Utrecht University

Graduate School of Natural Sciences

Inferring dynamics from data in rotavirus epidemiology

Master's Thesis

Author:
Alejandro Javier
Alarcón González

Supervisors:
Prof. dr. ir. Jason Frank
Prof. dr. Jacco
Wallinga
Dr. Don Klinkenberg

Second reader:
Dr. Fieke Dekkers

July 25, 2019

A Natalia

Abstract

The focus of this thesis is the inference of changes in the dynamics of rotavirus epidemiological data. As was discussed by S Hahné et al. [1], there was an exceptionally low rotavirus incidence in the Netherlands in the winter of 2013/2014. Motivated by an internal report from the National Institute of Public Health and the Environment (RIVM) [2] that provided a transmission model of rotavirus dynamics that suggested the appearance of bifurcations, we try to detect such bifurcations by analysing rotavirus time series with the use of Wasserstein distances (as is discussed by Michael Muskulus and Verduyn-Lunel in [3] for time series in general). Although we did not manage to detect the possible period doubling bifurcation affecting the Netherlands, we could use the Wasserstein distances approach to detect changes in the dynamics of rotavirus corresponding to the introduction of vaccination against the disease in Germany.

Acknowledgements

I want to thank my supervisors Prof. dr. ir. Jason Frank (UU) and Prof. dr. Jacco Wallinga (RIVM) for their guidance during the process of this thesis research project. Their commitment and ideas were a source of inspiration for my research. Dr Don Klinkenberg (RIVM) has also been of much support throughout the several months of the project.

I want to thank the CONACYT-Government of Tabasco scholarship for supporting me with the (partial) costs to pursue master studies at the University of Utrecht. Without this funding I would had never being able to attend postgraduate studies abroad.

I want to thank my parents for their love and example.

I am very grateful for the kindness and support of my friends in Utrecht. The multiple discussions with them provided me with better and broader understanding of my project. In particular, I want to thank Yuki, Guille, Jorge, Ernst and Marghe. Equally important, their company has given me many beautiful memories.

Being a student of the University of Utrecht has been a challenging experience, but highly recommendable.

Introduction

The purpose of this thesis research project is to test the method developed by Verduyn-Lunel and Muskulus in the paper "Wasserstein distances in the analysis of time series and dynamical systems" [3] as a numerical tool to detect qualitative changes in the underlying dynamics of time series of infections. Their method is promising as it has given interesting results when used to analyse neurological processes and time series of respiratory impedance ([4] and [5], respectively). For the time series of our interest, namely rotavirus incidence per week, we observe a cyclic pattern repeating each year (see figure 1), with high incidence reported around winter time and low incidence during the summer months. As the reason for this pattern is well known, we will pay special attention to the apparent transition from annual to biennial cyclic incidence occurring in the Netherlands in the year of 2014. As this might reflect a qualitative change in the dynamics of the epidemiological system of rotavirus, we will analyse this possible transition from the perspective of dynamical systems theory[1], and within this context, we pay special attention to the concept of *bifurcation* in the underlying dynamical system (as a previous research [2] leads to the hypothesis that the above mentioned transition reflects a bifurcation).

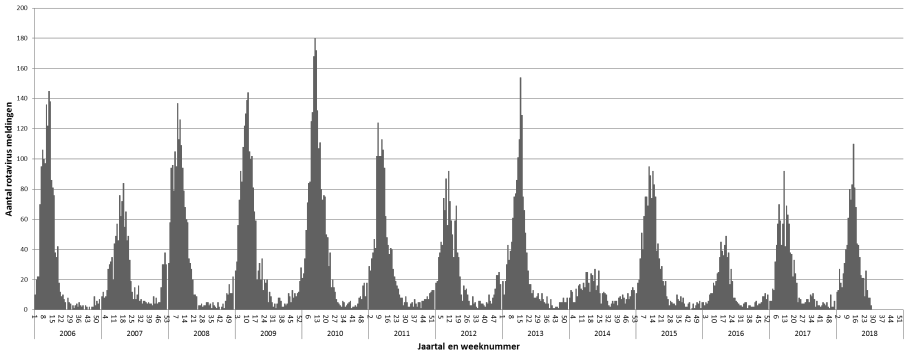


Figure 1: Weekly incidence of rotavirus in the Netherlands.

We propose to test the bifurcation hypothesis on the grounds that we do not know the set of differential equations describing the spread of rotavirus within a population. As this means that we first need to obtain a multidimensional dynamical representation that best corresponds to the behaviour of data (in one

[1]Which roughly speaking, is the branch of mathematics interested in the qualitative behaviour of a dynamical system.

variable), we follow the approach of Verduyn-Lunel and Muskulus in [3], where they choose for an attractor reconstruction technique called *delay embedding*. The delay coordinate time series thus obtained represent dynamical regimes.

The reason for the selection of the delay embedding comes from the search of a robust representation of a scalar time series. The vector time series thus obtained is less sensitive to both noise in measurements and the selection of initial conditions than the original scalar time series.

We assume the existence of an attractor for the underlying dynamical system of rotavirus transmission, an attractor that is changing with the conditions of the system. To determine whether there exists a vector time series representation that will be faithful to the dynamical behaviour of this attractor, we can rely on an important result by Takens, namely that a delay coordinates time series will show both topological and differential equivalence to a trajectory in the attractor.

Once this reconstruction is obtained, we will regard the space of reconstructed trajectories as a measure space, over which we identify each trajectory with a probability measure. This comes from the idea that two realisations of the same stochastic process will show similar distributions (or even the same if we care about the long-term behaviour). It is this kind of similarity that we want to (numerically) measure for trajectories of the rotavirus dynamical system.

Verduyn-Lunel and Muskulus choose for the Wasserstein distance among a variety of distances between probability measures because it takes into account both geometric and probabilistic aspects in the reconstructed trajectory (as discussed in [6]).

Our claim can be stated as follows: if the delay coordinate map is a good method, then the use of Wasserstein distances has the potential to detect bifurcations in the dynamics of rotavirus transmission.

In chapter 1, the mathematics supporting all methods are presented. In chapter 2, the model of Alexandra Teslya is presented [2], and the method of Verduyn-Lunel and Muskulus [3] is tested on synthetic time series obtained from it. In chapter 3, the same method is applied to German notification data. Finally, in chapter 4, all the results are discussed and conclusions from the research project are given.

Contents

Chapter 1

Mathematical background

In this chapter we start with a short review of dynamical systems theory (although we refer to [7] for a complete treatment of the subject) with the purpose of making the text as self-contained as possible. In the second part of the chapter we address a technique to reconstruct attractors from time series. In the last part, we introduce an optimisation problem that aims at providing a numerical method for detecting bifurcations.

Hereafter, we deal with concepts arising from a deterministic continuous in time dynamical system defined by a smooth vector field $f : M \to TM$, where M is a smooth manifold and TM is the tangent space of M. The integral curves of this vector field provide the *state space* M with a flow Ψ on it, such that $\Psi(M) = M$.

A continuous in time dynamical system can be defined by a system of differential equations of the form

$$\dot{x}(t) = f(x, \lambda), \tag{1.1}$$

where $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m$ is a free parameter, i.e. there are n coordinates and m parameters in the system.

An example of a differential equation of the form (1.1) is the SIR-type model (2.1), designed to describe the dynamics of rotavirus within a population. As we are interested in understanding the dynamical behaviour of this model's solutions, we include in the following a theoretical overview of the possible qualitative changes in the flow dynamics of systems of the form (1.1), when the parameter λ is varied. Such qualitative changes are called *bifurcations*, which in order to be discussed, the concepts of *trajectory* and *attractor* first need to be introduced.

Definition 1.0.1 (Trajectory) *The trajectory of the initial point $x(t_0) \in \mathbb{R}^n$ under the flow Ψ of the dynamical system* (1.1) *is given by the set $\{x(t)\}_{t>t_0}$.*

Definition 1.0.2 (Phase portrait) *The set of all trajectories of the dynamical system* (1.1) *forms the phase portrait.*

In the examples treated in this thesis project, we can understand that a time series made up of observed incidence of rotavirus $\{x_i\}_{i\in\mathbb{I}\subset\mathbb{N}}$ can be regarded as

a (discrete) trajectory for an unknown dynamical system, which is the one we try to obtain information from. We will deal only with limiting trajectories or attractors, since these are the components of the flow Ψ that characterize most faithfully a dynamical system.

Definition 1.0.3 (Attractor) *A set $A \subset \mathbb{R}^n$ is an attractor of the open set $U \subset \mathbb{R}^n$ under the flow Ψ if for every neighbourhood V of A, and an initial condition $x(t_0)$ of (1.1) lying in U, there exists a real number $k(V)$ such that the corresponding trajectory $\{x(t)\}_{t \geq k(V)} \subset V$.*

In the case of our interest, we assume that the given time series are obtained from observations of trajectories close to the attractor of their corresponding dynamical system. Such attractors might be of chaotic nature, given the known non-linearity of transmission models from epidemiology, which are the ones used for modelling the transmission of infectious diseases [8]. However, we assume that the attractor is topologically a circle (or more specifically a *limit cycle*), since this is what the cyclic pattern in the time series 1 and the solutions of the rotavirus transmission model in [2] suggest.

In this thesis project only two kinds of bifurcations are considered: the period doubling bifurcation and the limit point of cycles bifurcation. We restrict our treatment for the following reasons:

- The actual measurements of rotavirus incidence suggest that a period-doubling bifurcation in the dynamics might have occurred. This comes from the observation of time series 1.
- The cyclic solutions of the transmission model in [2] undergo both period-doubling and limit point of cycles bifurcations, as can be seen in the bifurcation diagram 2.3.

1.1 Bifurcations in Dynamical systems

Definition 1.1.1 (Bifurcation) *A bifurcation is the appearance of a topologically nonequivalent phase portrait of a dynamical system when the system parameters vary.*

For example, we observe this phenomenon in the case of a period-doubling bifurcation (to be explained in subsection 1.1.1), as there is no way to continuously deform two cycles into one.

The mathematical analysis of bifurcations (simply called *Bifurcation theory*) deals with the study of *normal forms* (thoroughly discussed in [9]), which are power series representations of dynamical systems that allow for a systematic characterisation of dynamical regimes in terms of the coefficients in the series (also called *normal form coefficients*). We remark that the defining criterion for identifying dynamical regimes is topological equivalence.

1.1.1 Codimension one bifurcations of limit cycles

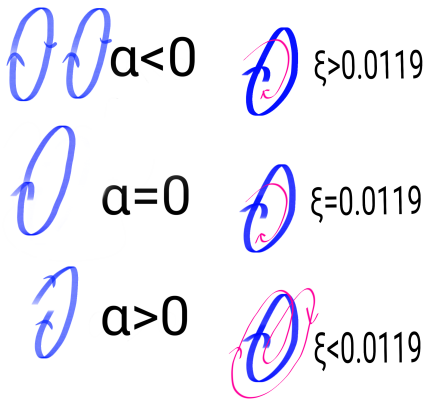
Bifurcations of the dynamics can be classified by their *codimension number*, which is the number of system parameters that need to be varied in order for

a bifurcation to occur. The two bifurcations types that are considered within this text correspond to codimension one bifurcations of limit cycles. These are described in the following.

If the solutions of a dynamical system (1.1) converge to one (or more) limit cycle(s) (as is the case in seasonally forced transmission SIR-models of epidemiology [8]), the **codimension one** bifurcations that such cycles may undergo are:

1. *Limit point of cycles* (also called fold bifurcation): this consists of the collision and disappearance of two cycles (one stable and one saddle) when one of the parameters in (1.1) crosses a critical value α^*. This concept is illustrated in figure 1.1a. In relation to this kind of bifurcation, the parameter region corresponding to the existence of two stable limit cycles and one saddle cycle will be referred to as a *bi-stability region*.

2. *Period-doubling bifurcations* (also called flip bifurcation): this consists of the change in the stability of a cycle from stable to unstable and the emergence of a stable limit cycle with double the period of the unstable cycle when one of the parameters in (1.1) crosses a critical value ξ^*. This concept is illustrated in figure 1.1b.



(a) (b)

Figure 1.1: In figure 1.1a there is illustrated the transition occurring in a limit point of cycles bifurcation when one of the parameter of the system (1.1) crosses the critical value $\alpha = 0$. The transition illustrated in figure 1.1b corresponds to a period-doubling bifurcation to be seen in chapter 2.

As we will perform numerical bifurcation analysis in the next chapter, we now provide an analytical method for detecting codimension one bifurcations of cycles based on the multipliers of the *monodromy matrix*.

Definition 1.1.2 *If $u(t) \in \mathbb{R}^n$ is a limit cycle solution of (1.1) with period τ, i.e. $u(t) = u(t+\tau)$ (for all $t \in \mathbb{R}^+$), then the cycle multipliers (also called Floquet multipliers) of $u(t)$ are the eigenvalues $\{\mu_i\}_{i=1}^n$ of the monodromy matrix $M(\tau)$, which is in turn defined as the solution of the system*

$$\begin{cases} \dot{M}(t) &= f_x(u(t), \lambda)M(t) \\ M(0) &= I_n, \end{cases} \tag{1.2}$$

where I_n is the $n \times n$ identity matrix.

The first feature to recognize in the multipliers $\{\mu_i\}_{i=1}^n$ is that one of them is equal to one, e.g. $\mu_n = 1$ without loss of generality. The significance of this is found in the *perturbations* along $u(\tau)$: if y_0 is a vector tangent to $u(\tau)$, then the linear approximation of the vector field around the cycle ($f_x(u(t), \lambda)$) will keep y_0 fixed, i.e. $f_x(u(\tau), \lambda)y_0 = y_0$. It turns out that such y_0[1] satisfies $M(\tau)y_0 = \mu_n y_0 = y_0$. As the derivation of this result relies on concepts from Floquet theory which we decide not to include in this thesis report, we refer to [10, p. 60] and [7, p. 25] for a discussion on the relation between perturbations of $u(t)$ and the set of multipliers of $M(\tau)$.

In the following, we give a criterion for assessing the stability of $u(t)$ in terms of the eigenvalues of $M(\tau)$.

- If $|\mu_i| < 1$ for all $i \neq n$, then the cycle $u(t)$ is stable.
- On the other hand, if $|\mu_i| > 1$ for any $i \in 1, \ldots, n-1$, then the cycle $u(t)$ is unstable.

The multipliers $\{\mu_i\}_{i=1}^n$ of $M(\tau)$ provide the next criterion for detecting codimension one bifurcations of the corresponding limit cycle.

- If $\mu_1 = 1$ at the parameter value λ^*, then the cycle $u(t)$ is undergoing a limit point of cycles bifurcation.
- If $\mu_1 = -1$ at the parameter value λ^*, then the cycle $u(t)$ is undergoing a period-doubling bifurcation.

As the numerical bifurcation package MATCONT [11] (to be extensively used in the next chapter) has the computation of cycle multipliers implemented in it, but only of their absolute values, MATCONT [11] also requires the computation of the normal form coefficients in order to numerically detect bifurcations of limit cycles.

1.2 Attractor reconstruction

The motivation for attractor reconstruction methods is that a time series can be regarded as an observable function of the solution trajectory near the attractor of a dynamical system (provided we count on one). In case the system is nonlinear (as is the case in dynamical systems describing interactions such as the transmission of a disease within a population), then we may be facing what is called *chaotic determinism*, in which two seemingly different trajectories might belong to the same attractor. Thus we choose for a vector time series representation that does not fall short in showing the bigger dynamical picture (as the

[1] y_0 can be expressed as $y_0 = e^{\lambda\tau}y^*$, where y^* is an eigenvector corresponding to the eigenvalue $\lambda = 0$ of the matrix R such that $M(\tau) = e^{R\tau}$.

scalar time series may do). Moreover, as real measurements are prone to noise, and the analysis of a noisy scalar time series is more likely to be biased, it then becomes natural to search for a vector time series representation that can work around some noise. Attention should be stressed to the fact that we do not do this in order to neglect (possible) randomness inherent to the rotavirus epidemiological system, as it is well understood that randomness and determinism pull the strings governing physical phenomena [12].

The discussion in the previous paragraph finds its relevance in chapters 2 and 3, where we deal with real-life observations of a dynamical system, namely weekly measurements of rotavirus incidence[2] in Germany. As will be illustrated in chapter 2, the observable is not even a variable of the system, and furthermore the epidemiological measurements only provide one-variable data. What we will proceed to do in the first part of this thesis project is to **construct** a higher dimensional representation of the system trajectories from the one-variable weekly incidence of rotavirus. Such multidimensional representation will provide us with richer information of the dynamical behaviour. In this section, we give the theoretical foundations of this reconstruction.

The attractor reconstruction technique we consider is called the *Delay coordinate map*. This was first discussed by Floris Takens [13], and it briefly works as follows: for a given time series $\{x_n\}_{n=1}^N$, a time delay τ and an *embedding* dimension k, k-delay coordinates are formed by sampling $\{x_n(t)\}_{n=1}^N$ every time τ starting from any $n \in \{1, \ldots, N-(k-1)\tau\}$. It will be shown that this procedure provides a representation that reconstructs the differential structure of the system attractor A, i.e. the delay map is an embedding of A.

1.2.1 Delay embedding

We recall the notion of an observable and formalise it as follows: an *observable* is a smooth function $h : \mathbb{R}^k \to \mathbb{R}$. For instance, an observable of the dynamics of rotavirus in the population is its incidence.

If we recall the expression (1.1) of a dynamical system, and substitute the domain of f by the more general n-dimensional smooth manifold M, we can understand that the solutions $x(t)$ will define a flow Ψ on M.

The time series considered in the next chapters are made by first sampling every time τ a solution trajectory $x(t) \subset M$ of a (whether known or unknown) dynamical system and then applying an observable function h on this sampling, i.e. from the sampled solution $\{x_n\}_{n=1}^N = \{x(\tau n)\}_{n=1}^N$, we consider the set $\{h(x_n)\}_{n=1}^N$. We will not analyse such time series directly, but the corresponding set of *delay coordinates*, a technique which is introduced in the following.

Definition 1.2.1 (Delay coordinate map) *Given a manifold M equipped with a flow Ψ on it, a time delay $\tau \in \mathbb{R}^+$, and h an observable function on M, we define the delay coordinate map $F(h, \Psi, T) : M \to \mathbb{R}^k$ as:*

$$F(h, \Psi, \tau)(x) = (h(x), h(\Psi_\tau(x)), \ldots, h(\Psi_{(k-1)\tau}(x))). \tag{1.3}$$

We now explain the components of the delay coordinate map $F(h, \Psi, \tau)$ for the cases considered in this thesis project.

[2] By *incidence* we mean the number of notifiable cases per 100,000 persons.

- In the case of synthetic time series $\{h_1(x_n)\}_{n=1}^{N}$ (to be analysed in chapter 2), the flow Ψ is defined by the vector field of the system (2.1), h_1 is the solution of the influx into compartment I_1, and τ is equal to one week.

- In the case of rotavirus incidence in German regions (the time series $\{h_2(y_i)\}_{i=1}^{M}$ to be analysed in chapter 3), Ψ is again the vector field of the (unknown) underlying dynamical system, the image of h_2 is the rotavirus incidence, and τ is equal to one week.

The image of $F(h, \Psi, \tau)$ on a time series $\{x_n\}_{n=1}^{N}$ is also a time series consisting of k coordinates for each point in it. If we set $f_n := F(h, \Psi, \tau)(x_n)$ for $1 \leq n \leq N - (k-1)\tau$, then the resulting delay coordinates time series is

$$\{f_n\}_{n=1}^{\bar{N}} \subset \mathbb{R}^k, \tag{1.4}$$

where $\bar{N} = N - (k-1)\tau$ is the maximum length of the new time series. The space $\mathbb{R}^k$ will be called *reconstruction space* for the reason that this is the space where the reconstructed trajectory is obtained.

Remark. It is not the purpose of this thesis project to properly reconstruct the attractor for a given time series, but rather to compare time series via their (possibly degenerate) reconstructions. Therefore, we will not care for finding the optimal embedding dimension k.

A theoretical result of high value for the purpose of this thesis project would be the map $F(h, \Psi, \tau)$ to define an embedding of the attractor into reconstructed space. The first theorem in such direction was provided by Whitney [14], who proved the embedding to exist for smooth maps $\Phi : \mathbb{R}^k \to \mathbb{R}^{2d+1}$ if the attractor lies within a compact smooth manifold of dimension d. Takens [13], in an effort to help experimentalists, extended this theorem by addressing the delay coordinate map again on a compact smooth manifold in $\mathbb{R}^k$. However, it might be the case that the attractor has a fractal dimension, thereby becoming what is called *a strange attractor*. It is in the work of Sauer, Yorke and Casdagli [15] that fractal attractors were considered, and it is their result we will rely on.

One uncommon concept is present in the series of theorems mentioned before. The last theorem concludes that $F(h, \Psi, \tau)$ is an embedding with *probability one*. We place attention on this since the functional space of smooth maps is infinite dimensional. However, the Lebesgue concept of probability one (or *almost every*) is defined on linear spaces with finite dimension, and therefore the need for an extension should be considered. This extension comes in the concept of prevalence:

Definition 1.2.2 (Prevalence) *A Borel subset S of a normed vector space V is prevalent if there is a finite-dimensional vector subspace $E \subset V$ such that for $v \in V$, $v + e \in S$ for almost every e in E.*

Hereafter, we will interchangeably use the terms prevalence and almost every to refer to probability one over normed linear spaces.

On the other hand, when addressing the concept of dimensionality of an attractor, and if we suppose that this is a fractal, we resort to the *box counting dimension*, defined in the following.

Definition 1.2.3 (Box-counting dimension) *If A is a strange attractor covered by voxels of size ε, the dimension of A is given by the limit*

$$boxdim(A) = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{-\log \varepsilon}, \tag{1.5}$$

where $N(\varepsilon)$ is the number of voxels that make up the cover.

The embedding theorem of Sauer, Yorke and Casdagli [15] can now be stated.

Theorem 1.2.1 (Fractal Delay Embedding Prevalence Theorem) *Let Ψ be the flow of a dynamical system on an open subset U of $\mathbb{R}^k$, and let A be a compact subset (possibly a fractal) of U of box counting dimension d. Let $k > 2d$ be an integer, and let $\tau > 0$. Assume that A contains at most a finite number of equilibria, no periodic orbits of Ψ of period τ or 2τ, at most finitely many periodic orbits of period $3\tau, 4\tau, \ldots, n\tau$, and that the linearisations of those periodic orbits have distinct eigenvalues. Then for almost every smooth function h on U, the map $F(h, \psi, \tau)$ is:*

1. *One-to-one on A.*
2. *A C^1-immersion on each compact subset C of a smooth manifold contained in A.*

Remark. Theorem 1.2.1 (as it appears in [15]) is given in terms of a backwards (in time) delay coordinate map $F(h, \psi, \tau)$, but their proof also holds for our choice of a forward (in time) delay coordinate map.

We elaborate on some mathematical implications of Theorem 1.2.1. For a given observable h, the map $F(h, \Psi, \tau)$ satisfies (with probability one) the following:

1. F is one-to-one on A: this ensures that in case a solution trajectory (close to the attractor) is being intertwined with h, then the delay map F will disentangle it back to its original topology in reconstructed space.
2. F is a C^1-immersion on each compact subset C of a smooth manifold contained in A. This ensures that the Monodromy matrix $M(\tau)$ (the solution of (1.2)) exists for the reconstructed cycle and furthermore it has full rank, which allows for the codimension one bifurcation analysis of cycles based on Floquet multipliers (explained in subsection 1.1.1).

The practical relevance of theorem 1.2.1 is the following. By applying the delay coordinate map $F(h, \psi, \tau)$ on scalar time series, with $k > 2d$ we obtain a multidimensional time series that with probability one will satisfy the same dynamical properties of the actual trajectory (of the unknown dynamical system), if we assume that the observations are taken close to the attractor.

1.3 Wasserstein distances

In this section we introduce a measure-theoretic representation of the delay coordinates, and a distance (called *Wasserstein distance*) to compare the measures thus created. Such distance will in turn provide us with a numerical value

that aims to reflect the similarity between dynamical regimes. This naturally suggests the use of the distance as a numerical tool for detecting bifurcations in the dynamics.

Hereafter, we assume that the delay coordinate map has been applied to a given time series $\{x_i\}_{i=1}^{N}$, and furthermore the reconstructed space $\Omega = \mathbb{R}^k$ will be regarded as a measurable space (equipped with the Borel $\sigma-$algebra $\mathbb{B}$). The probability measure $\mu_{\{f_i\}} : (\Omega, \mathbb{B}) \to [0, 1]$ is defined as:

$$\mu_{\{f_i\}}(A) := \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \delta_{f_i}(A), \tag{1.6}$$

where $A \in \mathbb{B}$ and $\delta_{f_i}(A)$ is the Dirac measure of the delay-coordinates point f_i, with $1 \leq i \leq \bar{N}$. Note that A in definition (1.6) can be an attractor in phase space $\mathbb{R}^k$, since all attractors are closed and $\mathbb{B}$ is the σ-algebra generated by closed sets in Ω.

The interpretation we give to the measure $\mu_{\{f_i\}}$ is the following. For a Borel set $A \in \mathbb{B}$, $\mu_{\{f_i\}}(A)$ indicates the (normalized) number of visits paid by the series $\{f_i\}$ to A. Equivalently, $\mu_{\{f_i\}}(A)$ is the *counting measure* of the set $A \cap \{f_i\}$. We remark that $\mu_{\{f_i\}}$ defines a probability distribution in $(\Omega, \mathbb{B})$.

In definition (1.6), $\bar{N}$ represents the length of the reconstructed time series, and in general we would require $\bar{N} \to \infty$ if we want μ to become an *invariant measure*, i.e. a probability distribution satisfying $\mu(A) = \mu(\Psi(A))$ for every $A \in \mathbb{B}$. For instance, in the reconstruction exercises of the following chapter we will obtain a time series from a limit cycle A (which assumes $\bar{N} \to \infty$). Since $\Psi(A) = A$ (by definition of limit cycle), it follows that $\mu(A) = \mu(\Psi(A))$, and therefore μ becomes an invariant measure under this particular scenario. For the incidence of rotavirus time series to be analysed later, the conditions are the same up to small random perturbations, since we assume that actual measurements of rotavirus are the result of a noisy observable function on trajectories close to a limit cycle. The issue arising here would be the non-uniqueness of μ, as the random fluctuations may affect the counting measure in $A \cap \{f_i\}_{i=1}^{\bar{N}}$. However, the perturbations being small implies the unicity of μ (for a proof this we refer to Lasota [16]). Therefore, the measure μ in (1.6) is called *natural invariant measure* when $\bar{N} \to \infty$.

We proceed to consider the functional space composed of invariant measures on $(\Omega, \mathcal{B}, \Psi)$ and provide it with the *Wasserstein metric*.

For two probability measures $\mu_{\{f_i\}}$ and $\mu_{\{g_i\}}$ (generated by (1.6) from the times series $\{x_i\}$ and $\{y_i\}$, respectively), we introduce the concept of *transporting* the probability distribution in $(\Omega, \mathbb{B}, \mu_{\{f_i\}})$ to the probability distribution in $(\Omega, \mathbb{B}, \mu_{\{g_i\}})$.

Definition 1.3.1 (Transportation plan) *A transportation plan consists of a configuration specifying how much probability density will be moved between the Borel sets of* $(\Omega, \mathbb{B}, \mu_{\{f_i\}})$ *and* $(\Omega, \mathbb{B}, \mu_{\{g_i\}})$ *in order to move one whole distribution to the other. This plan takes the form of a (product) measure* π *on* $(\Omega \times \Omega)$ *that satisfies*

$$\int_{y \in \Omega} d\pi[A, y] = \mu_{\{f_i\}}(A) \text{ and } \int_{x \in \Omega} d\pi[x, B] = \mu_{\{g_i\}}(B), \tag{1.7}$$

for all $A, B \in \mathbb{B}$.

Furthermore, the total cost of the plan π *is given by the functional form*

$$C(\pi) = \int_{\Omega\times\Omega} \|x - y\|_2 d\pi[x, y], \tag{1.8}$$

where $\|\cdot\|_2$ *denotes Euclidean distance.*

We denote the product measure space over which $C(\pi)$ is defined as $(\Omega \times \Omega, \mathbb{B} \times \mathbb{B}, \Pi(\mu_{\{f_i\}}, \mu_{\{g_i\}}))$, where $\Pi(\mu_{\{f_i\}}, \mu_{\{g_i\}})$ is the set of product probability measures (defined in general in [17]) generated by $\mu_{\{f_i\}}$ and $\mu_{\{g_i\}}$, and satisfying (1.7). The Euclidean distance in between x and y in Ω gives the *cost* of moving one probability mass point to the other, and this is why $C(\pi)$ is called the *cost functional*.

We introduce now the Wasserstein distance W between two measures $\mu_{\{f_i\}}$ and $\mu_{\{g_i\}}$ on $(\Omega, \mathbb{B})$ as follows:

$$\begin{aligned} W(\mu_{\{f_i\}}, \mu_{\{g_i\}}) &= \inf_{\pi\in\Pi(\mu_{\{f_i\}},\mu_{\{g_i\}})} \int_{\Omega\times\Omega} \|x - y\|_2 d\pi[x, y] \\ &= \inf_{\pi\in\Pi(\mu_{\{f_i\}},\mu_{\{g_i\}})} C(\pi). \end{aligned} \tag{1.9}$$

In the literature, this discussion is referred to as the *optimal transportation problem*.

Since the time series to be considered in this thesis project are all of finite length, the corresponding measure (1.6) will consist of finite sums of Dirac measures. This converts the problem (1.9) into a discrete transportation problem, as described in detail by Moeckel and Murray [6].

1.3.1 Discrete transportation problem

The probability measures (1.6) obtained from finite times series become probability mass functions. The shipping plan of the corresponding probability mass points within reconstructed space is explained in the following. We regard a box $B \in \mathbb{B}$ containing both delay coordinates, i.e. $\{f_i\}_{i=1}^{\bar{N}}, \{g_i\}_{i=1}^{\bar{N}} \subset B \subset \mathbb{R}^k$. If we create a regular division of B into the set $\{B_n\}_{n=1}^{m}$, then we will care for the invariant measures of the sub-boxes by introducing the set of probabilities $p_n = \mu_{\{f_i\}}(B_n)$ and $q_n = \mu_{\{g_i\}}(B_n)$, with $n = 1, \ldots, m$.

A *transportation plan* in this context would provide us with a configuration that specifies the amount of probability mass that is moved in between all pairs of boxes in order to move one whole mass distribution to the other. If f_{ij} is the amount of mass points being moved from B_i to B_j, then the optimal transportation problem is stated in the following.

Definition 1.3.2 (Transportation problem) *The discrete optimal transportation problem consists on finding the values of the matrix* (f_{ij})*, which minimizes the expression*

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} f_{ij}, \tag{1.10}$$

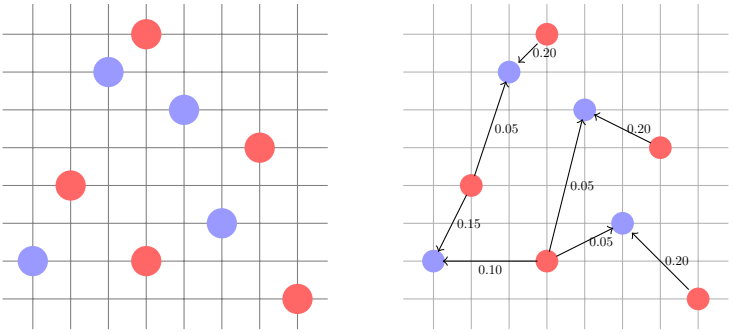
where c_{ij} is the Euclidean distance between the centres of B_i and B_j, and the matrix (f_{ij}) is subject to the constraints

$$\begin{aligned} \sum_{j=1}^{m} f_{ij} &= p_i \text{ for } i = 1, 2, ..., m \\ \sum_{i=1}^{m} f_{ij} &= q_j \text{ for } j = 1, 2, ..., m \\ f_{ij} &\geq 0 \text{ for all } i \text{ and } j. \end{aligned} \tag{1.11}$$

Definition 1.3.3 *By taking $m \to \infty$ over the set of regular sub-boxes $\{B_n\}_{n=1}^{m}$, then the minimum of (1.10) over all transportation plans (f_{ij}) is the Wasserstein distance between the probability mass functions $\{p_n\}_{n=1}^{\infty}$ and $\{q_n\}_{n=1}^{\infty}$.*

Remark. As the time series that we want to compare ($\{f_i\}_{i=1}^{\bar{N}}$ and $\{g_i\}_{i=1}^{\bar{N}}$) are finite, then there are also a finite number of non-zero probabilities in $\{p_n\}_{n=1}^{\infty}$ and $\{q_n\}_{n=1}^{\infty}$. By considering only these non-zero probabilities in the optimal transportation plan 1.3.2, a linear programming problem is set up, which is solved with the *revised simplex* algorithm presented in [18].

In the following we give an example of a discrete optimal transportation problem, which is solved with functions from the R-package *transport*.



In this regular grid (with each cell having edges of length one) we observe an example of two discrete distributions of mass points. If we wish to move the uniform distribution represented by the red circles to the uniform distribution of blue circles, we find an *optimal transportation plan* given by the arrows and numbers on the right. The distance in the behavioural space (the Wasserstein distance) is 2.4072.

Chapter 2

Wasserstein distances on synthetic time series

In this chapter we consider oscillatory solutions of a model developed to describe the dynamics of rotavirus transmission. We first analyse the model with the aid of the MATCONT package [11] for Matlab (which performs numerical bifurcation analysis). Afterwards, we apply the method of Verduyn-Lunel and Muskulus [3] with the purpose of identifying bifurcations in the dynamics, when the solutions of the model are regarded as synthetic time series.

Previous work by Alexandra Teslya [2] resulted in the Susceptible-Infected-Recovered-Susceptible-Infected epidemiological model (2.1). This is a system of ordinary differential equations (ODE's) that describes the transmission dynamics of rotavirus. The five variables of the model represent primary susceptible individuals (denoted by S_1), individuals infected for the first time (denoted by I_1), individuals recovered from the first infection (denoted by R), individuals who lost immunity and became susceptible again (denoted by S_2), and individuals who got infected once more (denoted by I_2).

$$\begin{aligned}
\dot{S_1} &= -\beta(1+\eta\cos(2\pi t))S_1(I_1+\rho I_2) + \xi(1-c-S_1) \\
\dot{I_1} &= \beta(1+\eta\cos(2\pi t))S_1(I_1+\rho I_2) - (\alpha+\xi)I_1 \\
\dot{R} &= \xi c + \alpha I_1 + \gamma I_2 - (\kappa+\xi)R \\
\dot{S_2} &= \kappa R - \nu\beta(1+\eta\cos(2\pi t))S_2(I_1+\rho I_2) - \xi S_2 \\
\dot{I_2} &= \nu\beta(1+\eta\cos(2\pi t))S_2(I_1+\rho I_2) - (\gamma+\xi)I_2.
\end{aligned} \tag{2.1}$$

Since the time scale used in this model is years and the forcing term $(1+\eta\cos(2\pi t))$ reaches its maximum value at times $t = 0, 1, 2, 3, \ldots$, then we set the initial time $t = 0$ to be January 1^{st} to reflect that rotavirus is more transmissible in winter. In other words, the transmission rate β is seasonally driven with amplitude η

The system (2.1) is a *multi-compartment* kind of model (the theory of which is thoroughly discussed in [8]), since the total population is divided into subpopulation compartments represented by S_1, I_1, R, S_2 and I_2. Furthermore, any of these variables gives the proportion of the population belonging to the corresponding compartment (implying that $S_1 + I_1 + R + S_2 + I_2 = 1$).

The model 2.1 describes influx/outflux into/from the different compartments, which represent changes in the epidemiological status of the individuals in the population. This system is better illustrated in figure 2.1.

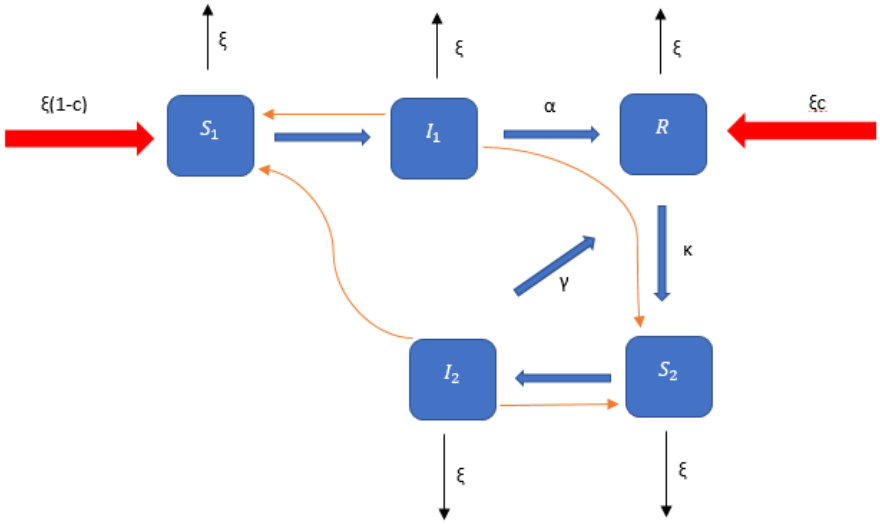


Figure 2.1: Flow diagram of the compartmental model (2.1). Red arrows represent influx given by newborns, black arrows represent out-flux given by deaths, orange arrows represent inter-compartmental interactions that may lead to rotavirus transmission, and blue arrows represent movement of individuals from one compartment to another. Only rates corresponding to non-linear terms of (2.1) are not shown in the diagram.

Movements between compartments are given by rates, which are explained in the following.

- Birth/death rate, denoted by ξ. The model assumes that all newborns belong to the compartment S_1.
- Transmission rate, denoted by β. It gives the (mean) probability of infection for an individual in S_1 when in contact with and infected individual.
- Relative susceptibility of S_2, denoted by ν. This parameter is added because an individual who got once infected with rotavirus is less vulnerable to a second infection.
- Recovery rates for primary and secondary infected individuals, denoted by α and γ, respectively.
- Waning rate of immunity, denoted by κ. This rate describes how immunity (in recovered individuals) decreases in time.
- Vaccination uptake, denoted by c. This takes up values in $[0, 1]$, where $c = 0$ represents no vaccination at all, and $c = 1$ means full vaccination coverage.

- Relative infectivity in I_2, denoted by ρ. This rate takes up values in $[0, 1]$ since an individual (might) becomes less infectious during the second (or subsequent) rotavirus infection.

The model parameters are described in table 2.1, and these are obtained from several sources (WHO [19], Dafilis et al. [20] & Pitzer et al. [21]). Throughout this chapter, only the case of no vaccination ($c = 0$) is considered, since it represents the current situation in the Netherlands.

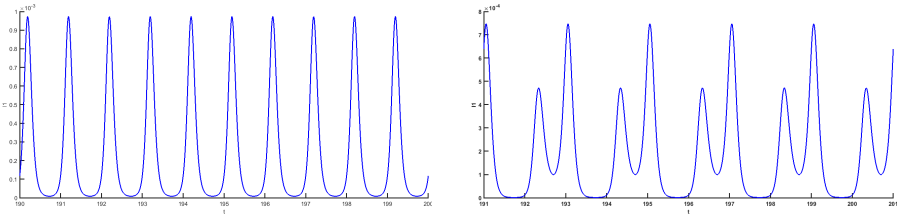
Parameter	Units	Value	Source
ξ	year^{-1}	$[1/125, 1/25]$, mean is $1/81$	[19]
β	(number of people $\times$ year)$^{-1}$	1040	Calculated by setting $R_0 = 20$ and $c = 0$.
ν		$[0.05, 1]$	
ρ		$[0, 1]$	
η		$[0.001, 1]$	[20]
α	year^{-1}	52	[21]
γ	year^{-1}	90	[21]
κ	year^{-1}	$4/3$	[21]
c		$[0, 1]$	Vaccination uptake rate $\times$ vaccine efficacy.

Table 2.1: Table of parameters taken from Teslya's report [2].

As the basic reproduction number (defined in appendix B) of the system (2.1) satisfies $R_0 > 1$, the disease is able to spread in the population, and therefore it is important to analyse the long term solutions.

2.1 Numerical bifurcation analysis

We first simulate model solutions using fixed parameter values $\nu, \rho = 0.2728$, $\eta = 0.1$, taken from table 2.1. We start off with a birth rate per year of $\xi = 0.0125$, and the initial conditions $S_1(0) = 0.9$ and $I_1(0) = 0.1$. After a transient time of 190 years, the solution converges to a cycle with period of one year, as can be seen in figure 2.2a. We proceed to continue this cycle with MATCONT [11] over the domain $\xi \in [0.0075, 0.013]$.



(a) (b)

Figure 2.2: Oscillations of the $I_1(t)$ compartment (y-axis on both plots) of system (2.1). In both plots, the x-axis represents time. In 2.2a the birth rate per year is $\xi = 0.0125$, and in 2.2b it is $\xi = 0.01$.

By observing the bifurcation diagram 2.3 we see that solution dynamics of the model (2.1) undergo three bifurcations when the birth rate ξ is varied within the interval $[0.0075, 0.013]$. One bifurcation on the dynamics occurs at the parameter value $\xi_1 = 0.0119$, and MATCONT [11] labels it as a period-doubling (PD) bifurcation.

Remark. As the software only calculates the moduli of the multipliers at ξ_1 (which is equal to one for two of them), MATCONT [11] relies on the computation of the normal form coefficient as a numerical criterion to detect this kind of bifurcation. Such normal form coefficient has the value $c = -1.34$ [1].

As was explained in 1.1.1, we observe the following effect on the limit cycle solutions of (2.1): the period one cycle becomes unstable and a new stable period two cycle appears when the birth rate decreases and crosses the value ξ_1. We observe in figure 2.2b that the period two cycle is comprised of two peaks.

If the value of ξ keeps decreasing during the continuation, we observe from figure 2.3 that the amplitude of the highest peak smoothly decreases as well, until reaching the next critical value at $\xi_2 = 0.0086$, which MATCONT [11] labels as a limit point of cycles (LPC) bifurcation (with corresponding normal form coefficient -1.11591×10^{16}), followed by another LPC bifurcation at $\xi_3 = 0.0092$ (with corresponding normal form coefficient $3.0066e + 17$).

The two bifurcations $\xi_2 = 0.0086$ and $\xi_3 = 0.0092$ provoke the next effect in the dynamics: we observe the creation of a so-called *bi-stability region*, as is seen in more detailed from picture 2.4 (created with MATCONT [11] as well). The behaviour of solutions within this bi-stability region is as follows: for different values of initial conditions, the corresponding solution trajectory may converge to one or another limit cycle. We explain how to numerically integrate such cycles in section 2.2.

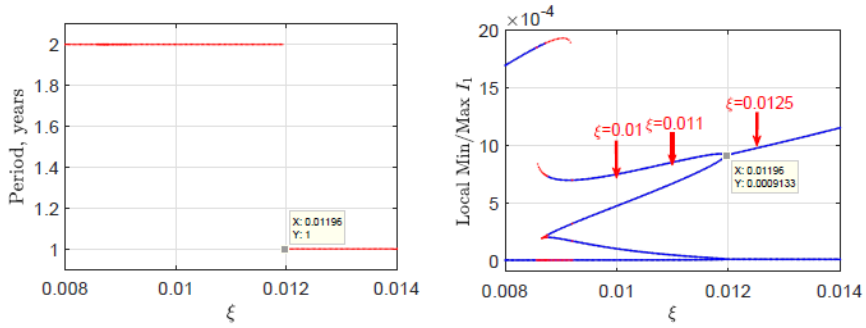


Figure 2.3: On the plot to the left we observe the period of the limit cycle solutions of system (2.1) when the birth rate is varied between $[0.0080, 0.0140]$. On the right plot, we observe the bifurcation diagram. For the interval $\xi \in [0.0086, 0.0119]$, the line above represents the global peak incidence of period-two limit cycles and the line just below represents the other peak incidence, both repeating every two years (as can be seen in the example of figure 2.2b). Diagrams taken from Teslya's report [2].

[1] We note that (in general) the normal form coefficients depend very much on the details of the computations and can be reliably used only if the bifurcation points are computed to high accuracy.

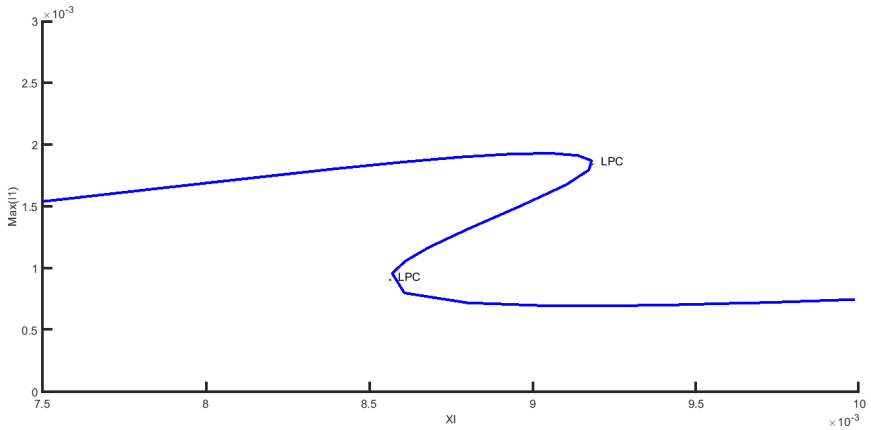


Figure 2.4: The bi-stability region for birth rates within the interval $\xi \in [0.0086, 0.0092]$ is shown in here. The plot represents birth rates on the $x-$ axis, and on the y-axis we observe maximum values of the I_1 compartment for the corresponding cyclic solutions of system (2.1).

2.2 Wasserstein distance analysis

We now perform another numerical analysis, but this time it is based on Wasserstein distances, with the purpose of detecting bifurcations in the dynamics of synthetic rotavirus data. The generation of synthetic data from the model (2.1) is first explained, followed by the Wasserstein distance calculation on these data.

For the sake of consistency with the actual observable function (rotavirus incidence per week), we only address the variable corresponding to observations of rotavirus per week, and for this we consider the second equation in (2.1), the one giving the rate of change in the compartment of primary infected individuals (I_1). If we take the first term in the *r.h.s.* of this equation, and define the function of time $U(t)$ as follows

$$\dot{U} = \beta(1 + \eta\cos(2\pi t))S_1(I_1 + \rho I_2), \tag{2.2}$$

we are left with the influx into this compartment. The solution $U(t)$ of (2.2) can also be regarded as the cumulative (primary) incidence of rotavirus infection given by the transmission model (2.1). In order to make a time series from this, we discretise the time variable with the constant time step $\Delta_t = 1/52$, (so as to approximately divide a year into weeks) obtaining the discrete time variable $t_i = i\Delta_t$, with $i = 0, \ldots, 52$. This time sampling is used to define ten years of synthetic data $I(t_i)$ as:

$$I(t_j) := U(t_{j+1}) - U(t_j),\ j = 1, \ldots, 520. \tag{2.3}$$

In the following we describe the plan used to generate the probability mass functions that will be used in the Wasserstein distance analysis.

1. Numerically integrate the model (2.1) and the influx equation (2.2) with the use of the *ode*45 algorithm of Matlab (for specifics about this algorithm we refer to [22]). After a transient time of 190 years, the numerical (cyclic) solution corresponding to the years $[190, 200]$ is stored. We apply this first step to the following set of initial conditions.

   - For the initial values $S_1(t_0 = 0) = 0.0130$, $I_1(t_0 = 0) = 6.1916e-08$, $R(t_0 = 0) = 0.1019$, $S_2(t_0 = 0) = 0.8851$, $I_2(t_0 = 0) = 1.3215e-06$, and $\xi = 0.0075$. The corresponding solution trajectory converges to a cycle with amplitude greater than one, represented by a point on the line above in the plot of figure 2.4.
   - For the initial values $I_1(t_0 = 0) = 0.000001$, $S_2(t_0 = 0) = .9999999$, and $\xi = 0.0086$. The corresponding solution trajectory converges to a cycle with amplitude less than one, represented by a point on the line below in the plot of figure 2.4.

2. Sample the function $U(t)$ with the time step $\Delta_t = 1/52$.

3. Introduce the variable $I(t)$ as explained in (2.3), this contains ten years of weekly incidence data.

4. Create the delay coordinates time series corresponding to $I(t)$ (from definition 1.2.1), using the embedding dimension $k = 3$.

5. Make a probability mass function from the delay coordinates time series by following the rule (1.6). The *R* package *transport* has this conversion implemented in the function *wpp*.

In order to obtain all the probability mass functions, we proceed with a *numerical continuation* of the cycles obtained in the first step, i.e., we take the last value of each cycle (we may call it x_{last}), move the parameter by Δ_ξ, and use x_{last} as the initial condition for the new numerical integration ($x_{last} = x_0$). This process makes the computation of limit cycles more efficient, since we know from the bifurcation diagram that there is no discontinuity in the amplitude of the cycles.

The above described plan is implemented for a grid of the parameter space given by dividing the interval $[0.0075, 0.013]$ of ξ with a constant step size of $\Delta_\xi = 0.0001$. As a result, 62 time series $I(t)$ are obtained. This number does not equal the number of values that ξ takes on in the grid because (we know from the numerical bifurcation analysis) one can extract two different time series $I(t)$ for values of ξ within the bi-stability region.

2.2.1 Period doubling bifurcation

In the following, we compute the Wasserstein distances corresponding to synthetic time series of incidence obtained through the previous discussion. The parameter values used to generate these discrete solutions are close to the actual values and include the PD bifurcation value $\xi_1 = 0.0119$ (obtained with MATCONT [11]). In figure 2.5, we can observe the Wasserstein distances matrix of time series generated from cycles corresponding to birth rate values ξ ranging from 0.0086 to 0.013.

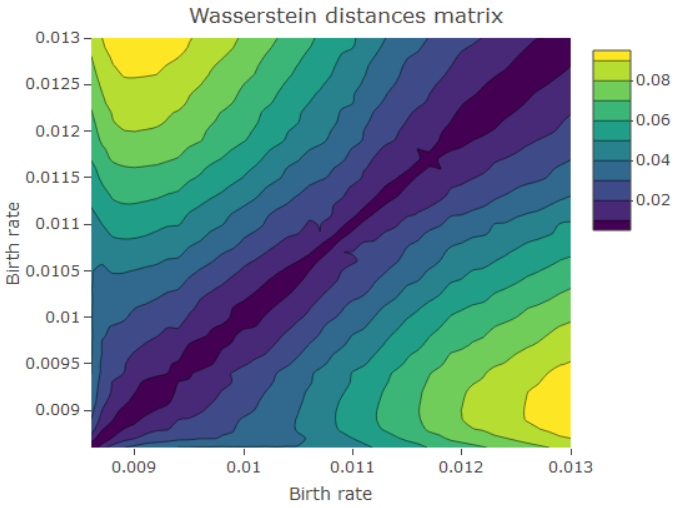


Figure 2.5: The map represents the Wasserstein distance matrix for stable cycles corresponding to birth rate values between $\xi = 0.0086$ to $\xi = 0.013$. Dark colours represent small distances, light colours represent long distances.

2.2.2 Limit point of cycles bifurcation

The next bifurcation to be considered is the one that creates a bi-stability region in the parameter space. Such region is defined as the interval between the two limit point of cycles bifurcations $\xi_2 = 0.0086$ and $\xi_3 = 0.0092$. The effect of these two bifurcations can be seen in detail in figure 2.4: for different values of initial conditions, the solution trajectory may converge to one of the two limit cycles. If we numerically integrate these cycles, we can compare them with the use of the Wasserstein distance, and moreover, compare them with all the cycles considered in figure 2.5.

2.2.3 Complete Wasserstein distances matrix

In figure 2.6 we observe how the Wasserstein distance behaves for all values of ξ within $[0.0075, 0.013]$. We describe this plot by dividing it into the following subregions.

1. Region $A := [0.0075, 0.0092] \times [0.0075, 0.0086]$. Since we only count with delay time series corresponding to large amplitude cycles (as can be seen in the plot of 2.4), we plot the Wasserstein distances matrix between these same cycles, which we call *self distance Matrix*. This gives the dark purple-coloured region at the bottom-left corner of 2.6.

2. Region $B := [0.0075, 0.0092] \times [0.0086, 0.013]$. In this region we plot the Wasserstein distances between delay time series corresponding to large amplitude cycles (x-axis) and small amplitude cycles (y-axis). We observe that the Wasserstein distance increases as the value of ξ on the y-axis increases. In particular, there is a clear discontinuity when ξ exceeds the LPC bifurcation value $\xi_2 = 0.0086$.

3. Region $C := [0.0092, 0.013] \times [0.0075, 0.0086]$. In this region we plot the Wasserstein distances between delay time series corresponding to small amplitude cycles (on the x-axis) and large amplitude cycles (y-axis). The situation is similar to the one observed in region B, as the Wasserstein distance increases when the value of ξ on the x-axis increases. This time the discontinuity occurs at the LPC bifurcation value $\xi_3 = 0.0092$.

4. Region $E := [0.0092, 0.013] \times [0.0086, 0.013]$. Since we count with delay time series corresponding to small amplitude cycles only, we plot the self distance Wasserstein matrix for this region. As the amplitudes corresponding to small amplitude cycles do not vary much, the Wasserstein distances remain rather low.

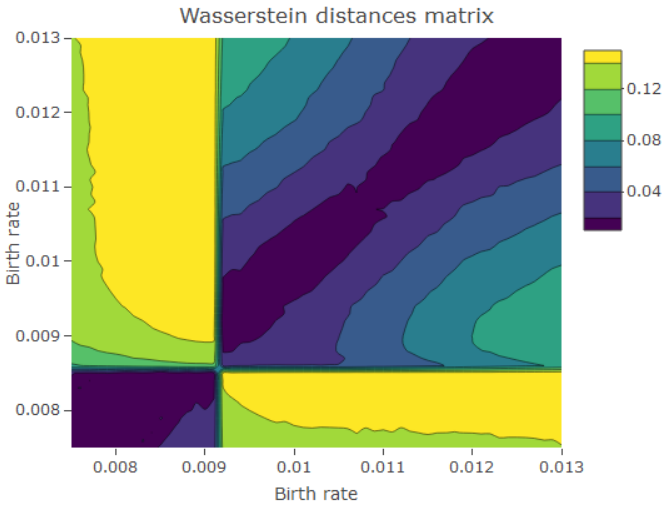


Figure 2.6: The map represents the Wasserstein distance matrix between cycles with high amplitude and low amplitude cycles. There is a discontinuity in the matrix when the high amplitude cycles are compared with small amplitude cycles from outside of the bi-stability region.

Chapter 3

Wasserstein distances for rotavirus incidence in German states

We obtained from the *Robert Koch Institute* website a dataset of rotavirus incidence in the 16 German states [23]. The measurements are recorded on a weekly basis, starting from the first week of the year 2001 up until the last week of 2017. Recall that by *incidence* we mean the number of notified cases per $100,000$ persons.

In this chapter, the time series analysed with the use of Wasserstein distances are of the following form:

1. Measurements of rotavirus incidence along time for each of the German states, so one time series of weekly reports starting from the beginning of year 2001 and comprising a total of 934 weeks.

2. For each of the German states, we split the rotavirus measurements in two halves, and we consider the corresponding time series.

3.1 Wasserstein distances between German states

We start with the 17 year-long time series for each of the German states, which can be seen in the series of figures 3.1 and 3.2. As we can observe in these, the incidence seems to follow a cyclic pattern with peaks repeating every winter. If we regard the incidence as an observable function h defined on an epidemiological dynamical system in the population (one example being the model of Teslya (2.1)), it would then make sense to use the theory of chapter 1 in order to find the degree of similarity between the regional dynamics of rotavirus.

In the following we make use of the delay embedding map $F(h, \Psi, \tau = \text{one week})$ on the regional time series plotted in the figures 3.1 and 3.2. The report of Teslya [2] indicates the existence of limit cycles in the dynamics of rotavirus (as it was discussed in chapter 2). Since the dimension of a limit cycle

is one, the Delay Embedding Theorem 1.2.1 suggests that the use of the embedding dimension $k = 3$ makes the delay coordinates time series thus obtained a good dynamical representation of the actual attractor.

Remark. We do not care for the optimal value of the embedding dimension k, since it is not the purpose of this thesis project to properly reconstruct the attractor.

From the delay coordinates thus created we obtain the probability mass functions $\{\mu_i\}_{i=1}^{16}$ (as defined in the expression (1.6)), one for each of the German states, upon which the Wasserstein distances will be computed. For this we use the *R* package *transport*, which has implemented in it the conversion from (coordinate) time series to probability mass functions of the form (1.6). Furthermore, the package includes the function *wasserstein* which we use to solve the corresponding discrete transportation problem introduced in the subsection 1.3.1.

We store the Wasserstein distances between all states in a matrix M (shown in table 3.1). Since it is rather difficult to interpret this matrix directly, we choose to analyse it by using *classical multidimensional scaling* (abbreviated as cMDS), a method which places a point for each state in multidimensional space, where these points satisfy (in a Euclidean fashion) the distances in matrix M (as it is explained in detail in appendix A). As a result of applying cMDS on the matrix M, we get the two-dimensional plot of figure 3.3, which we compare with a map of Germany 3.4.

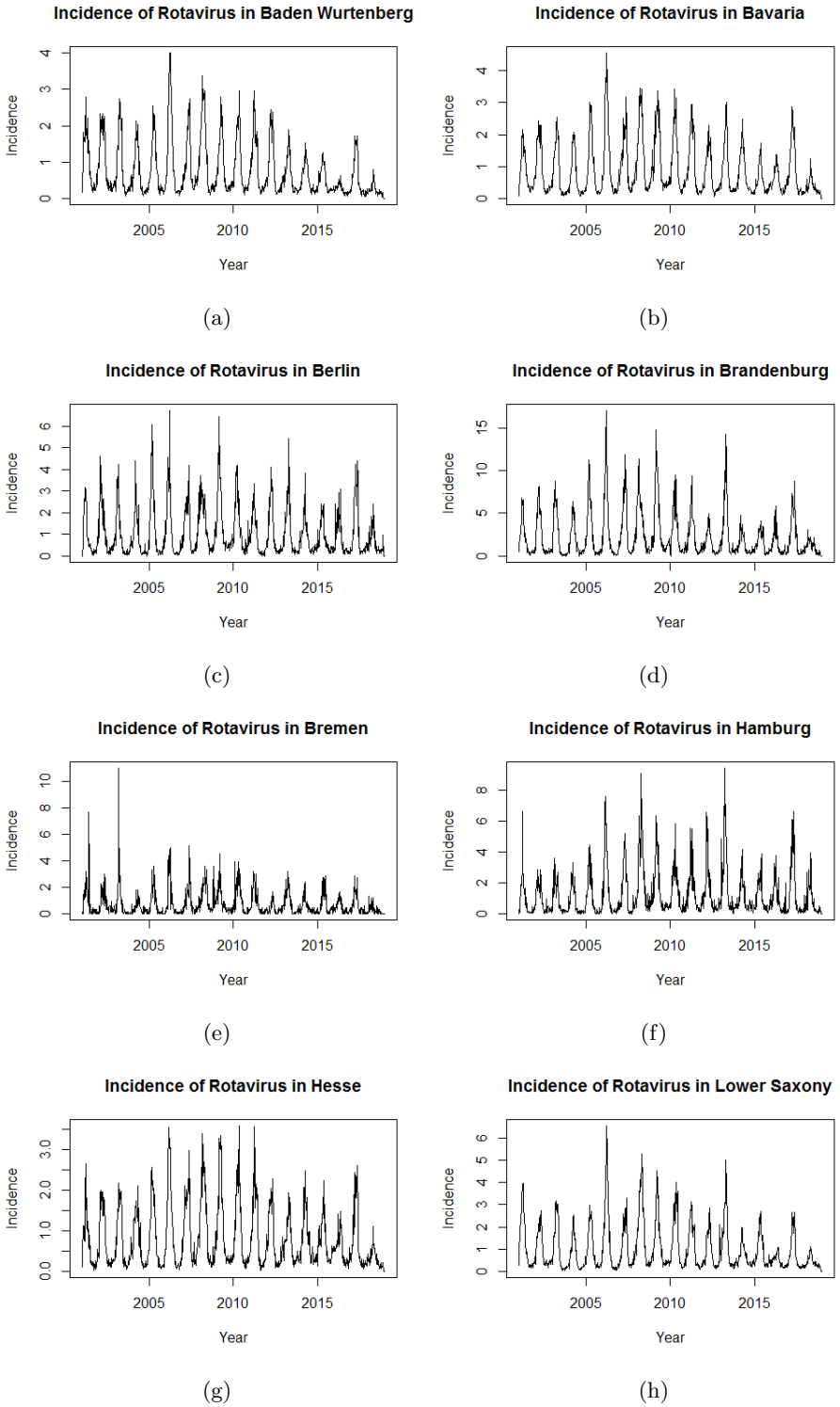


Figure 3.1: Notified incidence (number of cases per $100,000$) of rotavirus on each German state. The years covered range from 2001 to 2018.

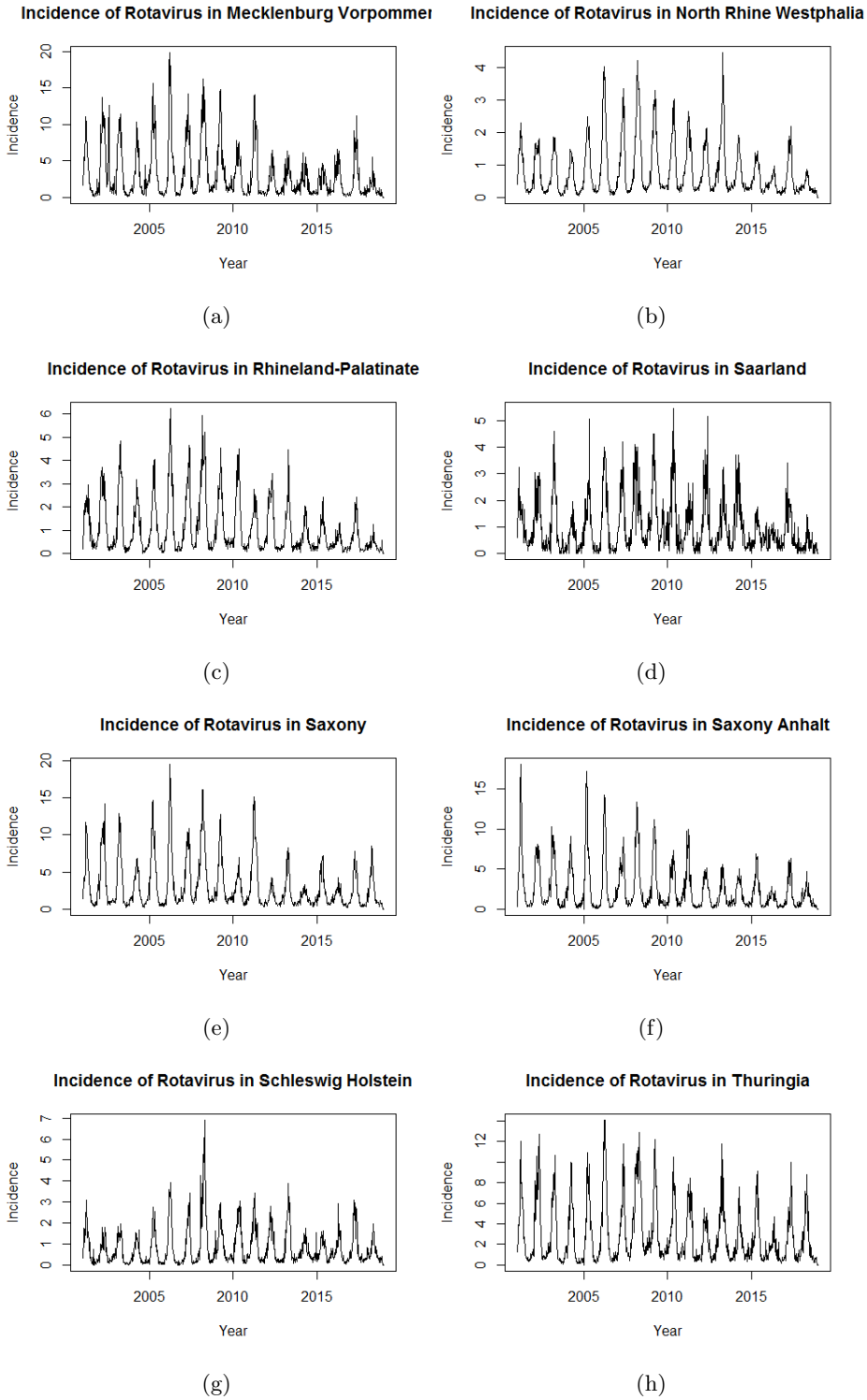


Figure 3.2: Notified incidence (number of cases per $100,000$) of rotavirus on each German state. The years covered range from 2001 to 2018.

	Baden	Bavaria	Berlin	Brandenburg	Bremen	Hamburg	Hesse	Mecklenburg	Lower-Saxony	North-Rhine-Westphalia	Rhineland	Saarland	Saxony	Saxony-Anhalt	Schleswig-Holstein	Thuringia
Baden	0.00	0.22	0.60	2.76	0.39	1.00	0.16	4.24	0.44	0.13	0.57	0.55	4.17	3.13	0.23	3.84
Bavaria	0.22	0.00	0.43	2.56	0.47	0.82	0.18	4.03	0.25	0.18	0.39	0.40	3.96	2.92	0.22	3.63
Berlin	0.60	0.43	0.00	2.19	0.64	0.45	0.52	3.66	0.24	0.54	0.18	0.31	3.60	2.56	0.46	3.26
Brandenburg	2.76	2.56	2.19	0.00	2.75	1.85	2.68	1.50	2.33	2.70	2.21	2.29	1.45	0.51	2.62	1.18
Bremen	0.39	0.47	0.64	2.75	0.00	0.94	0.40	4.22	0.57	0.40	0.65	0.55	4.18	3.12	0.32	3.83
Hamburg	1.00	0.82	0.45	1.85	0.94	0.00	0.91	3.31	0.62	0.94	0.48	0.56	3.27	2.22	0.82	2.92
Hesse	0.16	0.18	0.52	2.68	0.40	0.91	0.00	4.16	0.37	0.16	0.49	0.46	4.09	3.05	0.21	3.76
Mecklenburg	4.24	4.03	3.66	1.50	4.22	3.31	4.16	0.00	3.81	4.17	3.68	3.76	0.61	1.17	4.09	0.74
Lower-Saxony	0.44	0.25	0.24	2.33	0.57	0.62	0.37	3.81	0.00	0.37	0.21	0.33	3.74	2.70	0.34	3.41
North-Rhine-Westphalia	0.13	0.18	0.54	2.70	0.40	0.94	0.16	4.17	0.37	0.00	0.50	0.50	4.10	3.06	0.20	3.77
Rhineland	0.57	0.39	0.18	2.21	0.65	0.48	0.49	3.68	0.21	0.50	0.00	0.35	3.62	2.58	0.44	3.28
Saarland	0.55	0.40	0.31	2.29	0.55	0.56	0.46	3.76	0.33	0.50	0.35	0.00	3.72	2.67	0.41	3.37
Saxony	4.17	3.96	3.60	1.45	4.18	3.27	4.09	0.61	3.74	4.10	3.62	3.72	0.00	1.11	4.03	0.68
Saxony-Anhalt	3.13	2.92	2.56	0.51	3.12	2.22	3.05	1.17	2.70	3.06	2.58	2.67	1.11	0.00	2.99	0.98
Schleswig-Holstein	0.23	0.22	0.46	2.62	0.32	0.82	0.21	4.09	0.34	0.20	0.44	0.41	4.03	2.99	0.00	3.70
Thuringia	3.84	3.63	3.26	1.18	3.83	2.92	3.76	0.74	3.41	3.77	3.28	3.37	0.68	0.98	3.70	0.00

Table 3.1: Wasserstein distances matrix generated from probability mass functions corresponding to rotavirus incidence in German states.

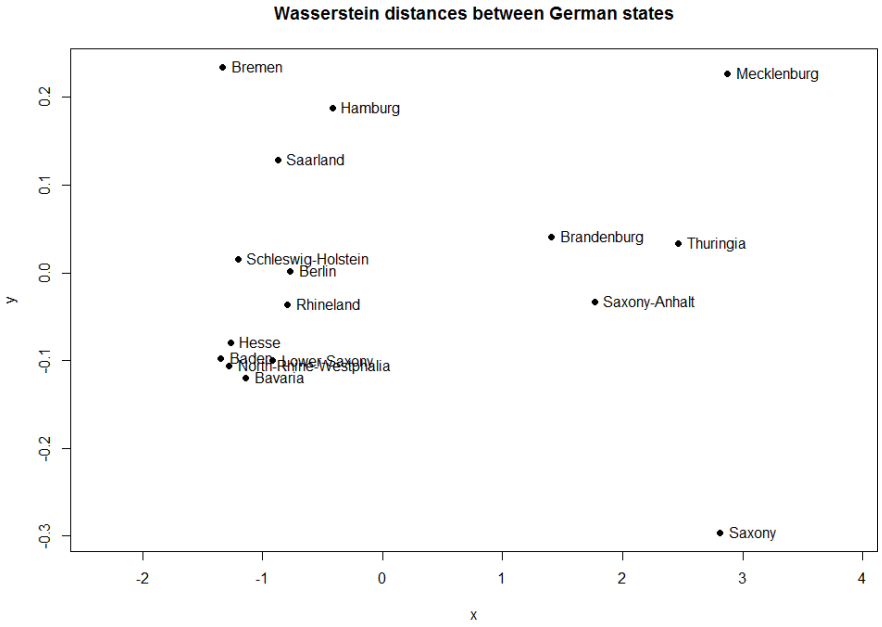


Figure 3.3: Classical MDS for the Wasserstein distances matrix 3.1. The years covered range from 2001 to 2017, and the weights of these two dimensions are given by the GOF $(0.9842, 0.9857)$.

Figure 3.4: German states.

3.2 Wasserstein distances within German states

We follow to split the 934 week-long time series into two halves, for each of the German states. The first half corresponds to incidence of rotavirus from the first week of the year 2001 to the week 49 of 2009, and the second half covers from the week 50 of 2009 to the week 45 of the year 2018. We use the resulting time series to compute Wasserstein distances within region. The motivation to make this analysis comes from observing the time series 3.1 and 3.2, where we can see an overall difference in the cyclic pattern between the first and second halves of records.

The Wasserstein distances within regions can be seen in the table 3.2, where the largest distances correspond to the states of the former Eastern Germany, namely Brandenburg, Mecklenburg Vorpommern, Saxony, Saxony Anhalt, and Thuringia. When observing the time series corresponding to these states, we can confirm that these are the ones with less regularity, and with an overall tendency to lower rotavirus incidence in the second half of the measurements. As the values of these Wasserstein distances are so high that are similar to the largest distances between regions (as seen from the table 3.1), it is natural to suspect that a major change happened within the internal conditions of these states. In [24] we find the dates when rotavirus vaccination was included into the German state guidelines.

State	Wasserstein distance	Inclusion date for rotavirus vaccination
Mecklenburg-Vorpommern	3.35	July-2009
Saxony	2.87	January-2008
Saxony-Anhlat	2.41	July-2013
Thuringia	1.84	October-2009
Brandenburg	1.6	January-2009
Rhineland-Palatinate	1.01	July-2013
Baden-Württemberg	0.58	July-2013
Lower Saxony	0.54	July-2013
Saarland	0.49	July-2013
Berlin	0.47	July-2013
Bremen	0.45	July-2013
Bavaria	0.43	July-2013
Hamburg	0.42	July-2013
North Rhine-Westphalia	0.35	July-2013
Hesse	0.33	July-2013
Schleswig-Holstein	0.32	March-2011

Table 3.2: Correspondence between Wasserstein distances within German states and introduction dates of rotavirus vaccination.

Chapter 4

Discussion and conclusions

4.1 Wasserstein distance between synthetic series

Unfortunately, in the Figure 2.6 we do not observe any discontinuity between the distances around the PD-parameter value $\xi = 0.0119$. As the period doubling bifurcation produces cycles occupying very similar Borel sets for the double-cycle and for the one-cycle, the invariant measure is not really a good method to distinguish between these two different dynamical regimes. Thus, we cannot rely on the method of Verduyn-Lunel and Muskulus [3] to detect a period doubling bifurcation on the underlying dynamics of data.

On the other hand, we observe in Figure 2.6 a clear discontinuity when the two stable cycles within the bi-stability region are compared. This is explained by the computation of the Wasserstein distance: the invariant measures corresponding to delay coordinates time series are defined over Borel sets, therefore, as the limit cycles within the bi-stability region occupy different Borel sets (this can be observed in the bifurcation diagram 2.4), it makes sense that the transportation plan will imply larger cost than a corresponding one for cycles that occupy Borel sets very close to each other (which is what happens at the PD bifurcation). The same (as for the bi-stability region) can be said for the Wasserstein distances between large and small amplitude limit cycles in general, which are created when LPC bifurcation of cycles occurs. Thus, the method of Verduyn-Lunel and Muskulus [3] can be used for the numerical detection of LPC bifurcations of cycles from data.

4.2 Wasserstein distances between German states

Before starting with this discussion, it is important to remark that the goodness of fit (denoted as GOF and defined in A.0.1) corresponding to the plot 3.3 is $GOF = (0.9842, 0.9857)$, which implies that we can rely on this plot to draw conclusions from the Wasserstein distances analysis.

We observe a cluster of points on the left side of figure 3.3, which implies similar rotavirus dynamics in the corresponding regions, e.g. the Wasserstein distance analysis suggests that the spread of rotavirus in the regions of $North-$

Rhine − Westphalia and *Bavaria* show similar dynamics. A claim that turns out to be remarkably accurate when we observe the time series corresponding to these regions: the qualitative behaviour in 3.2b and 3.1b is very similar, and furthermore, if we note that the scales in both plots are the same, then we realize that the quantitative character is also very similar. On the other hand, points far away from each other in the plot 3.3 correspond to time series with different behaviour, e.g. the Wasserstein distance analysis suggests that the spread of rotavirus in the regions of *Saxony* and *Bremen* follows different dynamics. This claim is confirmed when we observe the figures 3.1e and 3.2e: these plots show major qualitative and quantitative differences in the regional incidences.

When we compute cMDS with $l = 1$, we obtain a GOF value of $(0.977, 0.978)$. Therefore, we can neglect the rather big distance between Mecklenburg and Saxony in the plot 3.3, as their x-coordinates are 2.8747 and 2.8141, respectively. Indeed, as the Wasserstein distance between these two states is 0.61 (the lowest for these two, as seen in table 3.1) and we can observe a strong similarity in the time series 3.2a and 3.2e, the relation stated in the last paragraph is confirmed.

4.3 Wasserstein distances within German states

As the Wasserstein distances within German states is remarkably high for former Eastern Germany states, we looked into a major change in the composition of the population, such as the introduction of vaccination against rotavirus. As vaccination intake results in newborn individuals being considered directly into the R compartment of the flow chart 2.1, it is understandable that the introduction of vaccination would decrease the disease transmission. Therefore, it is no coincidence that high Wasserstein distances correspond to the introduction of vaccination in states with high rotavirus incidence.

4.4 Conclusions

- We cannot know by using the approach of Verduyn-Lunel and Muskulus [3] if a period doubling bifurcation occurred in the dynamics of rotavirus.
- We can use their approach to find major changes in the patterns, such as the sudden decrease in rotavirus incidence that suggested the introduction of vaccination in Eastern Germany.
- It is not necessary for the actual trajectory (from which the measurements of rotavirus are taken) to lie on the attractor, as it can be close to it and the Takens embedding theorem will still hold.

4.5 Recommendations for further research

- As the Takens embedding theorem holds for strange attractors as well, it might be worth to try using the method of Verduyn-Lunel and Muskulus [3] for time series more irregular than the ones for rotavirus.

- As we could see from the clear detection of the LPC bifurcation, we can rely on Wasserstein distances to spot bifurcations that produce a radical change in the dynamical regime (in terms of Borel sets), such as the Andronov-Hopf bifurcation, and therefore, dynamical systems that may have been affected by this kind of bifurcations can be analysed with the method.

Bibliography

[1] S. Hahné, M. Hooiveld, H. Vennema, A. van Ginkel, H. de Melker, J. Wallinga, W. van Pelt, and P. Bruijning-Verhagen, "Exceptionally low rotavirus incidence in the netherlands in 2013/14 in the absence of rotavirus vaccination," *Eurosurveillance*, vol. 19, no. 43, 2014.

[2] A. Teslya, "Rotavirus: insights from mathematical modeling," 2018. Internship report. National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

[3] M. Muskulus and S. Verduyn-Lunel, "Wasserstein distances in the analysis of time series and dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 240, no. 1, pp. 45 – 58, 2011.

[4] M. Muskulus, S. Houweling, S. Verduyn-Lunel, and A. Daffertshofer, "Functional similarities and distance properties," *Journal of Neuroscience Methods*, vol. 183, no. 1, pp. 31 – 41, 2009. BrainModes: A Principled Approach to Modeling and Measuring Large-Scale Neuronal Activity.

[5] M. Muskulus, A. M. Slats, P. J. Sterk, and S. Verduyn-Lunel, "Fluctuations and determinism of respiratory impedance in asthma and chronic obstructive pulmonary disease," *Journal of Applied Physiology*, vol. 109, no. 6, pp. 1582–1591, 2010. PMID: 20813978.

[6] R. Moeckel and B. Murray, "Measuring the distance between time series," *Physica D: Nonlinear Phenomena*, vol. 102, no. 3, pp. 187 – 194, 1997.

[7] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. Springer-Verlag New York, 1983.

[8] O. Diekmann, H. Heesterbeek, and T. Britton, *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton University Press, 2013.

[9] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory (second ed.)*. Berlin, Heidelberg: Springer-Verlag, 1998.

[10] P. Hartman, *Ordinary Differential Equations*. Society for Industrial and Applied Mathematics, second ed., 2002.

[11] W. Govaerts, Y. A. Kuznetsov, and B. Sautois, *MATCONT*, 2006 (Accessed May 16, 2019). `http://www.scholarpedia.org/article/MATCONT`.

[12] J. P. Crutchfield, "Between order and chaos," *Nature Physics*, vol. 8, no. 1, p. 17, 2012.

[13] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980* (D. Rand and L.-S. Young, eds.), (Berlin, Heidelberg), pp. 366–381, Springer Berlin Heidelberg, 1981.

[14] H. Whitney, "Differentiable manifolds," *Annals of Mathematics*, vol. 37, no. 3, pp. 645–680, 1936.

[15] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579–616, Nov 1991.

[16] A. Lasota and M. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*. Applied Mathematical Sciences, Springer New York, 1998.

[17] R. G. Bartle, *The Elements of Integration and Lebesgue Measure*. New York: John Wiley & Sons, 1995.

[18] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*. Springer Publishing Company, Incorporated, 2015.

[19] "World health organisation: Netherlands." `https://www.who.int/countries/nld/en/`. Accessed: 2018-07-20.

[20] M. P. Dafilis, F. Frascoli, J. McVernon, J. M. Heffernan, and J. M. McCaw, "The dynamical consequences of seasonal forcing, immune boosting and demographic change in a model of disease transmission," *Journal of Theoretical Biology*, vol. 361, pp. 124 – 132, 2014.

[21] V. E. Pitzer, J. Bilcke, E. Heylen, F. W. Crawford, M. Callens, F. De Smet, M. Van Ranst, M. Zeller, and J. Matthijnssens, "Did large-scale vaccination drive changes in the circulating rotavirus population in belgium?," *Scientific reports*, vol. 5, p. 18585, 2015.

[22] L. F. Shampine and M. W. Reichelt, "The matlab ode suite," *SIAM J. Sci. Comput.*, vol. 18, pp. 1–22, Jan. 1997.

[23] "Database of cases of notifiable diseases." `https://www.rki.de/EN/Content/infections/epidemiology/SurvStat/survstat_node.html`. Accessed: 2018-12-12.

[24] S. Dudareva-Vizule, J. Koch, M. an der Heiden, D. Oberle, B. Keller-Stanislawski, and O. Wichmann, "Impact of rotavirus vaccination in regions with low and moderate vaccine uptake in germany," *Human Vaccines & Immunotherapeutics*, vol. 8, no. 10, pp. 1407–1415, 2012.

[25] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. New York, NY: Springer, 2005.

[26] "Classical (metric) multidimensional scaling." `https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/cmdscale`. Accessed: 2019-06-21.

Appendix A

Multidimensional scaling

The objective of *classic multidimensional analysis* (denoted by cMDS) is to retrieve a low-dimensional visualisation of points, such that the distances between any two of these points satisfy a given distance matrix M_{ij}. For a thorough discussion of cMDS we refer to the book of Borg and Groenen [25].

The set of points $\{x_n\}$ retrieved in chapter 4 (figure 3.3) lies in an Euclidean space, but it represents a set of dynamical systems. Therefore, we shall interpret the Euclidean distance between any two of the points in $\{x_n\}$ as a measure of similarity between the corresponding dynamical systems.

Classical multidimensional analysis works as follows. We first assume that a given matrix M_{ij} of $n \times n$ shows the Euclidean distance between n (unknown) points lying in $\mathbb{R}^m$ (with $m < n$), i.e. the (unknown) set $\{x_i\}_{i=1}^n \subset \mathbb{R}^m$ (with $m < n$) will satisfy $d(x_i, x_j) = M_{ij}$, for $1 \leq i, j \leq n$. We want to find the $(n \times m)$ matrix X such that $\{X_{ij}\}_{j=1}^m = x_i$, for $1 \leq i \leq n$. We do this by first obtaining the matrix $B = XX'$ from the relation

$$B = -\frac{1}{2} J M^2 J, \tag{A.1}$$

with $J = I - \frac{1}{n} 1_n (1_n)'$. Note that in here, I is the $n \times n$ identity matrix and 1_n is a column vector made up of n ones. Furthermore, $'$ denotes the transpose of a matrix.

The cMDS coordinate points X to be obtained from B come from the eigendecomposition of B. If we take $B = Q \Delta Q'$, where Δ is a diagonal matrix composed of the eigenvalues of B, and Q is the matrix whose columns are the corresponding eigenvectors, then we may write $B = (Q\Delta^{1/2})(Q\Delta^{1/2})' = XX'$.

We proceed to order the matrix Δ so that $\Delta_{ii} \geq \Delta_{jj}$ if $i < j$, and to rearrange the matrix Q accordingly. Furthermore, we choose a number l so that Q_l is the matrix whose columns are the first l columns of Q. The matrix of coordinate points obtained from cMDS is thus:

$$X := Q_l \Delta_l^{1/2}. \tag{A.2}$$

If $l > 3$ we will face the problem of not being able to plot X, and therefore we need to choose $l \leq 3$. As this might imply the loss of information in the

graphic setting, we consider the percentage explained by l dimensions of the classical multidimensional scaling. We will do this by reporting the so called *goodness of fit*, which is implemented in the R function *cmdscale* (as seen in the documentation webpage [26]).

Definition A.0.1 *Goodness of fit is a numeric vector of two coordinates* (g_1, g_2) *defined as*

$$g_i = \frac{\sum_{j=1}^{l} \lambda_j}{\sum_{j=1}^{n} T_i(\lambda_j))}, \ i = 1, 2, \tag{A.3}$$

where l *is the dimension chosen for the visual representation of the cMDS process,* $\lambda_j = \Delta_{jj}$ *for* $j = 1, \ldots, n$ *(the ordered eigenvalues in* Δ*), and* $T_1(\lambda_j) = abs(\lambda_j)$ *and* $T_2(\lambda_j) = max(\lambda_j, 0)$.

Appendix B

Epidemiology

The basic reproduction number (denoted by R_0), plays a central role in the study of the long term dynamics of an infectious disease in the population, and it is defined as follows.

Definition B.0.1 (Basic reproduction number of an infectious disease) *If we assume an initial demographic setting where a population consisting of $N+1$ individuals is comprised of one infected individual and N individuals being susceptible to become infected, then R_0 is defined as the average number of individuals infected (in the long run) by the primary infected.*

Once the computation of the basic reproduction number is obtained, a first criterion to assess whether the corresponding disease might be dangerous is given by the following.

- If $R_0 < 1$ then the disease will not be spread in the population.
- If $R_0 > 1$ then the disease might spread within the population.