when building geographically explicit ABMs. The aim is to create spatial agent-based models that have the necessary complexity to reproduce empirically observed spatial-temporal patterns. Here we evaluate the complexity of these models and of their outputs.

Table 7-2 Overview table characteristics of the models

	Cholera model	**Pertussis model**	**Informal settlement model**
Drivers	Two types of infection were modelled, environment to human infection (EH), and human to environment to human infection (HEH).	Mobility was used as driver and different types of mobility (adolescent and adult) were producing different patterns	Alignment or dispersion were used as drivers and indeed they were leading to completely different patterns.
Multiple drivers and complex patterns	The combination of the two infection types leads to more complex patterns compared to only one type of infection.	Yes, the combined mobility patterns of adolescent and adult commuters lead to different patterns.	Yes, the combined alignment and dispersion drivers lead to more complex pattern.
Scaleless behavior and self-similarity	Area simulated was too small to perform multi-scale analysis.	Although the area simulated was large enough this analysis has not been conducted. Applying SOMs in space should be able to detect self-similarity.	Area simulated was too small to perform multi-scale analysis.
State-transitions	State transition from non-epidemic to epidemic is simulated.	All simulations show transition to epidemic state. When using only adult commuting, the disease becomes endemic. Only adolescent shows epidemic behavior	When the informal settlement becomes full, no new houses can be added, transition from a state of growth to a state of consolidation.

In chapter 1 a number of characteristics of patterns produced by complex systems were described. In short: the system consists of multiple drivers and the combined effect of these drivers leads to a pattern that is more complex than patterns produced by a single driver. The system shows scaleless behaviour as patterns repeat themselves at different scales, and the systems

experiences state transitions. In table 7.2 we evaluate each of these factors against the three models developed.

When the structural realism of a model is correct we assume that it can reproduce the identified patterns. All models are designed to include different drivers and these drivers, indeed, lead to different patterns. The pertussis model is able to reproduce state transitions. To some extend this also applies to the informal settlement model. The limited duration of the cholera model (90 days) only allows for one epidemic but not for the occurrence of lasting transitions.

Table 7-3 Further reflection

	Cholera model	**Pertussis model**	**Informal settlement model**
Missing elements in model	Agents are not risk aware. No behavior change during the simulation.	No short distance mobility for all age groups. No/limited interventions by health units.	No growth of new roads (static road network). No social network.
Difficulties in pattern comparison	The pattern on duration of infection was not correctly matched.	The pattern for combined commuting shows transition from epidemic to endemic yet, the epidemics in the endemic phase are not pronounced enough.	
Limitations in empirical data	Data for a single epidemic. No boundaries communities.	Empirical data is always incomplete dues to a-symptomatic and a-typical disease cases.	Data collection at fixed intervals.

The biggest problem is the reproduction of scaleless behaviour and self-similarity in space. The spatial extent of the models for cholera and informal settlements is not large enough to observe patterns at different spatial scales. The only model simulating a spatial extent large enough to observe this type of behaviour is the pertussis model. This is because the model is not an agent-based model but designed as a hybrid model and we are using only the population based diffusion model for the analysis

Models are always an abstraction of reality and in all three models important elements are missing that could hinder the detection of patterns related to complexity (Table 7.3). In the pertussis model, interventions performed by health units are missing. The mobility model is also incomplete because we could have included mobility of elderly people and added random mobility for all age groups. In the cholera model we did not include the risk perception of the households. We also did not take movement of people into account. In the model of urban informal settlements, we did not include the creation of new roads but only focused on building houses. The choice of a location to build a house may also be influenced by the social network of the agents. The question is not if the models are complete, because they are certainly not but if they are complete enough to capture the complex systems they represent. As a modeller you make a choice based on which patterns you will validate your model. In case one of these patterns is not correctly reproduced this indicates a mismatch between the complex system and the model. But you will never know if you included enough patterns to show a possible mismatch.

In the cholera model there are two interesting elements: the pattern of duration was not correctly matched and the model shows that the same pattern can be reproduced by different types of models.

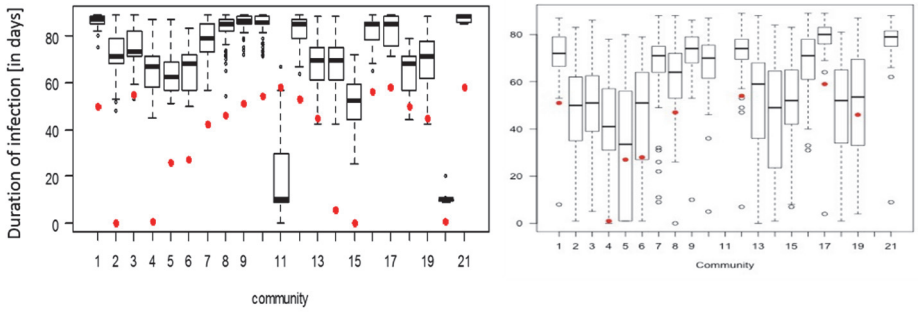

Figure 7-4 Comparison duration of infection. Result cholera model (left) result Abdulkareem et al.(Abdulkareem et al., 2018) (right). Red dots indicate empirical data, boxplots the simulated data.

As a modeller you might know that a certain pattern has not been correctly matched yet this does not immediately indicate how to solve the problem. For the cholera model, matching the percentage of infection per community was not difficult however, the duration of infection was overestimated in all communities (Figure 7.4 left). A later study by Abdulkareem et al. (2018) enhanced the agents with risk perception and behaviour change. This had a large impact on the duration of infection and the results match the empirical data more closely (Figure 7.4 right). This highlights the importance of using

multiple patterns including spatio-temporal patterns next to spatial and temporal patterns to evaluate the performance of (agent-based) models.

Many models for cholera diffusion exist. Most of these models do not model the way the cholera infection reaches the river, like the model discussed in this thesis, but the diffusion of cholera via the river system to downstream locations. The epidemic curves generated by our model are showing similar patterns compared to other models. The fact that completely different models can generate the same patterns asks for cautiousness. The fact that the model reproduces a pattern does not automatically validate this model. It also underlines the importance of matching patterns in time and space.

The informal settlements model is a vector based model where raster based models are more common in this domain. The model was developed as a vector based model to allow for a comparison with the empirical data. This way, model output can be directly compared with the empirical data and this is very useful during the design process. Where this method works for informal settlements, the spatial footprint of epidemics is lacking and such a method cannot be applied for disease modelling.

The pertussis model is a able to simulate a state transition from epidemic to endemic but during this endemic phase, epidemic outbreaks occur. These epidemics are not pronounced enough in the simulated data. In this case the patterns seem to indicate that short distance commuting for more age groups should be implemented. We can only check if this is the case by integrating this type of commuting into the model and comparing the patterns. We should realize that fixing this problem in a temporal pattern, will have an impact on the spatial patterns.

7.3 Reflection on pattern comparison

In the previous two reflections we discussed methods to detect spatiotemporal patterns in empirical data (7.1) and the models that re-produce the patterns (7.2). This section addresses the third sub-objective: how to compare simulated and empirical data. An overview of the patterns for the three models used can be found in Table 7.4. Not all patterns (spatial, temporal and spatio-temporal) were analysed for all models as our main aim were the spatio-temporal patterns.

The methods developed in chapter 2 and 3 were not applied to compare the simulated and empirical patterns for all simulation models. Pattern comparison for the informal settlement model was based on metrics calculated for different time periods (as independent steps). Periods were dictated by the time stamps of the available empirical data. For the cholera model, the infection per

community, the time of first infection (order in which the communities were infected) and the duration of infection per community was monitored via boxplots. In addition the temporal patterns were identified via epidemic curves. Only for the analysis of the patterns generated by the pertussis model described in chapter 6 the SOM based Sammon's projection method was used.

The Sammon's projection was not usable for the informal settlements as this is not a diffusion process (but growth) and the number of time steps and spatial areas was too small. It could not be applied for the cholera simulation as the empirical data was not good enough to apply this method. The method requires a training dataset with multiple epidemics and only data for 2005 were available for this case study.

Table 7-4 simulated patterns

	Cholera model	**Pertussis model**	**Informal settlement model**
Spatial Patterns	Percentage of cases per community	**-**	Housing patterns via metrics
Temporal patterns	Epidemic curves for different drivers were evaluated	**-**	**-**
Spatio-temporal patterns	Start of infection and duration of infection per community	The trajectories of the Sammon's projection were used to compare the simulated and empirical spatio-temporal patterns	Spatial patterns were evaluated in discrete time steps.
Robustness of patterns	**-**	Was evaluated by comparing different epidemics of both empirical and simulated data. In both empirical and simulated data. Model produced a variety of patterns.	**-**

There are some advantages and disadvantages of the use of SOM based Sammon's projection to compare empirical and simulated data. One of these advantages is that the division of time is done based on characteristic states of the data itself (the time steps mapped back to the same SOM neuron). This is more flexible than the method applied for the Information settlements where the time periods were fixed. Another advantage is that a trajectory contains a number of characteristics that can be evaluated individually. Although the shape of the complete diffusion trajectory may be different for two epidemics, they may start in a similar way. The step from non-epidemic/endemic to

epidemic (state transition) can be linked to a particular spatial pattern, which can be very useful.

The method also has clear disadvantages, including the need for a long training datasets that includes multiple epidemics, and the fact that alignment of time series can a problem.

The largest disadvantage however, is the fact that a choice has to be made concerning the training datasets on which both empirical and simulated data have to be mapped back to allow comparison of simulated and empirical patterns. There are differences in level of detail, completeness, and noise between empirical and simulated patterns. Mapping empirical data back to a SOM trained based on empirical data or mapping simulated data back to a SOM trained with simulated data will always lead to a better result. However, this will not allow for direct comparison of the diffusion trajectories.

The Sammon's projection method still requires human interpretation of the similarity between the patterns. Comparison of vectors will always remain more difficult than comparison of numeric values like e.g. the metrics calculated for the informal settlements. The patterns found via the Sammon's projection consisted of a wide range of aspects like the number of neurons visited, the order in which the neurons were visited, but also back and forward transitions between two connected neurons etc.. When we get a better knowledge on which aspects of the patterns are really important, it might be possible to compare patterns in an automated way.

In this PhD thesis, the evaluation of patterns was based on model output, and the evaluation itself was not an integrated part in the model. In this way we demonstrated that the models could reproduce a variety of spatial temporal patterns. However, integration of the pattern evaluation in the model itself will greatly enhance the possibilities to judge if the model contains the correct level of complexity. It would allow for a direct evaluation of the impact model changes have on the results. Eventually we need to understand how "juggling" with different behaviour rules influences the spatial temporal patterns produced by a model. This can be done by specifying a large set of agent behaviour rules prior to the model development, and defining the output patterns that should be regenerated. By mixing all possible rules and including an evaluation mechanism in the model to evaluate the results of the combinations of rules with the pre-defined patterns, the essential rules can be deduced. Work in this line has been conducted by Wagner et al. (2015) using self-adaptive complexification, Wagner et al. (2013) by using evolutionary algorithms and the work by Stonedahl and Wilensky (2010) on evolutionary approach for finding emergent behaviour.

7.4 Conclusions

Patterns can be considered as footprints left behind by complex systems. Analysing empirical and simulated data to reveal patterns can lead to a better understanding of the systems. The main research objective of this PhD thesis is "*To design new approaches to the use of spatio-temporal patterns for building and validating ABMs*". For operational reasons, this objective was split into three sub-objectives related to the development of methods to identify spatio-temporal patterns, the application of these patterns to build and validate ABMs and the comparison of simulated and empirical data. The research questions that stem from each of these sub-objectives are answered below.

Answers to the research questions

Q1. How can time-series clustering be used to identify diffusion hierarchies and spatial and spatio-temporal self-similarity?

In chapter 2 we have shown how SOMs can be used to detect hierarchy in empirical disease time series data. Using a measles case study for Iceland, we identified diffusion from Reykjavik to the surrounding areas and an ordered diffusion pattern to the larger cities in the north of the island. The spatial patterns found using SOMs are closely linked to structures described by Cliff et al. (1981b) as being typical of hierarchical diffusion.

Besides hierarchy between spatial locations, we also identified similarities in time. This was done by combining the SOM with a Sammon's projection and connecting the neurons to form a diffusion trajectory that enables a visual comparison between epidemics. The Iceland case study revealed several epidemics with similar spatio-temporal diffusion patterns.

We also compared diffusion trajectories for pertussis epidemics in the Netherlands (chapter 6). We found less similarity between the diffusion trajectories of pertussis epidemics than in the measles case study. This can be due to the fact that pertussis is a less stable disease compared to measles, or because of the complex commuting patterns caused by the high population density of the Netherlands, or the fact that the Netherlands is not a closed system (i.e. the close interactions with Belgium and Germany can lead to re-infection). However, where the complete trajectories differed, all diffusion patterns started with a raise in infection in the west of the country. This indicates that our clustering-based approach can reveal critical spatial states at the beginning of epidemics, which may be very useful for applications like early warning systems.

Q2. *How should the data be spatially aggregated to maximize the likelihood of obtaining meaningful clusters?*

Disease surveillance data is often available in aggregated form due to privacy issues. Aggregation units often correspond to administrative boundaries that have no particular meaning for the diffusion processes. This level of aggregation might or might not be optimal to find diffusion patterns. When the aggregation is too fine, the level of detail may blur overall patterns, and when aggregation is too coarse (units are to large) more detailed patterns are invisible. When patterns are scaleless, aggregation to a range of different levels might be needed. With disease data becoming available at lower spatial scales (e.g. postal codes) it is important to find new re-aggregation methods that are based on the data itself (or important components) and aggregate to different scales.

For the Iceland epidemics, data aggregation was based on health units. The advantage of using data aggregation per administrative zones or health units is that this type of aggregation corresponds to the data collection. However the disadvantage is that these units do not follow naturally formed boundaries or homogeneous populations.

The Critical Community region (CCR) that was identified in chapter 3 is an aggregation to the area in which the disease is endemic. Input for the CCR were the percolation zones. We also used the percolation method for the SOMs discussed in chapter 6. This method is based on the hierarchy of the road network. As hierarchical disease diffusion patterns are closely linked to the distribution of the population and their commuting patterns, the aggregation of disease data to percolation zones seemed appropriate.

The application of the percolation method was straightforward and let to the identification of 16 urbanized zones in the Netherlands. The extracted zones were visually compared to the adult work related commuting data of 2013, and we found a good match between these datasets. A benefit of the percolation method is that it can be applied in situations where no mobility data is available because it only requires a road dataset as input. Another benefit of using the percolation method is that it identifies units that are larger than municipalities and smaller than provinces.

When using these zones in our SOM analysis, we noticed that in many cases several zones have the same infection level yet, these are not always the same zones. In the Netherlands, many large cities are so close together that they function as one urbanised unit in the disease diffusion process and not as individual cities. The Netherlands seems to be so urbanised that spatial unit of a municipality or city does not seem to be a useful spatial aggregation level for

analysing infectious disease any longer. Hierarchical diffusion theory still has individual cities as its main diffusion units yet, perhaps the highest hierarchical level should be replaced by an urbanised zone consisting of large groups of cities.

Q3. *How should time series be aligned to be able to compare epidemics using clustering(-based) approaches?*

The alignment of time series is important when comparing time series of different epidemics, especially when the comparison is based on the shape of the time series. This is because infection in a certain area may start earlier or later in different epidemics. In the SOM based approach introduced in chapter 2 the time series of the different health units were aligned manually. It would however be preferred to find a method that can automatically align time series.

In chapter 3, we made a comparison between two time series clustering methods (SBD and DTW). Although SBD is a global alignment method (aligns time series by moving one of them over the time axis) and DTW is a local alignment method (alignment of individual time series nodes) they both performed similarly. The two tested methods are not the only options, other alignment algorithms could be tested in combination with different clustering methods.

Combining the SBD and SOM method is not uncommon in other application domains. By first aligning the time series using the SBD followed by a SOM and Sammon's projection a combination could be made of all positive elements of both methods. Unfortunately none of the datasets used was large enough to test such a double clustering.

It is important to understand the characteristics of the input datasets but also the clustering goal, as this might change the preference for a certain clustering method completely. In this PhD thesis, we focused on shape-based methods, knowing that we compromise in the phase (timing) and the amplitude. As we were using input time series for percolation zones, our input areas were more homogeneous in population density than administrative units like municipalities would have been. For a large diversity in the input time-series, a method that combines several different characteristics (like the SOM) would be preferable. This however, would require more attention to be paid to the alignment of sub-sequences for different epidemics.

Q4. How can models that generate more detailed (vector based) simulation outputs help to compare simulated and empirical data?

In chapter 4 we present an ABM for informal settlements in Dar es Salaam. Where many urban simulation models are raster based, this is a vector based model that creates new houses as polygons and changes the shape of other houses (extension). By creating a vector based simulation the output of this model shows a great level of spatial detail. In the model, complexity was defined as spatial (form) realism. Is it necessary that the output of simulation models show a high level of spatial detail in order to detect complexity? In chapter 1 we defined different types of patterns. We noted that in urban simulation empirical data has a high level of spatial realism and this is not the case in disease data as it has no true spatial footprint. When we have empirical data with a high level of spatial realism, modelling output with the same level of detail can be useful. Yet, we should not forget that models are always an abstraction of reality. It is definitely so that spatial realism is necessary for the detection of some patterns. The patterns generated by the informal settlement model would not have been so detailed if we would have used a raster implementation at a coarser scale. However, this level of detail is not necessary or beneficial for all application domains and scales. A disease simulation like the cholera model presented in chapter 4 only produces point data of diseased individuals. This type of simulation is more independent of spatial realism.

Q5. How important is the use of spatio-temporal patterns, compared to temporal and spatial patterns, when building geographic ABMs?

In chapter 5 we developed a cholera model to test the hypothesis of runoff from dumpsites being the source of infection for the 2005 epidemic in Kumasi, Ghana. Results show that the model is able to reproduce epidemic curves (temporal patterns) that are realistic for cholera. The model is also able to reproduce the spatial pattern of infection per community. However, differences were found in the spatio-temporal patterns, especially for the during of infection in the different community. A later study showed that the introduction of risk perception in agents can overcome this mismatch. As agent-based models facilitate the simulation of spatial processes and social behaviour, we anticipate an increasing interest in including spatio-temporal patterns when modelling the behaviour of more dynamic agents. This is because agent behaviour changes the outcome of agent-based models over space and time.

Where the model was able to show that diffusion can be due to run-off from dumpsites, this cannot be proven conclusively. This would require the integration of alternative diffusion mechanisms in the model and comparison over multiple epidemics. The model does however show the importance of also including spatio-temporal patterns in the validation process. Both temporal and

spatial patterns did not show a mismatch between simulated and empirical data. The designer will have to include enough patterns to conclude that the model is valid. These patterns should be temporal, spatial and spatio-temporal and cover multiple epidemics to show the robustness of the pattern.

Q6. What are the factors that hinder validation of agent-based models based on spatio-temporal patterns and how can the comparison of empirical and simulated patterns be improved?

In chapter 1, we indicated that pattern oriented modelling requires the reproduction of multiple patterns (Grimm et al., 2005). In particular, the model should:

a. have multiple drivers that produce different patterns.
b. be able to produce spatial and spatio-temporal patterns that show self-similarity.
c. be able to produce output showing state transitions.

The models presented in chapters 4, 5 and 6 had multiple drivers that indeed produced different patterns. State transitions could also be simulated, for example in the pertussis model where a transition from an epidemic to an endemic state with periodic epidemics were simulated. The informal settlement model, could reproduce a variety of settlement patterns and the cholera model showed different patterns when simulating only environmental to human infection, compared to environment to human to environment infection.

Self-similarity was more difficult to reproduce. In space the small map extent of the model output was a problem to see spatial self-similarity. Self-similarity in time was also not evident, as the similarity between consecutive epidemics was not very large e.g. for pertussis in the Netherlands.

Different factors hinder the comparison of patterns found in empirical and simulated data, which is a prerequisite for validating models and improving our understanding of the system and/or process that generates those patterns. Those factors roughly belong to any of the following three categories/types:

1. The models – ABMs map extent is too small to show the scaleless behavior
2. The empirical data – time series should be long enough with enough detail, especially when applying machine learning techniques.
3. The methods – especially for spatio-temporal patterns
4. Our understanding of the patterns (especially spatio-temporal patterns)

The SOM based method can be further improved by introducing automatic alignment of the time series. Disadvantage of the method is that comparison

of patterns is only possible when mapping data to the same trained SOM. In general more methods should be developed based on machine learning techniques when sufficient training data is available. Integration of spatio-temporal pattern detection in the models itself can further enhance the usefulness of these patterns during the model building process. The SOM based method does allow for real time mapping, an important element when integrated in the model.

The Sammon's projection based on the SOM revealed that we lack understanding of which characteristics of the trajectory are (most) important. Trajectories for the empirical data showed that the patterns of pertussis are not robust, they differ considerably between the different epidemics, however all trajectories pass through a similar step at the beginning of the trajectory. This neuron corresponds to the state transition from non-epidemic to epidemic. This information is very important as it can help to understand which spatial pattern is at the core of a disease diffusion process and should be reproduced by the models.

Research achievements

The aim of this PhD thesis was three-fold: the development of methods for the detection of spatio-temporal patterns (a) the use of these patterns while designing agent-based models (b), and the comparison of patterns based on simulated and empirical data (c). Based on the studies presented in previous chapters, the following main research achievements were realized:

a. SOMs, SBD and DTW were used to cluster time series based on shape similarity. This approach was successful to reveal spatio-temporal patterns in diffusion processes when combining the clusters with the Sammon's projection. Two difficulties were observed: the alignment of time series of different epidemics and the interpretation of the different elements of the diffusion trajectories. A limitation of this research is that only a small number of clustering methods were tested and that not enough training data was available to test the methods for all case studies.
b. Spatio-temporal patterns provide important and useful information for the design and validation of ABMs. The use of spatio-temporal patterns is especially important when agents adapt their behavior or learn new behavior during the simulation. In this thesis I also identified a number of open issues when modelling complex systems. Some characteristics of complex systems were observable in our simulations, including the effect of multiple drivers and state transitions. However, the scaleless behavior of complex systems could not be reproduced due to the small map extent of the ABMs. This could either be resolved by enlarging the map extent or by looking for other solutions like multi-scale or linked models.

c. Comparison of patterns in empirical and simulated was possible using the Sammon's projection. This revealed useful information e.g. that all pertussis epidemics in the Netherlands move via the same critical neuron that can be linked to a state transition. Structural differences hinder the comparison of simulated and empirical data. Some of the identified spatio-temporal patterns were not robust. A limitation of the current work is that the pattern detection took place on the model output and not as an integrated part in the model itself. This limits that possibilities to fine-tune the model to observe direct changes in the patterns during simulation runs. The Sammon's projection is compared visually and not numerically, this may be difficult when implementing the pattern comparison in the model. Besides the SOM based method presented here, other (supervised) learning algorithms may be tested.

7.5 Directions for future work

To bring the use of spatio-temporal patterns to the next level we need to:

1. Generate multi-scale models that combine the advantages of ABMs with larger map extents.
2. Develop more methods to describe and detect spatio-temporal patterns and to compare empirical and simulated patterns.
3. Integrate pattern comparison methods directly into the models to allow for a better evaluation and comparison.

Multi-scale models

The weakest element in the check of patterns of complex systems is the check on scaleless behaviour. This is because the map extent of ABMs is often too small to reproduce patterns at different scales. Hybrid models, combining agent-based modelling with other types of models, like the pertussis model can overcome this issue. The way hybrid models integrate agent-based modeling differs considerably. For example in Bobashev and Epstein (Bobashev and Epstein, 2007) individual human agents are simulated and when the number of infected individuals reaches a treshold the model switches to a compartmental approach. In the pertussis model used in this reasearch the agent-based part is limited to agents representing health units that conduct interventions yet no individual disease agents are simulated.

Within the field of epidemiology, the focus seems to be shifting to hybrid models (Martin and Schlüter, 2015, Bobashev and Epstein, 2007) although there are some example of coupled models (Quang Nghi et al., 2016). A hybrid model combines an agent-based approach with another type of modeling e.g.

mathematical modeling within a single model. A coupled model combines different models into an overarching framework were feedbacks are allowed to pass between the models.

Both the coupled and the hybrid approach can lead to multi-scale models, simulating at different spatial and temporal scales. Multi-scale models are an emerging field in epidemiology (Garira, 2017). They are especially important for modelling infectious diseases because they spread over large areas and affect large populations (Garira, 2017). One of the advantages of multi-scale models mentioned by Garira is that they provide higher levels of detail and accuracy in characterizing infectious disease systems. Banos et al. (2015) state that multi-scale models are preferable because they are better for testing control strategies. This relates especially to ABMs that are able to model preventive strategies in the response behaviour of agents. Modelling response behaviour has a large impact on the simulated number of disease cases as was shown by Abdulkareem et al. (Abdulkareem et al., 2017). Little research has been done regarding the impact of human's cognition and preventive behaviour on disease diffusion (Funk et al., 2010).

Other reasons for focusing on multi-scale modeling might come from the importance of monitoring geographical disease incidence in relation to climate change. This is a relatively undeveloped field according to Rosenthal (2009) whereas climate change can affect the survival, reproduction and life cycle of pathogens, for example through changes on temperature and precipitation that impact hosts. Extreme weather events caused by climate change (e.g. floods, droughts) may also have a direct impact on disease transmission (Wu et al., 2016). Two types of studies are relevant in this context (Wu et al., 2016): studies that predict the impact of climate change on health and studies that identify how climate variables will alter the life cycle of pathogens and the transmission of infectious diseases. Especially the first type of studies require coupling disease and climate models, and this in turn requires performing multi-temporal analysis. Thus it is important to include options in disease models to experiment with various climate change scenarios (Manogaran and Lopez, 2017).

Methods to detect spatio-temporal patterns

The simulation of large map extents should be followed by a multi-scale analysis to detect and compare patterns at different scales. In this PhD thesis no multi-scale analysis were conducted. However, for a case study like the pertussis diffusion (based on hierarchical diffusion) this could have been done. The diffusion trajectories of individual percolation zones could have be analysed assuming that at micro scale a diffusion from a core city to the surrounding areas takes place. Even more interesting would have been to

study the disease diffusion at a European scale to reveal the role of the Netherlands in the diffusion of the disease. As stated before, this however, would mean that disease diffusion should be simulated for a larger population.

At present a system of disease monitoring, anomaly detection, and an intervention (e.g. vaccination) is practiced. However, we should ultimately seek to understand the drivers of the disease and prevent disease diffusion before it even starts. This includes a shift from control to preventive methods (Christaki, 2015, Liao et al., 2017). Data mining and big data approaches for early detection of infectious disease has been suggested by many authors (Han and Drake, 2016). This is partly triggered by the availability of new datatypes like social media, mobility and environmental data. Data mining of environment factors (and pathogen adjustment to these factors) is often linked to climate change. Further investigation of data mining and big data approaches that are applicable within the health domain is necessary if we want to use of all the available data in an integrated and useful way.

Many machine learning algorithms have been developed over the past years that might be useful in the detection of spatio-temporal patterns. This PhD only focused on unsupervised methods like the Self-Organizing Maps. However when enough training data is available supervised methods might proof to be very useful.

Integration of pattern comparison methods in models

In this PhD thesis we employed empirical validation of agent-based models, based on pattern similarity. Although POM as introduced by Grimm et al. (2005) improves model calibration and validation, the discussion on the validation of ABMs continues. Getting insights into the patterns a model is able to generate seems crucial for both calibration and validation purposes. The manual comparison of patterns generated by simulation models with empirical data is a tedious undertaking. Integration of pattern comparison directly in the model would greatly enhance the usefulness of the method. This will make the exploration of varying model parameters to create a range of patterns possible. Three directions are indicated by Lee et al. (2015) for the exploration of the solution space (the variety of patterns the model can simulate) and model calibration: meta-heuristics, optimization algorithms and machine learning. Surrogate models have been suggested by Lamperti et al. (2017). Different types of machine learning algorithms are used including genetic algorithms (Stonedahl and Wilensky, 2010, Calves and Hutzler, 2005), Bayesian Approaches (Korb et al., 2013) and hill-climbing (Stonedahl and Wilensky, 2010). The work presented in this thesis on SOM trajectories could be integrated in an ABM to be used to explore the variety of patterns the model is able to generate and to compare simulated patterns to empirical patterns.

References

ABBOTT, J. 2002. An analysis of informal settlement upgrading and critique of existing methodological approaches. *Habitat International,* 26**,** 303-315.

ABDULKAREEM, S., AUGUSTIJN, E.-W., MUSTAFA, Y. T. & FILATOVA, T. 2017. Integrating Spatial Intelligence for risk perception in an Agent Based Disease Model *Geocomputation 2017.* Leeds, UK.

ABDULKAREEM, S. A., AUGUSTIJN, E.-W., MUSTAFA, Y. T. & FILATOVA, T. 2018. Intelligent judgements over health risks in a spatial agent-based model. *International Journal of Health Geographics,* 17**,** 8.

ABUL, S. 2010. Environmental and health impact of solid waste disposal at Mangwaneni dumpsite in Manzini: Swaziland. *Journal of Sustainable Development in Africa,* 12**,** 64-78.

AGHABOZORGI, S., SEYED SHIRKHORSHIDI, A. & YING WAH, T. 2015. Time-series clustering – A decade review. *Information Systems,* 53**,** 16-38.

AL-AHMADI, K., SEE, L., HEPPENSTALL, A. & HOGG, J. 2009. Calibration of a fuzzy cellular automata model of urban dynamics in Saudi Arabia. *Ecological Complexity,* 6**,** 80-101.

ANDREW LIEBHOLD, WALTER D. KOENIG & BJØRNSTAD, O. N. 2004. Spatial Synchrony in Population Dynamics. *Annu. Rev. Ecol. Evol. Syst.,* 35**,** 467–90.

ANDRIENKO, G., ANDRIENKO, N., BAK, P., BREMM, S., KEIM, F., LANDESBERGER, T. V., POLITZ, C. & SCHRECK, T. 2010. A Framework for Using Self-Organzing Maps to Analyze Spatio-Temporal Patterns, Exemplified of Mobile Phone Usage. *Journal of Location based services,* 4**,** 200-221.

APOLLONI, A., POLETTO, C. & COLIZZA, V. 2013. Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic. *BMC Infectious Diseases,* 13.

ARCAUTE, E., MOLINERO, C., HATNA, E., MURCIO, R., VARGAS-RUIZ, C., MASUCCI, P., WANG, J. & BATTY, M. 2015. Hierarchical organisation of Britain through percolation theory. *arXiv:1504.08318*.

ARNAUD BANOS, A., CORSON, N., GAUDOU, B., LAPERRIÈRE, V. & COYREHOURCQ, S. R. 2015. The Importance of Being Hybrid for Spatial Epidemic Models: A Multi-Scale Approach. *Systems,* 3**,** 309-329.

ARNOLD, J. G., SRINIVASAN, R., MUTTIAH, R. S. & WILLIAMS, J. R. 1998. Large area hydrologic modelling and assessment part I: model development. *Journal of American Water Resources Association,* 34**,** 73-89.

ASOMANIN ANAMAN, K. & BERNICE NYADZI, W. 2015. Analysis of Improper Disposal of Solid Wastes in a Low-Income Area of Accra, Ghana. *Applied Economics and Finance,* 2**,** 66 - 75.

AUGUSTIJN, E.-W. & ZURITA-MILLA, R. 2013. Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International Journal of Health Geographics,* 12**,** 60.

AUGUSTIJN, P. W. M., ZURITA-MILLA, R. & VAN DER MAAS, N. Using self - organizing maps to analyse spatial temporal diffusion of infectious diseases. Geocomputation 2015, 2015 Dallas, Texas.

AYOMOH, M. K. O., OKE, S. A., ADEDEJI, W. O. & CHARLES-OWABA, O. E. 2008. An approach to tackling the environmental and health impacts of municipal solid waste disposal in developing countries. *Journal of Environmental Management,* 88**,** 108-114.

BAK, P., CHEN, K. & CREATZ, M. 1989. Self-Organized Criticality in the "Game of Life". *Nature,* 342**,** 780-782.

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,* 106**,** 21484-21489.

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**,** 132-145.

BARROS, J. X. 2004. *Urban Growth in Latin American Cities, Exploring urban dynamics through agent-based simulation.* PhD, University of London.

BARTLETT, M. S. 1957. Measles periodicity and community size. *Journal of the Royal Statistical Society. Series A (General),* 120**,** 48-70.

BARTLETT, M. S. 1960. The Critical Community Size of Measles in the United States. *Journal of the Royal Statistical Society. Series A (General),* 123**,** 37-44.

BASARA, H. & YUAN, M. 2008. Community health assessment using self-organizing maps and geographic information systems. *International Journal of Health Geographics,* 7**,** 67.

BATTY, M. 2005. *Cities and complexity : understanding cities with cellular automata, agent - based models, and fractals,* Cambridge etc., MIT.

BATTY, M. 2012. Building a science of cities. *Cities,* 29, Supplement 1**,** S9-S16.

BELIK, V., GEISEL, T. & BROCKMANN, D. 2011. Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. *Physical Review X,* 1.

BENENSON, I. 2004. Agent-Based Modeling: From Individual Residential Choice to Urban Residential Dynamics. *In:* M. F. GOODCHILD, D. G. J. E. (ed.) *Spatially Integrated Social Science: Examples in Best Practice.* Oxford University

BENENSON, I. & TORRENS, P. M. 2004. Geosimulation: Object-based modeling of urban phenomena. *Computers, Environment and Urban Systems,* 28**,** 1-8.

BERTUZZO, E., AZAELE, S., MARITAN, A., GATTO, M., RODRIGUEZ-ITURBE, I. & RINALDO, A. 2008. On the space-time evolution of a cholera epidemic. *Water Resources Research,* 44.

BERTUZZO, E., CASAGRANDI, R., GATTO, M., RODRIGUEZ-ITURBE, I. & RINALDO, A. 2009. On spatially explicit models of Cholear epidemics. *Journal of the Royal Society Interface,* 7**,** 321-333.

BHASKARAN, K., GASPARRINI, A., HAJAT, S., SMEETH, L. & ARMSTRONG, B. 2013. Time series regression studies in environmental epidemiology. *International Journal of Epidemiology,* 42**,** 1187-1195.

BJORNSTAD, O. N., IMS, R. A. & LAMBIN, X. 1999. Spatial population dynamics: analyzing patterns and processes of population synchrony. *Trends in Ecology & Evolution,* 14**,** 427-432.

BLUMBERG, S. & LLOYD-SMITH, J. O. 2013. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. *PLOS Computational Biology,* 9**,** e1002993.

BOBASHEV, G. V. & EPSTEIN, J. M. 2007. A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches. *In:* HENDERSON, S. G., BILLER, B., HSIEH, M.-H., SHORTLE, J., TEW, J. D. & BARTON, R. R. (eds.) *Proceedings of the 2007 Winter Simulation Conference.*

BOONSTRA, J. 1994. Estimating Peak Runoff Rates *In:* RITZEMA, H. (ed.) *Drainage principles and applications.*

BROCKMANN, D., HUFNAGEL, L. & GEISEL, T. 2006. The scaling laws of human travel. *Nature,* 439**,** 462-465.

BROWN, D. G., ROBINSON, D. T., AN, L., NASSAUER, J. I., ZELLNER, M., RAND, W., RIOLO, R., PAGE, S. E., LOW, B. & WANG, Z. 2008. Exurbia from the bottom-up: Confronting empirical challenges to characterizing a complex system. *Geoforum,* 39**,** 805-818.

CALVES, B. & HUTZLER, G. Automatic Tuning of Agent-Based Models using Genetic Algorithms. MABS 2005: Proceedings of the 6th International Workshop on Multi-Agent-Based Simulation, 2005.

CAPASSO, V. & PAVERI-FONTANA, S. L. 1979. A mathematical model for the 1973 cholera epidemic in the european mediterranean region. *Rev Epidem et Sante Pub,* 27**,** 121-132.

CAZELLES, B., CHAVEZ, M., MAGNY, G. C. D., GUEGAN, J.-F. & HALES, S. 2007. Time-dependent spectral analysis of epidemiological time-series with wavelets. *J. R. Soc. Interface,* 4**,** 625–636.

CGIAR. 2010. *srtm* [Online]. Available: www.cgiar.org 2010].

CHEN, L. C., KAMINSKY, B., TUMMINO, T., CARLEY, K. M., CASMAN, E., FRIDSMA, D. & YAHJA, A. Aligning simulation models of smallpox outbreaks. *In:* CHEN, H. Z. D. D. M. R. L. J., ed., 2004. 1-16.

CHEN, Y. 2015. A New Methodology of Spatial Cross-Correlation Analysis. *PLOS One,* 10**,** e0126158.

CHEN, Y. & ZHOU, Y. 2008. Scaling laws and indications of self-organized criticality in urban systems. *Chaos, Solitons & Fractals,* 35**,** 85-98.

CHOISY, M. & ROHANI, P. 2012. Changing spatial epidemiology of pertussis in continental USA. *Proceedings of the Royal Society B: Biological Sciences,* 279**,** 4574-4581.

CHRISTAKI, E. 2015. New technologies in predicting, preventing and controlling emerging infectious diseases. *Virulence,* 6**,** 558-565.

CLARAMUNT, C. & THÉRIAULT, M. 1996. Toward semantics for modelling spatio-temporal processes within GIS *Seventh International Symposium on Spatial Data Handling,.*

CLIFF, A. D., HAGGET, P. & SMALLMAN-RAYNOR, M. 2008. An exploratory method for estimating the changing speed of epidemic waves from historical data. *International Journal of Epidemiology,* 37**,** 106-112.

CLIFF, A. D., HAGGETT, P., ORD, J. K. & VERSEY, G. R. 1981a. Identifying diffusion processes - the historical record. *In:* FARMER, B. H., GROVE, A. T., HAGGETT, P. & WRIGLEY, E. A. (eds.) *Spatial Diffusion. An Historical Geogrphy of Epidemics in an Island Community.* New York: Press Syndicate of the University of Cambridge.

CLIFF, A. D., HAGGETT, P., ORD, J. K. & VERSEY, G. R. 1981b. *Spatial Diffusion, An Historical Geography of Epidemics in an Island Community,* Cambridge, USA, Press Syndicate of the University of Cambridge.

CLIFF, A. D., HAGGETT, P. & SMALLMAN-RAYNOR, M. 2009a. The changing shape of island epidemics: historical trends in Icelandic infectious disease waves, 1902–1988. *Journal of Historical Geography,* 35**,** 545-567.

CLIFF, A. D., HAGGETT, P. & SMALLMAN-RAYNOR, M. 2009b. The changing shape of island epidemics:historical trends in Icelandic infectious disease waves, 1902-1988. *Journal of Historical Geography,* 35**,** 545-567.

CODEÇO, C. 2001. Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. *BMC Infectious Diseases,* 1**,** 1.

COLIZZA, V., BARRAT, A., BARTHELEMY, M., VALLERON, A.-J. & VESPIGNANI, A. 2007. Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLOS Medicine,* 4**,** e13.

CROOKS, A. T. 2010. Constructing and implementing an agent-based model of residential segregation through vector GIS. *International Journal of Geographical Information Science,* 24**,** 661-675.

CURRIERO, F. C., PATZ, J. A., ROSE, J. B. & LELE, S. 2001. The Association Between Extreme Precipitation and Waterborne Disease Outbreaks in the United States, 1948–1994. *American Journal of Public Health,* 91**,** 1194-1199.

DALZIEL, B. D., POURBOHLOUL, B. & ELLNER, S. P. 2013. Human mobility patterns predict divergent epidemic dynamics among cities. *Proceeding of the Royal Society B,* 280.

DANIEL A. BAGAH, ISSAKA K. OSUMANU & OWUSU-SEKYERE, E. 2015. Persistent "Cholerization" of Metropolitan Accra, Ghana: Digging into the Facts. *American Journal of Epidemiology and Infectious Disease,* 3**,** 61-69.

DE BRUIJN, C. A. 1987. Monitoring a large squatter area in Dar es Salaam with Sequential Aerial Photography. *ITC journal,* 3**,** 233-238.

DE MELKER, H. E., VERSTEEGH, F. G. A., SCHELLEKENS, J. F. P., TEUNIS, P. F. M. & KRETZSCHMAR, M. 2006. The incidence of Bordetella pertussis infections estimated in the population from a combination of serological surveys. *Journal of Infection,* 53**,** 106-113.

DEVAS, N. & KORBOE, D. 2000. City governance and poverty: the case of Kumasi. *Environment & Urbanization,* 12**,** 124-136.

DNA 2010. Low power voltage affecting water supply in kumasi. GhanaHomePage/NewsArchive/artikel.php?ID=177728

DOLDERSUM, T. 2013. *The role of water in cholera diffusion. Improvements of a cholera diffusion model for Kumasi, Ghana.* MSc, University of Twente, The Netherlands.

DOMMAR, C. J., LOWE, R., ROBINSON, M. & RODO, X. 2014. An agent-based model driven by tropical rainfall to understand the spatio-temporal heterogeneity of a chikungunya outbreak. *Acta Tropica,* 129**,** 61-73.

DSDC 2005. REPUBLIC OF GHANA URBAN POVERTY REDUCTION PROJECT (POVERTY II) - APPRAISAL REPORT. African Development Fund.

EISENBERG, J. N. S., DESAI, M. A., LEVY, K., BATES, S. J., LIANG, S., NAUMOFF, K. & SCOTT, J. C. 2007. Environmental Determinants of Infectious Disease: A Framework for Tracking Causal Links and Guiding Public Health Research. *Environmental Health Perspectives,* 115**,** 1216-1223.

EPSTEIN, J. M., GOEDECKE, D. M., YU, F., MORRIS, R. J., WAGENER, D. K. & BOBASHEV, G. V. 2007. Controlling Pandemic Flu: The Value of International Air Travel Restrictions. *PLOS One,* 2**,** e401.

FEACHEM, R., BRADLEY, D. J., GARELICK, H. & MARA, D. D. 1983. Vibrio cholerae and cholera. *Sanitation and disease. Health aspects of excreta and waastewater management.*: John Wiley & Sons.

FUNK, S., SALATHÉ, M. & JANSEN, V. A. A. 2010. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of the Royal Society Interface,* 7**,** 1247.

G. ANDRIENKO, N. ANDRIENKO, S. BREMM, T. SCHRECK, T. VON LANDESBERGER, P. BAK & D.KEIM. Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. *In:* G. MELANÇON, T. MUNZNER & WEISKOPF, D., eds. Eurographics/ IEEE-VGTC Symposium on Visualization 2010, 2010.

GARIRA, W. 2017. A complete categorization of multiscale models of infectious disease systems. *Journal of Biological Dynamics,* 11**,** 378-435.

GATRELL, A. C., BAILEY, T. C., DIGGLE, P. J. & ROWLINGSON, B. S. 1996. Spatial poin pattern analysis and its application in geographical epidemiology. *Trans Inst Br Geogr***,** 256-274.

GAUDART, J., REBAUDET, S., BARRAIS, R., BONCY, J., FAUCHER, B., PIARROUX, M., MAGLOIRE, R., THIMOTHE, G. & PIARROUX, R. 2013. Spatio-Temporal Dynamics of Cholera during the First Year of the Epidemic in Haiti. *PLoS Negl Trop Dis,* 7**,** e2145.

GERMANN, T. C., KADAU, K., LONGINI, I. M. & MACKEN, C. A. 2006. Mitigation strategies for pandemic influenza in the United States. 103**,** 5935-5940.

GHASSEMPOUR, S., GIROSI, F. & MAEDER, A. 2014. Clustering Multivariate Time Series Using Hidden Markov Models. *Int. J. Environ. Res. Public Health* 11**,** 2741-2763.

GLADWELL, M. 2000. *The Tipping Point: How Little Things Can Make a Big Difference,* Boston, New York, London, Little, Brown & Company.

GRAD, Y. H., MILLER, J. C. & LIPSITCH, M. 2012. Cholera Modelling: Challenges to Quantitative Analysis and Predicting the Impact of Interventions. *Epidemiology,* 23**,** 523-530.

GRENFELL, B. T., BJORNSTAD, O. N. & KAPPEY, J. 2001. Travelling waves and spatial hierarchies in measles epidemics. *Nature,* 414**,** 716-723.

GRENFELL, B. T., PYBUS, O. G., GOG, J. R., WOOD, J. L. N., DALY, J. M., MUMFORD, J. A. & HOLMES, E. C. 2004. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science,* 303.

GRIMM, V., BERGER, U., DEANGELIS, D. L., POLHILL, J. G., GISKE, J. & RAILSBACK, S. F. 2010. The ODD protocol: A review and first update. *Ecological Modelling,* 221**,** 2760–2768.

GRIMM, V., REVILLA, E., BERGER, U., JELTSCH, F., MOOIJ, W. M., RAILSBACK, S. F., THULKE, H.-H., WEINDER, J., WIEGAND, T. & DEANGELIS, D. L. 2005. Pattern-Oriented Modeling of Agent-based Complex Systems: Lessons from Ecology. *Science,* 310**,** 987-991.

GSS 2008. Ghana living standards survey of the fifth round. Accra, Ghana: Ghana Statistical Service.

GSS 2012. 2010 Population & Housing Census Summary report of Final results. Accra: Ghana Statistical Service.

HAGERSTRAND, T. 1953. *Innovation Diffusion as a Spatial Process,* Chicago, University of Chicago Press.

HAGGETT, P. 1976. Hybridizing Alternative models of an epidemic diffusion process. *Economic geography,* 52**,** 136-146.

HAMEL, L. & BROWN, C. W. 2011. Improved Interpretability of the Unified Distance Matrix with Connected Components. *In:* STAHLBOCK, R. (ed.) *7th International Conference on Data Mining (DMIN'11).* Las Vegas: CSREA Press.

HAN, B. A. & DRAKE, J. M. 2016. Future directions in analytics for infectious disease intelligence. *EMBO reports,* 17**,** 785-789.

HARGROVE, W. W., HOFFMAN, F. M. & HESSBURG, P. F. 2006. Mapcurves: a quantitative method for comparing categorical maps. *Journal of geographical systems,* 8**,** 187-208.

HARRIS, J. B., LAROCQUE, R. C., QADRI, F., RYAN, E. T. & CALDERWOOD, S. B. 2012. Cholera. *Lancet,* 379**,** 2466-2476.

HARTLEY, D. M., MORRIS, J. G. & SMTIH, D. L. 2006. Hyperinfectivity: A Critical Element in the Ability of V. cholerae to cause epidemics? *PLoS Med,* 3**,** 63-69.

HATNA, E. & BENENSON, I. 2015. Combining segregation and integration: Schelling model dynamics for heterogeneous population. *Journal of Artificial Societies and Social Simulation,* 18**,** 15.

HAYDON, D. T., CLEAVELAND, S., TAYLOR, L. H. & M.K., L. 2002. Identifying Reservoirs of Infection: A Conceptual and Practical Challenge. *Emerging Infectious Diseases,* 8**,** 1468-1473.

HEROLD, M., SCEPAN, J. & CLARKE, K. C. 2002. The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. *Environment and Planning A,* 34**,** 1443-1458.

HIMBERG, J. Enhancing SOM-based data visualization by linking different data projections. International Symposium on Intelligent Data Engineering and Learning (IDEAL), 1998 Hong Kong. 427-434.

HOEN, A. G., HLADISH, H. J., EGGO, R. M., LENCZNER, M., BROWNSTEIN, J. S. & MEYERS, L. A. 2015. Epidemic Wave Dynamics Attributable to Urban Community Structure: A Theoretical Characterization of Disease Transmission in a Large Network. *J Med Internet Res,* 17.

HOLMNER, Å., MACKENZIE, A. & KRENGEL, U. 2010. Molecular basis of cholera blood-group dependence and implications for a world characterized by climate change. *FEBS Letters,* 584**,** 2548-2555.

HSU, W.-T., MORI, T. & SMITH, T. E. 2014. Spatial Patterns and Size Distributions of Cities. *Research Collection School Of Economics***,** 1-48.

HUANG, Q., PARKER, D. C., FILATOVA, T. & SUN, S. 2013. A review of urban residential choice models using agent-based modeling. *Environment and Planning B: Planning and Design,* 40.

HUFNAGEL, L., BROCKMANN, D. & GEISEL, T. 2004. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academic Society USA,* 101**,** 15124-15129

JOSÉ, M. V. & BISHOP, R. F. 2003. Scaling properties and symmetrical patterns in the epidemiology of rotavirus infection. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 358**,** 1625-1641.

KAJAGI, W. L. H. 1982. *Unplanned Settlements Improvement: Aerial Photography as a Data Source.* MSc, ITC.

KEELING, M. J. & GRENFELL, B. T. 1997. Disease Extinction and Community Size: Modelling the Persistence of Measles. *Science,* 275**,** 65-67.

KEOGH, E., LIN, J. & FU, A. 2005. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. *Proceedings of the Fifth IEEE International Conference on Data Mining.*

KEOGH, E. & RATANAMAHATANA, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge and Information Systems,* 7**,** 358-386.

KIRONDE, J. & RUGAIGANISA, D. 2002. Urban land Management Regularization and Local Development Policies in Tanzania. Dar es Salaam: UCLAS.

KOCABAS, V. & DRAGICEVIC, S. 2009. Agent-based model validation using Bayesian networks and vector spatial data. *Environment and planning B: Planning and design,* 36**,** 787-801.

KOHONEN, T. 2001. *Self-Organizing Maps,* New York, USA: Springer-Verlag.

KOK, K., A., F., VELDKAMP, A. & VERBURG, P. H. 2001. A method and application of multi-scale validation in spatial land use models. *Agriculture Ecosystems and Environment,* 85**,** 223-238.

KORB, K. B., GEARD, N. & DORIN, A. 2013. A Bayesian Approach to the Validation of Agent-Based Models. *In:* TOLK, A. (ed.) *Ontology, Epistemology, and Teleology for Modeling and Simulation: Philosophical Foundations for Intelligent M&S Applications.* Berlin, Heidelberg: Springer Berlin Heidelberg.

KOUA, E. & KRAAK, M.-J. 2004. Geovisualization to support the exploration of large health and demographic survey data. *International Journal of Health Geographics,* 3**,** 12.

KWASI OWUSU BOADI & KUITUNEN, M. 2005. Environmental and Health Impacts of Household Solid Waste Handling and Disposal Practices in Third World Cities: The Case of the Accra Metropolitan Area, Ghana. *Journal of Environmental Health,* 68**,** 32-36.

KYESSI, S. A. 1993. *Analysis of the physical and socio-economic characteristics of manzesse squatter settlement Dar es Salaam.* MSc, ITC.

LAMPERTI, F., ROVENTINI, A. & SANI, A. 2017. Agent-Based Model Calibration using Machine Learning Surrogates. *arXiv:1504.08318,* 3**,** 32 pages.

LEE, B. Y., BEDFORD, V. L., ROBERTS, M. S. & CARLEY, K. M. 2008. Virtual epidemic in a virtual city: simulating the spread of influenza in a US metropolitan area. *Translational Research,* 151**,** 275-287.

LEE, J.-S., FILATOVA, T., LIGMANN-ZIELINSKA, A., HASSANI-MAHMOOEI, B., STONEDAHL, F., LORSCHEID, I., VOINOV, A., POLHILL, J. G., SUN, Z. & PARKER, D. C. 2015. The Complexities of Agent-Based Modeling Output Analysis. *Journal of Artificial Societies and Social Simulation,* 18**,** 4.

LEVINE, E. & DOMANY, E. 2001. Resampling Method For Unsupervised Estimation Of Cluster Validity. *Neural Computation,* 13**,** 2573-2593.

LIAO , T. W. 2005. Clustering of time series data-a survey. *Pattern Recogn.,* 38**,** 1857-1874.

LIAO, Y., XU, B., WANG, J. & LIU, X. 2017. A new method for assessing the risk of infectious disease outbreak. *Sci. Rep.,* 7.

LIEBHOLD, A., KOENIG, W. D. & BJORNSTAD, O. N. 2004. Spatial synchrony in population dynamics. *Annual Review of Ecology Evolution and Systematics,* 35**,** 467-490.

LINARD, C., PONÇON, N., FONTENILLE, D. & LAMBIN, E. F. 2008. A multi-agent simulation to assess the risk of malaria re-emergence in southern France. *Ecological Modelling,* In Press, Corrected Proof.

LIU, Y. 2009. *Modelling urban development with geographical information systems and cellular automata,* Boca Raton Fla, CRC Press.

LUQUERO, F. J., BANGA, C. N., REMARTINEZ, D., PALMA, P. P., BARON, E. & GRAIS, R. F. 2011. Cholera Epidemic in Guinea-Bissau (2008): The importance of "Place". *PLOS* volume 6**,** e19005.

MANOGARAN, G. & LOPEZ, D. 2017. *Disease Surveillance System for Big Climate Data Processing and Dengue Transmission*.

MARI, L., BERTUZZO, E., RIGHETTO, L., CASAGRANDI, R., GATTO, M., I, R.-I. & RINALDO, A. 2012. Modelling cholera epidemics: the role of waterways, human mobility and sanitation. *Journal of the Royal Society Interface,* 9**,** 376-388.

MARTIN, R. & SCHLÜTER, M. 2015. Combining system dynamics and agent-based modeling to analyze social-ecological interactions—an example from modeling restoration of a shallow lake. *Frontiers in Environmental Science,* 3.

MATTHEWS, R., GILBERT, N., ROACH, A., POLHILL, J. & GOTTS, N. 2007. Agent-based land-use models: a review of applications. *Landscape Ecology,* 22**,** 1447-1459.

MCMICHAEL, A. J. 2000. The urban environment and health in a world of increasing globalization: issues for developing countries. Bulletin of the World Health Organization.

MEADE, M. & EMCH, M. 2010. *Medical Geography*.

METCALF, C. J. E., HAMPSON, K., TATEM, A. J., GRENFELL, B. T. & BJØRNSTAD, O. N. 2013. Persistence in Epidemic Metapopulations: Quantifying the Rescue Effects for Measles, Mumps, Rubella and Whooping Cough. *PLOS One,* 8**,** e74696.

MIKLER, A., VENKATACHALAM, S. & RAMISETTY-MIKLER, S. 2007. Decisions under uncertainty: a computational framework for quantification of policies addressing infectious disease epidemics. *Stochastic Environmental Research and Risk Assessment,* 21**,** 533-543.

MNISZEWSKI, S., DEL VALLE, S., STROUD, P., RIESE, J. & SYDORIAK, S. 2008. Pandemic simulation of antivirals+school closures: buying time until strain-specific vaccine is available. *Computational & Mathematical Organization Theory,* 14**,** 209-221.

MOECKEL, R., SPIEKERMANN, K. & WEGENER, M. Creating a synthetic Population. 8th Internatioanl Conference on Computers in Urban Planning and Mangement (CUPUM), 2003 Sendai: Center for Northeast Asian Studies. 1-8.

MORENO, N., WANG, F. & MARCEAU, D. J. 2009. Implementation of a dynamic neighborhood in a land-use vector-based cellular automata model. *Computers, Environment and Urban Systems,* 33**,** 44-54.

MORSE, S. S. 1995. Factors in the emergence of infectious diseases. *Emerging Infectious Diseases,* 1**,** 7-15.

OBIRI-DANSO, K., WEOBONG, C. A. A. & JONES, K. 2005. Aspects of health-related microbiology of the Subin, an urban river in Kumasi, Ghana. *Journal of Water and Health* 3**,** 69-76.

OSEI, F. & DUKER, A. 2008. Spatial dependency of V. cholera prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *International Journal of Health Geographics,* 7**,** 62.

OSEI, F. B. 2010. *Spatial statistics of epidemic data: the case of cholera epidemiology in Ghana.* Ph.D. PhD thesis, University of Twente.

OSEI, F. B., DUKER, A. A., AUGUSTIJN, E.-W. & STEIN, A. 2010. Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi, Ghana. *International Journal of Applied Earth Observation and Geoinformation,* 12 331-339.

PAPARRIZOS, J. & GRAVANO, L. 2015. k-Shape: Efficient and Accurate Clustering of Time Series. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data.* Melbourne, Victoria, Australia.

PARKER, D. C., EVANS, T. & MERETSKY, V. 2001. Measuring emergent properties of agent-based landcover/landuse models using spatial metrics. *Seventh Annual Conference of the International Society for Computational Economics.* New Haven, CT.

PATZ, J. A., DASZAK, P., TABOR, G. M., AGUIRRE, A. A., PEARL, M., EPSTEIN, J., WOLFE, N. D., KILPATRICK, A. M., FOUFOPOULOS, J., MOLYNEUX, D., BRADLEY, D. J. & MEMBERS OF THE WORKING GROUP ON LAND USE CHANGE DISEASE, E. 2004. Unhealthy Landscapes: Policy Recommendations on Land Use Change and Infectious Disease Emergence. *Environmental Health Perspectives,* 112**,** 1092-1098.

PENROSE, K., CASTRO, M. C. D., WEREMA, J. & RYAN, E. T. 2010. Informal Urban Settlements and Cholera Risk in Dar es Salaam, Tanzania. *PLoS Negl Trop Dis,* 4**,** e631.

PETRAITIS 2013. *Multiple stable states in natural ecosystems*, Oxford University Press, 2013.

PHILIPPE, P. 1999. The Scale-invariant Spatial Clustering of Leukemia in San Francisco. *Journal of Theoretical Biology,* 199**,** 371-381.

PISSOURIOS, I., LAFAZANI, P., SPYRELLIS, S., CHRISTODOULOU, A. & MYRIDIS, M. 2012. The Use of Point Pattern Statistics in UrbanAnalysis. *In:* GENSEL, J., JOSSELIN, D. & VANDENBROUCKE, D. (eds.) *Bridging the Geographic Information Sciences: International AGILE'2012 Conference, Avignon (France), April, 24-27, 2012.* Berlin, Heidelberg: Springer Berlin Heidelberg.

POLHILL, J. G., PARKER, D., BROWN, D. & GRIMM, V. 2008. Using the ODD Protocol for Describine Three Agent-Based Social Simulation Models of Land-Use Change. *Journal of Artificial Societies and Social Simulation,* 11**,** 3.

PONTIUS JR., R. G. & SCHNEIDER, L. C. 2001. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture Ecosystems and Environment,* 85**,** 239-248.

POST, J. 1999. The Problems and Potentials of Privatising Solid Waste Managment in Kumasi, Ghana. *Habitat International,* 2**,** 201-215.

RAFIEE, R., MAHINY, A. S., KHORASANI, N., DARVISHSEFAT, A. A. & DANEKAR, A. 2008. Simulating urban growth in Mashad City, Iran through the SLEUTH model (UGM). *Cities,* 26**,** 19-26.

RAILSBACK, S. F. & GRIMM, V. 2012. *Agent-Based and Individual-Based Modeling: a practical introduction,* Princeton and Oxford, Princeton University Press.

RAMADHANI, S. H. 2007. *Effects of tenure regularization programme on building investment in Manzese ward in Dar es Salaam.* MSc, ITC.

RAND, W. M. 1971. Objective criteria for evaluation of clustering methods. *Journal of American Statistical Association*.

REGO, R. F., MORAES, L. R. S. & DOURADO, I. 2005. Diarrhoea and garbage disposal in Salvador, Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 99**,** 48-54.

RHODES, C. J., JENSEN, H. J. & ANDERSEN, R. M. 1977. On the critical behaviour of simple epidemics. *Proceedings: Biological Sciences,* 264**,** 1639-1646.

RICKLES, D., HAWE, P. & SHIELL, A. 2007. A simple guide to chaos and complexity. *Journal of Epidemiology and Community Health,* 61**,** 933-937.

RILEY, S. 2007. Large-Scale Spatial-Transmission Models of Infectious Disease. *Science,* 316**,** 1298-1303.

RISCASSI, A. & SCHAFFRANEK, R. 2003. Flow Velocity Water Temperature, and Conductivity in Shark River Slough, Everglades National Park, Florida: August 2001-June 2002.: U.S. Department of the Interior, Reston.

ROSENTHAL, J. 2009. Climate Change and the Geographic Distribution of Infectious Diseases. *EcoHealth,* 6**,** 489-495.

SARDA-ESPINOSA, A. 2017. Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance.

SCHELLING, T. C. 1969. Models of segregation. *American Economic Review,,* 59**,** 488–493.

SCS 1986. Urban hydrology for small watersheds. United States Department of Agriculture, Soil Conservation Service Engineering Division.

SEDAS, V. T. P. 2007. Influence of environmental factors on thepresence of Vibrio cholerae in the marine environment: a climate link. *The Journal of Infection in Developing Countries,* 1**,** 224-241.

SEEK, N. H. 1983. Adjusting housing consumption: Improve or move. *Urban Studies,* 20**,** 455-469.

SHAW, E. M., BEVEN, K. J., CHAPELL, N. A. & LAMB, R. 2011. *Hydrology in Practice*, Spon Press, London and New York.

SHEKHAR, S., JIANG, Z., ALI, R., EFTELIOGLU, E., TANG, X., GUNTURI, V. & ZHOU, X. 2015. Spatiotemporal Data Mining: A Computational Perspective. *ISPRS International Journal of Geo-Information,* 4**,** 2306.

SHEUYA, S. 2009. Urban Poverty and Housing transformation in informal settlements, the case of Dar es Salaam, Tanzania. *International Development Planning,* 31**,** 81-108.

SIETCHIPING, R. Prospective Slum Policies: Conceptualization and Implementation of a Proposed Informal Settlement Growth Model. third urban research symposium on "Land Development, urban policy and poverty reduction", 4-6 April 2005 2005 Brasilia, Brazil.

SIETCHIPING, R. 2008. *Predicting and preventing slum growth: Theory method implementation and evaluation,* Saarbrücken, VDM-Verl. Müller.

SLIUZAS, R. 1988. *Problems in Monitoring the Growth of a Squatter Settlement: The housing Process in Manzese, Dar es Salaam.* MSc, ITC.

SLIUZAS, R. 2004. *Managing Informal Settlements, A Study using geo-information in Dar es Salaam.* PhD, Utrecht University and ITC.

SMYTHIES, J. R. 1957. A prelinimary analysis of the stroboscopic patterns. *Nature,* 4558**,** 523-524.

SOBREIRA, F. 2005. Modelling Favelas: Heuristic Agent Based Models for Squatter Settlements Growth and Consolidation. *In:* SCHRENK, E. (ed.) *10th International Conference on Information & Communication Technologies (ICT) in Urban Planning and Spatial Development and Impacts of ICT on Physical Space.* Vienna.

SOBREIRA, F. & GOMES, M. 2001. The Geometry of Slums: boundaries, packing and diversity. *UCL Working Papers Series.*

STANILOW, K. 2009. Typo-morphology and object-based automata: methodological advances towards a more accurate modelling of urban growth patterns. *Proceedings of the 11th international conference on computers in urban planning and urban management.* Hong Kong.

STONEDAHL, F. & WILENSKY, U. Finding Forms of Flocking: Evolutionary Search in ABM Parameter Spaces. *In:* VAN DER HOEK, W., KAMINKA, G. A., LESPÉRANCE, Y., LUCK, M. & SEN, S., eds. The Eleventh International Workshop on Multi-Agent-Based Simulation, May 10-14, 2010 2010 Toronto Canada. 59-73.

STRAATMAN, B., WHITE, R. & ENGELEN, G. 2004. Towards an automatic calibration procedure for constrained cellular automata. *Computers, Environment and Urban Systems,* 28**,** 149-170.

SUDHIRA, H. S., RAMACHANDRA, T. V., WYTZISK, A. & JEGANATHAN, C. 2005. Framework for integration of agent-based and cellular automata models for dynamic geospatial simulations. bangalore: Centre of Ecological sciences, Indian Institute of Science.

TAMERIUS, J. D., WISE, E. K., UEJIO, C. K., MCCOY, A. L. & COMRIE, A. W. 2007. Climate and human health: synthesizing environmental complexity and uncertainty. *Stochastic Environmental Research and Risk Assessment,* 21**,** 601-613.

TASDEMIR, K. & MERENYI, E. 2012. SOM-based topology visualisation for interactive analysis of high-dimensional large datasets. *In:* VILLMANN, T. & SCHLEIF, F.-M. (eds.) *Machine Learning Reports.* Mittweida and Bielefeld: University of Applied Sciences Mittweida, Dept. of Mathematics/Physics/Computer Sciences and University of Bielefeld, CITEC - AG Theoretical Computer Science.

TIGNOR, N., WANG, P., GENES, N., ROGERS, L., HERSHMAN, S. G., SCOTT, E. R., ZWEIG, M., CHAN, Y.-F. Y. & SCHADT, E. E. 2017. Methods for clustering timer series data acquired from mobile health Apps. *Pacific Symposium on Biocomputing 2017.*

TIZZONI, M., BAJARDI, P., DECUYPER, A., KON KAM KING, G., SCHNEIDER, C. M., BLONDEL, V., SMOREDA, Z., GONZÁLEZ, M. C. & COLIZZA, V. 2014. On the Use of Human Mobility Proxies for Modeling Epidemics. *PLoS Comput Biol,* 10**,** e1003716.

TJALMA, S. 2016. *AN AGENT-BASED MODEL TO COMPARE VACCINATION STRATEGIES FOR PERTUSSIS IN THE NETHERLANDS.* MSc, University of Twente.

UNITED NATIONS 2009. The millennium Development Goals Report 2009.

VAN DER MAAS, N. A. T., MOOI, F. R., DE GREEFF, S. C., BERBERS, G. A. M., SPAENDONCK, M. A. E. C.-V. & DE MELKER, H. E. 2013. Pertussis in the Netherlands, is the current vaccination strategy sufficient to reduce disease burden in young infants? *Vaccine,* 31**,** 4541-4547.

VESANTO, J. 1999. SOM-Based Data Visualization Methods. *Intelligent Data Analysis,* 3**,** 111-126.

VESANTO, J. & ALHONIEMI, E. 2000. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks,* 11**,** 568-600.

VIANA, M., MANCY, R., BIEK, R., CLEAVELAND, S., CROSS, P. C., LLOYD-SMITH, J. O. & HAYDON, D. T. 2014. Assembling evidence for identifying reservoirs of infection. *Trends in Ecology & Evolution,* 29**,** 270-279.

VIBOUD, C., BJORNSTAD, O. N., SMITH, D. L., SIMONSEN, L., MILLER, M. A. & GRENFELL, B. T. 2006. Synchrony, Waves, and Spatial Hierachies in the Spread of influenza. *Science,* 312**,** 447-451.

VLIET, J. V., WHITE, R. & DRAGICEVIC, S. 2009. Modeling urban growth using a variable grid cellular automaton. *Computers, Environment and Urban Systems,* 33**,** 35-43.

WAGNER, M., CAI, W. & LEES, M. H. Emergence by strategy: Flocking boids and their fitness in relation to model complexity. 2013 Winter Simulations Conference (WSC), 8-11 Dec. 2013 2013. 1479-1490.

WAGNER, M., CAI, W., LEES, M. H. & AYDT, H. 2015. Evolving agent-based models using self-adaptive complexification. *Journal of computational science,* 10**,** 351-359.

WANG, J.-F., GUO, Y.-S., CHRISTAKOS, G., YANG, W.-Z., LIAO, Y.-L., LI, Z.-J., LI, X.-Z., LIA, S.-J. & CHEN, H.-Y. 2011. Hand, foot and mouth disease: spatiotemporal transmission and climate. *International Journal of Health Geographics,* 10**,** 1-10.

WANG, N., BIGGS, T. W. & SKUPIN, A. 2013. Visualizing gridded time series data with self organizing maps: An application to multi-year snow dynamics in the Northern Hemisphere. *Computer, Environments and Urban Systems,* 39**,** 107-120.

WEHRENS, R. & BUYDENS, L. M. C. 2007. Self- and super-organizing maps in R: The kohonen package. *Journal of Statistical Software,* 21**,** 1-19.

WENDEL, J. & BUTTENFIELD, B. P. 2010. Formalizing Guidelines for Building Meaningful Self-Organizing Maps. *GIScience 2010.* Zurich.

WHITTINGTON, D., LAURIA, D. T., CHOE, K., HUGHES, J. A., SWARNA, V. & WRIGHT, A. M. 1993. Household sanitation in Kumasi, Ghana: A description of current practices, attitudes, and perceptions. *World Development,* 21**,** 733-748.

WHO. 2014. *Global Health Observatory (GHO)* [Online]. Available: http://www.who.int/gho/epidemic_diseases/cholera/cases_text/en/ [Accessed October 8, 2014 2014].

WIEGAND, T., JELTSCH, F., HANSKI, I. & GRIMM, V. 2002. Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *OIKOS,* 100**,** 209-222.

WILSON, S. W. 1986. Strobe Imagery: A scanning model . *Research Memo RIS.* The Rowland Institute for Science.

WORLDBANK. 2002. *Upgrading Low Income Urban Settlements, Country Assessment Report, Tanzania* [Online]. Available: http://web.mit.edu/urbanupgrading/upgrading/case-examples/overview-africa/country-assessments/reports/Tanzania-report.html [Accessed July 1, 2009.

WU, X., LU, Y., ZHOU, S., CHEN, L. & XU, B. 2016. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environment International,* 86**,** 14-23.

WU, X., ZURITA-MILLA, R. & KRAAK, M. J. 2013. Visual discovery of synchronisation in weather data at multiple temporal resolutions. *The cartographic journal,,* 50**,** 247-256.

XIE, Y. & BATTY, M. 2003. *Integrated urban evolutionary modelling,* London.

XIE, Y., BATTY, M. & ZHAO, K. 2005. Simulating emergent urban form: Desakota in China. *CASA working paper series.* London.

YAHJA, A. 2006. *Simulation Validation for Societal Systems.* PhD, Carnegie Mellon University.

YOUNG, G. 2010. *Socioeconomic Analysis of Informal Settlement Growth in Dar es Salaam: The concept for an agent based model.* MSc thesis, University of Twente.

ZURITA-MILLA, R., VAN GIJSEL, J.A.E., HAMM, N.A.S., AUGUSTIJN, P.W.M. , VRIELING, A. 2013. Exploring spatiotemporal phenological patterns and trajectories using self - organizing maps. *IEEE Transactions on geoscience and remote sensing,* 51**,** 1914-1921.

Summary

Patterns can be seen as footprints left behind by complex systems and, as such, their analysis leads to a better understanding of the underlying systems. This PhD thesis focuses on the following research objectives: (1) the development and evaluation of spatio-temporal pattern detection methods, (2) the use of these patterns to build and evaluate geographically explicit agent-based models (ABMs), and (3) the development of methods for the comparison of simulated and empirical patterns. These research objectives are illustrated by means of various case studies in the domains of urbanism and public health.

The methods used for the detection of spatio-temporal patterns in empirical data include Self-Organizing Maps (SOM) in combination with Sammon's projection, Dynamic Time Warping (DTW) and Shape-based distance (SBD). Three ABMs are discussed, including a model to simulate informal settlement development in Dar es Salaam (Tanzania), a cholera model for Kumasi (Ghana), and a pertussis model for the Netherlands. For the latter model, SOM and Sammon's projection are used to compare the patterns found in empirical and simulated data.

Chapter 2 demonstrates how Self-organizing maps (SOMs) in combination with Sammon's projection can be applied to identify spatiotemporal disease diffusion patterns. The method can be used to identify synchrony between spatial locations, to group epidemic waves based on similarity of diffusion pattern and to construct sequences of maps of synoptic states. The Sammon's projection was used to created diffusion trajectories from the SOM output. These methods were demonstrated with a dataset that reports Measles outbreaks that took place in Iceland in the period 1946-1970. Both stable and incidental synchronisation between medical districts were identified as well as two distinct groups of epidemic waves, a uniformly structured fast developing group and a multiform slow developing group. Diffusion trajectories for the fast developing group indicate a typical diffusion pattern from Reykjavik to the northern and eastern parts of the island. For the other group, diffusion trajectories are heterogeneous, deviating from the Reykjavik pattern.

In Chapter 3 time series clustering was applied to identify the critical community region (CCR), the threshold population above which an infectious disease maintains itself during inter-epidemic periods. Knowledge about the size and location of this area can be very useful for geographically targeted interventions, but delineation can be difficult in highly urbanized areas. In this study we use time series clustering to identify a collection of areas that together interact as a coherent unit that maintains the pathogen during non-epidemic periods and plays an important role in the disease diffusion process. As disease diffusion is triggered by human movement, we applied clustering

on percolation zones that are extracted from the road network. We apply this method on a case study of pertussis in the Netherlands for the period 1996-2013. Results reveal a pathogen reservoir that consists of a considerable number of un-connected zones distributed over different parts of the country. This region is robust over the majority of the tested epidemics.

Chapter 4 discusses the simulation of the growth of informal settlements. This type of simulation can be an essential building block to manage urbanization processes in cities of the developing world. We used agent-based modelling to develop a vector-based, micro-scale housing model to simulate the growth of informal settlements. A prototype of the housing model was implemented for Dar es Salaam, Tanzania. The results show that such a vector-based housing model built on three simple rules of spatial change (infilling, extension and enlargement of existing houses) can successfully simulate the housing pattern of informal settlements growth.

Chapter 5 introduces a spatially explicit agent-based simulation model for micro-scale cholera diffusion. The model simulates both an environmental reservoir of naturally occurring V. cholerae bacteria and hyperinfectious V. cholerae. Objective of the research is to test if runoff from open refuse dumpsites plays a role in cholera diffusion. A number of experiments were conducted with the model for a case study in Kumasi, Ghana, based on an epidemic in 2005. Experiments confirm the importance of the hyperinfectious transmission route, however, they also reveal the importance of a representative spatial distribution of the income classes. Although the contribution of runoff from dumpsites can never be conclusively proven, the experiments show that modelling the epidemic via this mechanism is possible and improves the model results.

Chapter 6 applies the SOM based diffusion trajectory method developed in chapter 2 to compare empirical diffusion patterns for pertussis in the Netherlands with simulated pertussis patterns. Diffusion patterns are compared in two ways, visual comparison and mapping empirical data back to a SOM trained based on simulated data. Trajectories for the empirical data showed that the patterns of pertussis are not robust, they differ considerably between the different epidemics, however all trajectories pass through a similar step at the beginning of the trajectory. This neuron corresponds to the state transition from non-epidemic to epidemic. Mapping both simulated and empirical data back to the same SOM allowed for a comparison of the diffusion patterns, yet also revealed that important information about the empirical patterns was lost. A specific objective of this research was to identify if introducing age-specific mobility (of adolescents) in simulation models influences the spatio-temporal diffusion patterns. Results show differences in spatio-temporal diffusion patterns for different commuting types.

In Chapter 7 three reflections are conducted related to the sub-objectives of this study. The first refection compares the methods developed in chapter 2 and 3. This is done by extracting the critical community region (CCR) for pertussis in the Netherlands using the SOM based method. Results of the different methods show good similarity.

The second reflection evaluates if characteristics of complex systems were used when building the three ABMs described in this thesis and evaluates the usefulness of these characteristics. Multiple drivers and state transitions proofed to be very helpful. The small map extent of the ABMs limited the detection of scaleless behaviour.

The third reflection compared the information derived based on temporal, spatial and spatio-temporal patterns to evaluated the added value of spatio-temporal methods. Especially for the cholera model, spatio-temporal patterns revealed a missing element in this model and therefore proved to be important.

Samenvatting

Patronen zijn de sporen van complexe systemen, en de analyse van deze sporen kan leiden tot een beter begrip van deze systemen. Dit proefschrift heeft de volgende onderzoeksdoelstellingen: (1) het ontwikkelen en evalueren van methoden voor het opsporen van ruimtelijk-temporele patronen, (2) het gebruik van deze patronen voor het ontwikkelen van geografisch expliciete agent-based modellen (ABMs), en (3) het ontwikkelen van methoden voor het vergelijken van empirische en gesimuleerde patronen. Deze onderzoeksdoelstellingen worden geïllustreerd door middel van een aantal case studies op het gebied van stedelijke ontwikkeling en verspreiding van ziektes.

De methoden die worden gebruikt voor het detecteren van ruimtelijk-temporele patronen in empirische data omvatten Self-Organizing Maps (SOMs) in combinatie met de Sammon projectie, Dynamic Time Warping (DTW) en Shape-based distance (SBD). Drie agent-based modellen worden bediscussieerd, een model voor het simuleren van het ontstaan van sloppenwijken in Dar es Salaam (Tanzania), een cholera model voor Kumasi (Ghana), en een model voor het simuleren van kinkhoest in Nederland. Voor het laatste model worden SOM en Sammon projectie gebruikt voor het vergelijken van de gesimuleerde patronen met empirische data.

Hoofdstuk 2 laat zien hoe SOMs in combinatie met de Sammon projectie kan worden toegepast om ruimtelijk-temporele patronen op te sporen. Deze methode kan synchronisatie van infectie tussen twee ruimtelijke locaties opsporen, en kan worden gebruikt om data van verschillende epidemieën te groeperen op basis van verspreidingspatroon en om series van kaarten van synoptische fasen te maken. De Sammon projectie wordt gebruikt om verspreidingsvectoren te generen op basis van de SOM uitkomsten. Deze methode wordt geïllustreerd met een dataset van mazelen epidemieën in IJsland gedurende de periode 1946-1970. Zowel stabiele als incidentele synchronisaties tussen medische districten werden gevonden, en epidemieën werden geclassificeerd in twee groepen, een uniforme snel verspreidende groep en een pluriforme langzaam verspreidende groep. De diffusie patronen van de snel verspreidende groep laten een verspreidingspatroon zien van Reykjavik naar het noorden en oosten van het eiland. De diffusiepatronen van de andere groep zijn heterogeen, en wijken af van het Reykjavik patroon.

In hoofdstuk 3 wordt tijdserieclustering door middel van DTW en SBD toegepast om de “critical community regio (CCR)” vast te stellen, het gebied waarin een infectie ziekte zichzelf kan handhaven in perioden tussen epidemieën. Kennis over de grootte en locatie van dit gebied kan nuttig zijn bij geografisch gerichte interventies, maar het kan moeilijk zijn om de CCR te bepalen in geürbaniseerde gebieden. In deze studie wordt clustering van

tijdseries gebruikt om een groep gebieden te identificeren waarin het pathogeen zich handhaaft gedurende niet-epidemische perioden. Dit gebied kan een belangrijke rol spelen in het ziekteverspreidingsproces. Omdat de verspreiding van infectie ziekten wordt versterkt door mobiliteit, gebruiken we percolatie zones die gebaseerd zijn op het wegennetwerk. We passen deze methode toe op een studie van kinkhoest in Nederland voor de periode 1996-2013. Resultaten laten zien dat het pathogeen reservoir bestaat uit een groot aantal, in het land verspreid liggende gebieden. Dit gebied is robuust over het grootste deel van de geteste epidemieën.

Hoofdstuk 4 richt zich op de groei van sloppenwijken. Dit type simulatie kan een belangrijke bijdrage leveren aan het beheersen van het urbanisatieproces in steden in ontwikkelingslanden. Een "Agent-based" modelleer techniek werd gebruikt om een simulatie te ontwikkelen die gebaseerd is op vector data, en gebruik maakt van een microschaal bebouwingsmodel om de groei van de nederzetting te simuleren. Een prototype van dit model werd geïmplementeerd voor Dar es Salaam in Tanzania. De resultaten laten zien dat zo'n, op vector data gebaseerd model op basis van drie simpele constructie regels (inbreiding, expansie en het uitbouwen van bestaande woningen), een succesvolle simulatie op kan leveren van de patronen die worden gevonden tijdens de groei van sloppenwijken.

In hoofdstuk 5 wordt een ruimtelijk expliciet agent-based model voor cholera diffusie beschreven. Dit model simuleert zowel het voorkomen van de bacterie V. cholerae in het milieu alsook de hoog besmettelijke variant van V. cholerae. De doelstelling van dit onderzoek is om te testen of afstroom van regenwater via vuilnisbelten een rol kan spelen bij de verspreiding van cholera. Een aantal experimenten werd uitgevoerd voor het case studie gebied in Kumasi, Ghana, gebaseerd op een epidemie in 2005. Deze experimenten laten zien dat de hoog besmettelijke variant een belangrijke rol heeft gespeeld tijdens deze epidemie, ze laten echter ook zien dat verspreiding sterk beïnvloed wordt door inkomensklassen. Hoewel verspreiding van cholera via afwatering over vuilnisbelten door dit model niet kan worden bewezen, laten de experimenten zien dat dit mechanisme een goede overeenkomst vormt met de empirische verspreidingspatronen.

In hoofdstuk 6 wordt de SOM methode die ontwikkeld werd in hoofdstuk 2 toegepast om empirische diffusie van kinkhoest in Nederland te vergelijken met gesimuleerde diffusie patronen. Dit gebeurt op twee manieren, door middel van visuele vergelijking en door middel van het projecteren van empirische data op een SOM die gebaseerd is op gesimuleerde data. De verspreidingsvectoren van de epidemieën laten zien dat er duidelijke verschillen zijn in ruimtelijk-temporele verspreiding van de epidemieën, maar alle verspreidingsvectoren gaan via eenzelfde traject (neuronen) aan het begin

van de epidemie. Deze begin vectoren komen overeen met de transitie van niet epidemisch naar epidemisch. Door zowel empirische als gesimuleerde data te projecteren op dezelfde SOM is het mogelijk om diffusie patronen direct te vergelijken. Hierbij gaat echter wel belangrijke informatie over het empirische verspreidingspatroon verloren. Een specifieke onderzoeksvraag bij dit onderzoek was of de introductie van leeftijdsspecifieke mobiliteit, met name van tieners, in het simulatie model invloed heeft op de ruimtelijk-temporele patronen. De resultaten laten inderdaad verschillen zien in verspreidingspatronen bij het simuleren van verschillende mobiliteitstypen.

In hoofdstuk 7 worden drie reflecties uitgevoerd die gerelateerd zijn aan de deeldoelstellingen van deze studie. De eerste reflectie vergelijkt de methoden die werden ontwikkeld in hoofdstuk 2 en 3. Dit werd gedaan door de critical community regio (CCR) voor kinkhoest in Nederland te bepalen met behulp van de SOM methode (hoofdstuk 2). Resultaten zijn vergelijkbaar met de originele uitkomsten uit hoofdstuk 3.

De tweede reflectie evalueert of de karakteristieken van complexe systemen werden gebruikt voor het bouwen van de drie beschreven simulatie modellen en reflecteert op de toepasbaarheid van deze karakteristieken. Het inbouwen van diverse oorzaken van verandering, en de aanwezigheid van een overgangsstaat (state transition) bleken heel nuttig te zijn. Het kleine simulatie gebied van agent-based modellen bemoeilijkt het detecteren van schaalloze patronen.

De derde reflectie vergelijkt de informatie die wordt verkregen op basis van de temporele, ruimtelijke en ruimtelijk-temporele patronen om zodoende de toegevoegde waarde van ruimtelijk-temporele patronen te bepalen. Deze patronen hebben met name toegevoegde waarde voor het cholera model, omdat op deze manier een belangrijk missend element, namelijk risico vermijdend gedrag van mensen, in het model kan worden opgespoord.