Sparse Identification of Epidemiological Models from Empirical Data

by

Jonathan Horrocks

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Applied Mathematics

Waterloo, Ontario, Canada, 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Current modelling practices in mathematical epidemiology are predicated on mechanisms stemming from theoretical assumptions, such as mass action incidence. Deterministic disease models can describe many patterns observed in empirical incidence data but challenges remain in creating accurate, parsimonious models that offer predictive value. Recent advances in data-driven techniques give rise to new model discovery methods that forego theoretical assumptions and attempt to create sparse, dynamic models directly from real-world data. Our goal is to apply these techniques to empirical case notification data of epidemiological systems, to either confirm current practices or give new insight not accessible by human intuition.

We adapt a recently developed technique called Sparse Identification of Nonlinear Dynamics (SINDy), which has demonstrated ability to recover governing equations of complex dynamical systems. To lend insight into this process, the SINDy algorithm was first applied to simulated data from various forms of the SIR model, a standard compartmental model of epidemics. Several conversion processes were then utilized to recover both the susceptible and infectious classes from raw incidence data. Finally, the SINDy algorithm was applied to empirical data from measles, varicella, and rubella datasets, three diseases that offer contrasting dynamic behaviour, and the resulting time-series and model coefficients were analysed.

The resulting models closely mimic the dynamics of the empirical data, most notably the frequency of epidemics, for all three diseases considered. The coefficients discovered exhibit sparsity, though not to the extent that current compartmental models do. Similarities between the discovered model equations and fitted SIR models can be noted, including a strong dependence on the cross-term corresponding with the mass action incidence mechanism. These encouraging results indicate this data-driven technique may be of use in verifying and improving current theoretical models in mathematical epidemiology.

Acknowledgements

I would, first and foremost, like to thank my supervisor Chris Bauch for his guidance and support over the last two years. The eventual success of this thesis would not have been possible without him.

Second, I would like to thank both my labmates and the members of the Toronto Maple Leafs Fan Club for making my tenure in Waterloo a memorable one.

Finally, thank you to my family, and especially my wife, for their love and support throughout my university career. There's been a lot of math in the last six years and not every moment has been exhilarating, but they are why I am where I am today.

Dedication

This thesis is dedicated to my wife, who is my stable equilibrium.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Mathematical Epidemiology

The work discussed in this thesis focuses on the application of dynamic statistical modelling techniques to the field of mathematical epidemiology. It is important, then, to have an understanding of the history and theory of this relatively modern field. Mathematical epidemiology applies mathematical methods, specifically methods of studying dynamical systems, to the spatiotemporal analysis of infectious diseases. The ability to create effective mathematical models is imperative to epidemiology, as experimental methods do not naturally lend themselves to the study of infectious diseases, especially in the case of epidemics.

1.1.1 History

In 1662, shortly before the Great Plague of London, the English statistician John Graunt published his book “Natural and Political Observations made upon the Bills of Mortality” [1]. In this work he estimated comparative risk of mortality caused by the current bubonic plague epidemic against other causes of death. It is this analysis that is considered the earliest attempt to use mathematical theory to explain epidemiological outbreaks. Towards the end of the 18th century, the Swiss mathematician Daniel Bernoulli published what is considered the first epidemiological model to advocate for inoculation against smallpox [2]. The first major stride in creating what is now known as mathematical biology was taken by William Hamel when he applied the Law of Mass Action (though this is alleged to be unwitting on his part) to create a simple epidemic model in discrete time [3, 4]. It is this law that motivated Kermack and McKendrick in 1927 to develop theory that is used as the foundation of many modern epidemiological models [5].

1.1.2 Compartmental Modelling

A *compartmental model* is a mathematical model of a population that classifies each member of the population as one of multiple categories, known as compartments. These models are structured such that members can be transferred from one compartment to another, where this transfer is governed by dynamic model equations [6, 7, 8].

The prototypical compartmental model is the SIR model, stemming from the work of Kermack and McKendrick in 1927 [5]. The system they constructed models the spread of an infectious disease by separating a given population into three compartments:

- *Susceptible* (S): Each member is susceptible to the disease.
- *Infectious* (I): Each member has been infected by the disease and is infectious to susceptible members they come in contact with.
- *Recovered* (R): Each member has recovered from the disease and is now immune.

The variables S, I, and R represent the number of members in the respective compartment. Given that members can be transferred between compartments, these variables are actually functions of time and should be written as $S(t), I(t)$, and $R(t)$. Under the most basic representation of the SIR model, these state variables are governed by the dynamical system given by the differential equations:

$$\begin{aligned} S'(t) &= -\beta S(t)I(t) && (1.1)\\ I'(t) &= \beta S(t)I(t) - \gamma I(t) && (1.2)\\ R'(t) &= \gamma I(t) && (1.3) \end{aligned}$$

In the standard formulation of the SIR model, several key assumptions are made. First, the population is assumed to be constant throughout the duration of the epidemic, such that

$$S(t^*) + I(t^*) + R(t^*) = N \equiv \text{constant}$$

for all $t^* \geq t_0$. Often the system will be scaled such that $N = 1$, meaning that the state variables now represent the proportion of the population that are members of the respective compartment.

Another important assumption is that the system adheres to the principle of *mass action mixing*, which is the primary mechanism behind the dynamic behaviour of the model [7, 8]. This principle was originally used to describe the rate of a well-mixed chemical reaction by relating it to the concentration of reactants [9]. In mathematical epidemiology it predicts that, given a homogeneous population, the rate at which members of the susceptible class become infected will be directly proportional to the size of the susceptible and infectious classes. The rate at which this infection occurs is dictated by the encounter rate β. In the most basic form of the SIR model this parameter is assumed to be temporally invariant, though later work shows this assumption to be an invalid one. Seasonally-varying alternatives are discussed in the next section. The only other parameter of the model (γ) controls the rate at which individuals recover from the disease and is assumed to be constant.

It should be noted that there is no closed form solution to the SIR system of equations (Eqs. 1.1 - 1.3), so numerical simulations must be used to obtain realizations of the system. Below is shown a simulation of the SIR model with certain parameters.

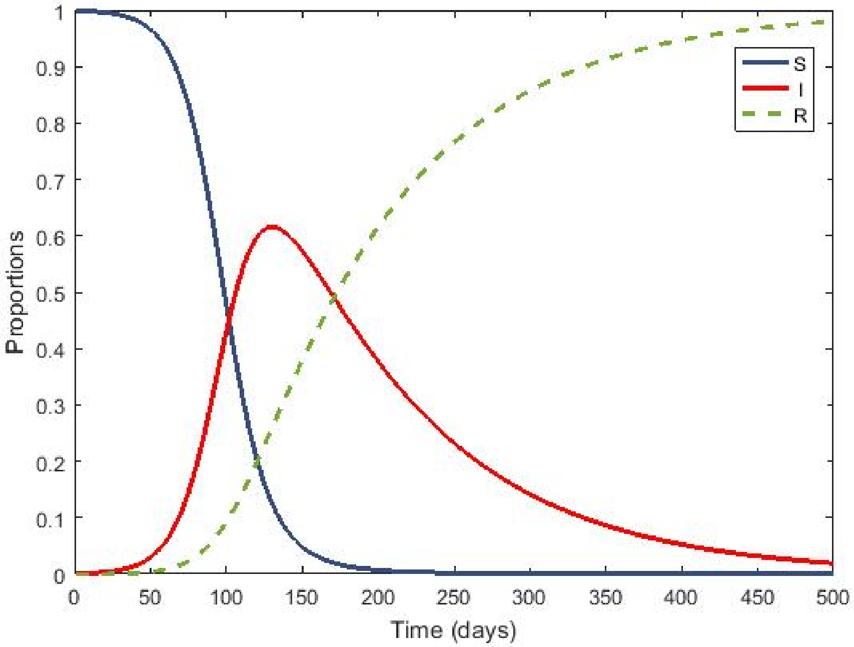


Figure 1.1: A simulation of the basic SIR model with initial conditions $S_0 = 0.999, I_0 = 0.001, R_0 = 0$ and parameters $\beta = 0.5$ day^{-1}, $\gamma = 0.1$ day^{-1}

This seminal model, while accurate in modelling some infectious diseases for a single infection in isolation, fails in recreating disease dynamics from reoccurring infections. This is due to two notable simplifying assumptions: a constant transmission rate and the lack of birth/death rates. Subsequent developments have expanded the model to relax both assumptions, resulting in more realistic dynamics [10, 11, 12, 13, 14].

1.1.3 Time-Varying Mass Action Transmission

Th mass action transmission parameter (β in Eqs. 1.1-1.3) has been noted to be non-constant for most notable infectious diseases, especially in diseases most prevalent among school-aged children [11, 15, 16, 17], as there is a notable shift in individual contact when the school season begins. Estimates of this parameter can be done by giving an approximate recursion relation for the disease incidence:

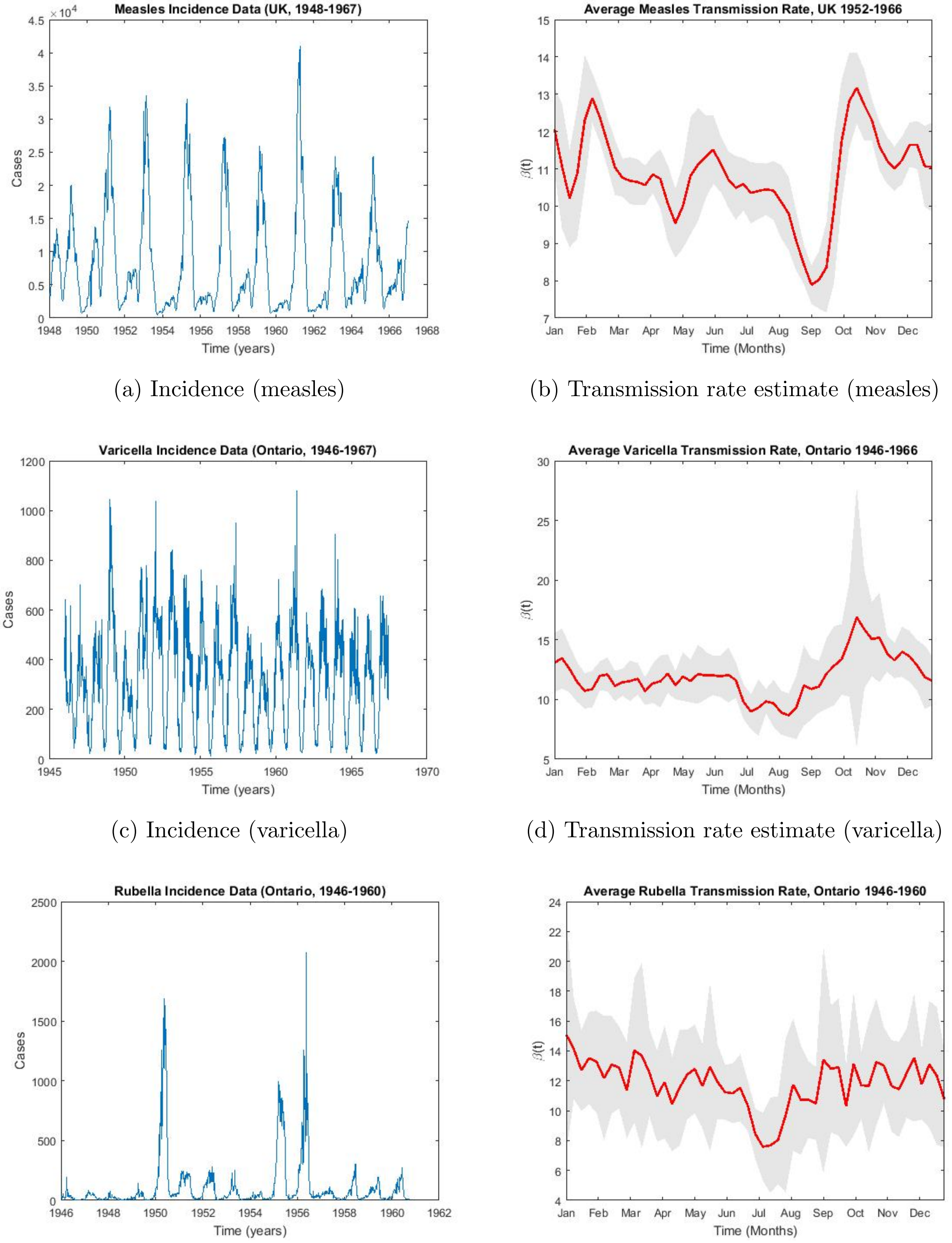
$$C_{t+1} = \beta S_t C_t \tag{1.4}$$

where C_t is the number of cases at time t. Therefore, given sufficient case and susceptible population data, rough estimates for a time-variant parameter $\beta(t)$ can be made. This does, of course, require a method of estimating the susceptible time series, as direct empirical data in this area requires an invasive serological survey and is rarely available, never in a sufficient temporal scale [18]. There exist several methods of estimating the susceptible time series [19, 20, 10]. The method used in this work was taken from Ref. [21] and is described in detail in the Methods chapter.

Solving Eq. 1.4 for β and iterating over empirical data for measles, varicella (chickenpox) and rubella give the time series found in Figure 1.2. Given that each of these diseases is most common amongst school-aged children [18, 10, 22] it is unsurprising that the period corresponding with the lowest transmission is in the summer, followed by a peak in September correlating with a return to school. These findings are further discussed and confirmed by Refs. [10, 23], though more analysis by Ref. [18] indicate the peak in transmission rate occurs several weeks earlier, alleging this effect may be better attributed to weather fluctuations. Regardless, there is undoubtedly a seasonal variance in the transmission rate for each of these diseases, one which occurs consistently each year. A common practice in disease modelling is to model the transmission rate functionally as sinusoidal [10, 24, 25, 26], of the form

$$\beta(t) = \beta_0(1 + \beta_1 \cos(2\pi t/T - \phi)), \tag{1.5}$$

where $T = 1$ year is the period of the oscillation and ϕ is the phase shift corresponding with the seasonal behaviour of the transmission rate. This method of seasonal forcing will be used for the remainder of this thesis, though other methods such as term-time forcing [11, 22, 12, 13] do exist.



(a) Incidence (measles)

(b) Transmission rate estimate (measles)

(c) Incidence (varicella)

(d) Transmission rate estimate (varicella)

(e) Incidence (rubella)

(f) Transmission rate estimate (rubella)

Figure 1.2: Reconstructed time-varying transmission rate $\beta(t)$ for three infectious diseases. Subpanels show weekly case notifications for (a) measles, (c) varicella, and (e) rubella and (b, d, f) their corresponding reconstructed $\beta(t)$. Red line in (b,d,f) shows mean value of reconstruction, and shaded areas show +/- one standard deviation.

1.1.4 Demographics

The basic SIR model (Eqs. 1.1-1.3) emits solutions that do not exhibit nontrivial long-term behaviour. The phase-plane analysis (shown in Section 1.1.6) shows that each realization of the system will converge to a steady-state, which does not mimic the endemic nature of many real-world infectious diseases. Part of this issue lies in the fact that the basic SIR model has no mechanism to recruit members to the susceptible class (often we call this "birth"), so the susceptible population becomes depleted over time. To remedy this, birth and death rates (ν and μ, respectively) are introduced to the model, resulting in the differential equations

$$
\begin{aligned}
S'(t) &= \nu - \beta(t)S(t)I(t) - \mu S(t) && (1.6)\\
I'(t) &= \beta(t)S(t)I(t) - \gamma I(t) - \mu I(t) && (1.7)\\
R'(t) &= \gamma I(t) - \mu R(t) && (1.8)
\end{aligned}
$$

Simulating this model results in much more complex dynamics than with the basic SIR model. Oscillations in each state variable can occur within a biologically relevant parameter space, where the system can exhibit annual, biennial, or multiennial attractors, each consistent with observed dynamics of real-world infections. In Figure 1.3, the above model was simulated using the time-varying transmission rate found in Eq. 1.5 to produce an annual attractor.

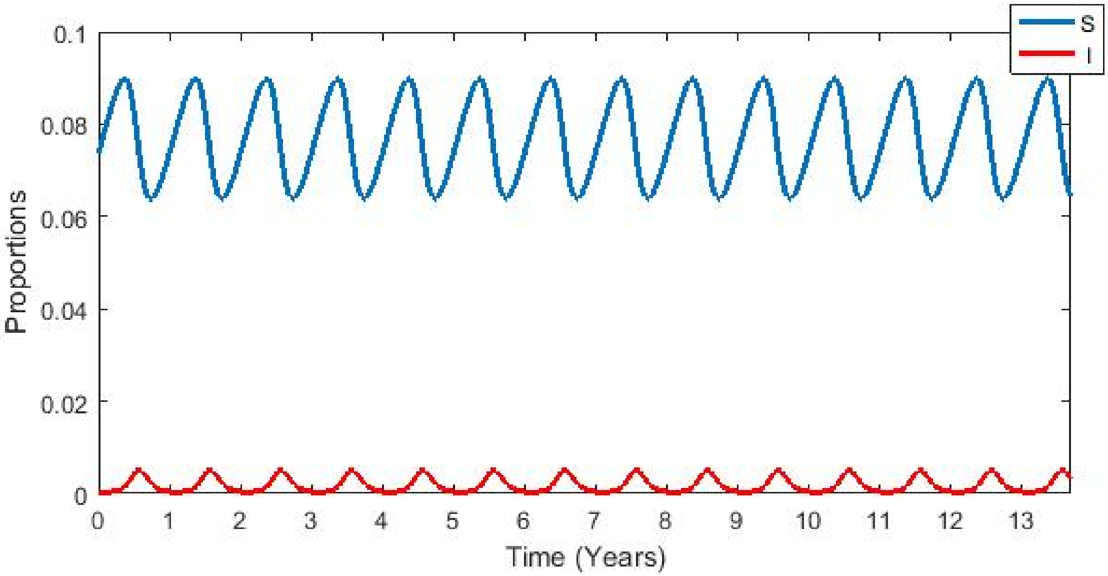


Figure 1.3: A simulation of the SIR model with demographics and seasonal forcing, with initial conditions $S_0 = 0.1, I_0 = 0.001, R_0 = 0.899$ and parameters $\beta_0 = 0.12$ day^{-1}, $\beta_1 = 0.08, \gamma = 0.1$ day^{-1}, and$\mu = \nu = 0.0002$ day^{-1}

1.1.5 Discrete Time Model

The SIR model presented in Eqs. 1.1 - 1.3 is a continuous time model, represented by differential equations and continuous state variables. However, given the discrete nature of the empirical data (case and birth data is often given in a weekly or biweekly format), it can be advantageous to approximate the SIR model as a discrete time model using difference equations:

$$S_{t+1} = S_t + \nu - \beta S_t I_t - \mu S_t \tag{1.9}$$

$$I_{t+1} = I_t + \beta S_t I_t - \gamma I_t - \mu I_t \tag{1.10}$$

$$R_{t+1} = R_t + \gamma I_t - \mu R_t \tag{1.11}$$

In the limit as $\Delta t \to 0$ the discrete model (Eqs. 1.9 - 1.11) converges to the continuous model (Eqs. 1.1 - 1.3). In our case, use of this approximation is advantageous as we require empirical data for the response of the system, which in the continuous case is the derivative vector $\boldsymbol{\dot{x}(t)} = \langle \dot{S}(t), \dot{I}(t), \dot{R}(t) \rangle$. As this requires numerical differentiation of a potentially noisy system, valuable information can be lost. However, when using the discrete system, the response vector is $\boldsymbol{x_{t+1}} = \langle S_{t+1}, I_{t+1}, R_{t+1} \rangle$, which is simply the next data point and thus is implicitly available without numerical approximations.

1.1.6 Stability Analysis

In order to understand the behaviour of any dynamical system it is important to analyse the existence and stability of equilibrium points of the system. An equilibrium point of an autonomous system $\boldsymbol{f}(\boldsymbol{x}(t))$ is any point $\boldsymbol{x^*}$ in the phase space such that $\boldsymbol{f}(\boldsymbol{x}^*) = 0$. There exists two main classifications defining behaviour of a system around an equilibrium point: *local* asymptotic stability and *global* asymptotic stability. To define these concepts, suppose $\boldsymbol{x}^*$ is an equilibrium point of the system $\boldsymbol{f}(\boldsymbol{x}(t))$. Then

- the system is **globally asymptotically stable** if, for every trajectory $\boldsymbol{x}(t)$ present in the phase space, $\boldsymbol{x}(t) \to \boldsymbol{x}^*$ as $t \to \infty$.
- the system is **locally asymptotically stable** near or at $\boldsymbol{x}^*$ if there exists an $R > 0$ s.t. $||\boldsymbol{x}(0) - \boldsymbol{x}^*|| \leq R \Rightarrow \boldsymbol{x}(t) \to \boldsymbol{x}^*$ as $t \to \infty$.

Below we find the equilibrium points of the SIR model with demographics (Eqs 1.6 - 1.8) and show that, when each exists, it is globally asymptotically stable.

Equilibrium Points

First we note that given the assumption of a fixed population size (i.e. $\nu = \mu$) and using a system scaled by population size we have $R = 1 - S - I$, so the dimensionality of the system may be reduced to the 2-dimensional system

$$S'(t) = \mu - \beta S(t)I(t) - \mu S(t) \tag{1.12}$$

$$I'(t) = \beta S(t)I(t) - \gamma I(t) - \mu I(t) \tag{1.13}$$

To simplify the analysis we will also assume the parameters β, μ, and γ are temporally invariant. We first find the equilibria of the system by setting the derivatives vector $[S'(t), I'(t)]$ identically to zero and solving the system:

$$\begin{aligned} 0 &= \mu - \beta S(t)I(t) - \mu S(t) \\ 0 &= \beta S(t)I(t) - \gamma I(t) - \mu I(t) \end{aligned}$$

which gives the two steady-states

$$\begin{aligned} \mathcal{E}_1 = (\bar{S}_1, \bar{I}_1) &= (1, 0) \\ \mathcal{E}_2 = (\bar{S}_2, \bar{I}_2) &= \left(\frac{\mu + \gamma}{\beta}, \frac{\mu(\beta - \mu - \gamma)}{\beta(\mu + \gamma)} \right). \end{aligned}$$

Note that for the equilibrium $(\bar{S}_2, \bar{I}_2)$ to be biologically relevant each of the state variables must be positive and ≤ 1. Thus in order for the steady-state to exist it is required that $\dfrac{\mu + \gamma}{\beta} \leq 1$. This gives rise to the threshold parameter $\mathcal{R}_0$, defined as

$$\mathcal{R}_0 = \frac{\beta}{\mu + \gamma}. \tag{1.14}$$

This parameter is called the **basic reproductive ratio** and is a crucial feature of any endemic infection. Biologically, it can be interpreted as the number of new cases produced by a single member of the infectious class in an otherwise susceptible population. Given this definition, it is logical to conclude that a disease will become endemic if and only if $\mathcal{R}_0 > 1$.

Classification of Equilibria

In order to classify the stability of each equilibrium point, we utilize the Lyapunov Theorem for Global Asymptotic Stability [27]:

Theorem 1.1.1. *Let $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{f}(\boldsymbol{x}^*) = 0$ for some $\boldsymbol{x}^* \in \Sigma \subset \mathbb{R}^n$. If there exists a $\mathcal{C}^1$ function $V : \mathbb{R}^n \to \mathbb{R}$ such that*

1. $V(\boldsymbol{x}^*) = 0$

2. $V(\boldsymbol{x}) > 0 \ \forall \ x \neq \boldsymbol{x}^*$

3. $\dot{V}(\boldsymbol{x}) < 0 \ \forall \ x \neq \boldsymbol{x}^*$

4. $V(\boldsymbol{x}) \to \infty$ *as* $||\boldsymbol{x}|| \to \infty$

then $\boldsymbol{x}^*$ *is a globally asymptotically stable equilibrium.*

Note that the time derivative of the Lyapunov function can be found using the Chain Rule:

$$\dot{V}(\boldsymbol{x}) = \frac{\partial V}{\partial \boldsymbol{x}} \dot{\boldsymbol{x}} = \sum_i \frac{\partial V_i}{\partial x_i} f_i(\boldsymbol{x}) \tag{1.15}$$

Case I: $\mathcal{R}_0 \leq 1$

In this case, the endemic equilibrium $\mathcal{E}_2 = (\bar{S}_2, \bar{I}_2) = \left(\frac{1}{\mathcal{R}_0}, \frac{\mu}{\mu + \gamma} \left[1 - \frac{1}{\mathcal{R}_0} \right] \right)$ is not biologically relevant, as $\bar{S}_2 = \frac{1}{\mathcal{R}_0} > 1$. Thus the only equilibrium to consider is the trivial $\mathcal{E}_1 = (\bar{S}_1, \bar{I}_1) = (1, 0)$.

The analysis done in Ref. [28] discovers the Lyapunov function

$$L(S, I) = \bar{S}_1 \left(\frac{S}{\bar{S}_1} - \ln \frac{S}{\bar{S}_1} \right) + I - 1 = S - \ln(S) + I - 1$$

which gives the time derivative

$$\dot{L}(S, I) = \frac{\mu}{S}(1 - S)^2 - (1 - \mathcal{R}_0) I \leq 0 \text{ when } \mathcal{R}_0 \leq 1 \text{ and } S, I \geq 0.$$

By inspection we also see that $L(1, 0) = 0$, $L(S, I) > 0 \ \forall \ (S, I) \neq (1, 0)$ and $L(S, I) \to \infty$ as $S, I \to \infty$. Thus by Theorem 1.1.1 the trivial equilibrium $\mathcal{E}_1$ is globally asymptotically stable.

Case II: $\mathcal{R}_0 > 1$

When the basic reproductive ratio exceeds 1 the disease becomes endemic and cycles of epidemics are allowed to occur. The endemic equilibrium point $\mathcal{E}_2$ is now biologically relevant, so we consider its stability. Ref. [28] presents a different Lyapunov function for this equilibrium, given by

$$V(S,I) = \bar{S}_2\left(\frac{S}{\bar{S}_2} - \ln\frac{S}{\bar{S}_2}\right) + \bar{I}_2\left(\frac{I}{\bar{I}_2} - \ln\frac{I}{\bar{I}_2}\right) - (\bar{S}_2 + \bar{I}_2)$$

which gives the time derivative

$$\dot{V}(S,I) = -\mu\frac{\bar{S}_2}{S}\left(1 - \frac{S}{\bar{S}_2}\right)^2 \leq 0 \text{ when } \mathcal{R}_0 > 1 \text{ and } S, I \geq 0.$$

By inspection we see that $V(\bar{S}_2, \bar{I}_2) = 0$ and that $V(S,I) \to \infty$ as $S, I \to \infty$. To note that $(\bar{S}_2, \bar{I}_2)$ is the global minimum of V it is sufficient to compute the partials

$$\frac{\partial V}{\partial S} = 1 - \frac{\bar{S}_2}{S}, \qquad \frac{\partial V}{\partial I} = 1 - \frac{\bar{I}_2}{I}.$$

Thus by Theorem 1.1.1 the endemic equilibrium $\mathcal{E}_2$ is globally asymptotically stable. As a result of this, it can be concluded that the trivial equilibrium $\mathcal{E}_1$ is unstable.

Figure 1.4 shows an example of this stability by plotting the direction field of the system (using $\mathcal{R}_0 = 8$ to ensure an endemic equilibrium point) as well as a specific trajectory of the model with $(S_0, I_0) = (0.12, 0.02)$.

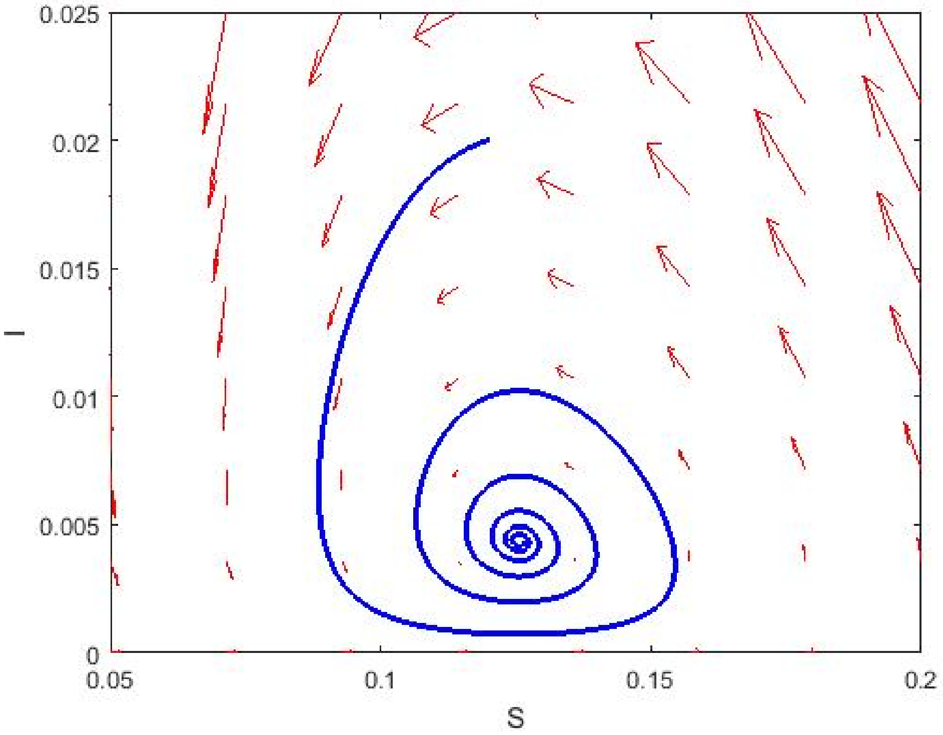


Figure 1.4: Direction field of a subset of the phase space for the system found in Eqs. 1.6 - 1.8 with parameters $\gamma = 0.01, \beta = 0.08, \mu = 0.00005$. Also plotted is a trajectory of the system with initial conditions $(S_0, I_0) = (0.12, 0.02)$. This demonstrates that the endemic equilibrium $\mathcal{E}_2 = \left(\frac{\mu+\gamma}{\beta}, \frac{\mu(\beta-\mu-\gamma)}{\beta(\mu+\gamma)}\right)$ is stable.

1.2 Data-driven Modelling and SINDy

The previous section outlines the evolution of creating epidemiological models based on theoretical assumptions and validated using empirical data, which historically has been the fundamental method of mathematically describing disease dynamics. In the next section we introduce a different approach: the derivation of dynamic models directly from prevalence data.

1.2.1 Origins of Data-Driven Dynamical Modelling

Complex nonlinear dynamics lie at the heart of many natural systems in science and engineering [29, 30, 31], including epidemiology [25, 26, 12]. Centuries of mathematical research have been devoted to creating models that accurately describe and predict the behaviour of these systems [32], usually in the form of deductive models from mechanics derived from pre-existing theory. In recent years, with advances in machine learning [33] and the increased availability and understanding of data [34, 35, 36] strides have been made in automating the model discovery process, creating inductive models. Rigorous techniques such as regression methods are currently in place to understand static data [37], but analogous advancements given dynamic data have not been developed as quickly.

One of the first and most influential attempts made to motivate dynamic models using empirical data was made by Edward Lorenz with his 1963 paper entitled "Deterministic Nonperiodic Flow" [38]. Through this famous work he developed theory that would lay the foundations for modern weather prediction, deriving nonlinear statistical modelling techniques from atmospheric data [39]. This led to a much better understanding of the chaotic dynamical systems which are often present in nature, including in epidemiology [17, 25, 12]. Other early attempts at reconstructing nonlinear dynamics that stem from chaos theory involved the methodology of delay-coordinate systems [40, 41]. This method, though successful at reconstructing features of the system such as dimensionality, Lyapunov exponents, and unstable periodic orbits [42], could not be used to recover a functional symbolic form that could be analysed using traditional phase-plane methods.

A breakthrough in modelling nonlinear dynamics functionally came in Ref. [43] and supplemented by Ref. [44] by applying symbolic regression (genetic programming [45]) to recover differential equations. This work was the first successful attempt at automating the process of finding the symbolic structure of the dynamical system governing a natural process. Being able to model a system symbolically rather than numerically is crucial due to the explanatory value of a model built with elementary functions. It is the goal of these symbolic modelling techniques to automatically uncover the nonlinearities active in the governing equations of a system, a process which traditionally is difficult for human intuition. However, these early attempts utilizing genetic programming were subject to overfitting, as well as being computationally expensive and lacking the ability to scale well to systems with higher dimensionality.

1.2.2 Sparsity and Regularization

The idea of achieving a high level of accuracy when creating a mathematical model is not quite as simple as it may seem. Traditional methods of obtaining accuracy are usually centred around the minimizing the residual squared error between the predicted response and the data. The most common and simplest method is known as ordinary least squares, or OLS. However, this is not always the most satisfactory result, for two reasons [46]. Firstly, simply minimizing the residuals results in a high level of variance, leading to inefficient prediction value from the resulting model. Despite high descriptive value, these models tend to overfit to any noise present in the data and lack the ability to identify the true predictors that drive the system in question. The idea of cross-validation [37] is useful in identifying instances where overfitting takes place, but does not in and of itself remedy the problem. Secondly, in an era where data is plentiful, creating a model that uses all possible predictors detracts from the interpretability of the model. These complex models may perform well but are beyond human analytic ability and are not useful in expanding theoretical knowledge. This temptation to create overly complicated models led to the reflection in Ref. [47] that "the best material model of a cat is another, or preferably the same, cat".

Most alternative methods to OLS that promote the ideas of predictability and interpretability fall under one of two classes [37]. The first is subset selection, which attempts to identify some subset of the predictors that adequately describes the system, disregarding the rest [48, 49]. The second is shrinkage (or regularization), which fits the model using all of the available predictors but forces the coefficients of select predictors towards zero, effectively performing an approximate form of variable selection. One such method is known as ridge regression which, rather than minimizing the residual squared error, instead minimizes the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{1.16}$$

where y_i are the response data, x_{ij} are the predictor data, β_i are the adjustable linear coefficients, and λ is the tuning parameter that is determined beforehand to adjust the level of shrinkage. Note that this method is simply the residual sum of squares with a second term subtracted, known as the shrinkage (l_2 regularized) penalty. While ridge regression is effective in obtaining shrinkage of coefficients that remedies overfitting, it will always include all predictors in the model and therefore is ineffective in variable selection and improving interpretive value. This leads to a slight alternative to this method, known as the LASSO (least absolute shrinkage and selection operator). This method is instead an l_1 regularized regression and minimizes the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{1.17}$$

The effect of an l_1 penalty is that now some coefficients will be forced to zero. Thus applying the LASSO can accomplish both subset selection and shrinkage, resulting in sparse models that have both predictive and interpretive value. This does not, of course, come without cost, as this method can tend to be computationally expensive when applied to large datasets [50].

1.2.3 Model Selection

There exists another group of rigorous statistical metrics that are used to balance goodness-of-fit with model complexity, called *information criteria*. These metrics are useful in the comparison and selection of models when first given a space of candidate models from which to choose. In the context of symbolic modelling this space is usually constructed from a functional basis, often heuristically defined given contextual theory [51, 52, 53]. Given a computationally tractable basis, each possible model would be fitted and the information criterion would be computed and used to select the model that best balances parsimony and predictive power. The information criterion used in this thesis is called the Akaike information criterion (AIC) [54] and is derived from use of maximum likelihood. The AIC value for a given candidate model i is defined by

$$AIC_i = 2k - 2\ln(L(\mathbf{x}, \hat{\mu}), \tag{1.18}$$

where L is the conditional probability of the observations $\mathbf{x}$ given the set of best-fit model parameters $\hat{\mu}$, and k is the number of free parameters in the model. During application, if sample size is a concern then (as noted by Ref. [55]) a correction should be applied of the form

$$AIC_c = AIC_i + \frac{2(k+1)(k+2)}{m-k-2}, \tag{1.19}$$

where m is the number of observations in the sample.

1.2.4 Sparse Identification of Nonlinear Dynamics

The previous sections have laid the groundwork for the main topic of this thesis: applications of the SINDy (Sparse Identification of Nonlinear Dynamics) algorithm to epidemiological data. This algorithm was developed by Brunton, Proctor, and Kutz in their 2016 paper entitled “Discovering governing equations from data by sparse identification of nonlinear dynamical systems” [50]. The methods outlined in this paper approach the problem of automating the discovery of dynamic equations that describe natural systems through the lens of sparsity-promoting regression techniques. Through the understanding that the governing equations of many of these systems are expected to be sparse in the space of all possible functions, they demonstrate it to be feasible to create an algorithm that automates the discovery of these active terms. The derivation of this algorithm as well as its application to epidemiological data is discussed in Chapter 2. Further work on this algorithm, including extensions and applications, can be found in Refs. [55, 56, 57, 58, 59, 60, 61, 62]. Of particular note is Ref. [55], which combines the notion of sparse regression with model selection to further promote parsimonious models with predictive and interpretive ability. This work also applies the SINDy algorithm to simulated data from a discrete compartmental disease model, successfully recovering the active terms in the model.

1.3 Objectives and Rationale

The central motivation for this thesis is to expand upon the efforts to automate the model discovery process and to create models simply from empirical data with minimal knowledge of the system. As discussed in Section 1.2.4, recent work by Ref. [50] has demonstrated that their SINDy algorithm is effective in recovering the governing equations of dynamical systems given simulated realizations of the system. Accurate model rediscovery from a simulated model deductively derived lends valuable insight towards determining whether discovering an inductive model of the system from data is feasible. However, to our knowledge this work and subsequent research using this algorithm has not successfully discovered the governing equations given empirical data of a natural system. My objective in this thesis, then, is to apply the SINDy algorithm to empirical disease data to discover functional forms for the nonlinear dynamics that govern epidemics, and to demonstrate the potential that data-driven techniques have to either confirm current epidemiological modelling practices or to enhance them with new insight.

The motivation to understand these epidemiological systems better are twofold. The obvious incentive stems from a public health importance, as accurate disease models are beneficial in advising government policy on vaccination strategy and predicting the occurrence and effect of infection outbreak [13, 63, 64, 65]. In the modern era of vaccination, these models are most often applied to the allocation of resources [66] to both temper the spread of severe epidemics and to completely eradicate infections that mass vaccination has drastically reduced. Through the development of epidemiological modelling techniques, mathematicians seek to understand the cause and spread of infections diseases, leading to the most efficient ways to control and eradicate them.

The second incentive is a more mathematical one, motivated by a desire to study complex dynamical systems. In the pre-vaccination era, epidemics exhibited a range of regular and irregular dynamics, depending on type of disease and geographical location. For example, measles prevalence could exhibit annual, biennial, or multiennial cycles, leading to new research into what mechanisms facilitated the sustainment of and transition between these oscillations [11, 13]. These epidemiological systems also can exhibit chaotic behaviour [12, 25] depending on the levels of biological realism (e.g. age structure, vital dynamics, etc.) and stochasticity.

The complex nature of dynamic disease systems, coupled with the lack of availability and quality of pre-vaccination data and inherent noise present, makes the idea of discovering a revolutionary new symbolic mathematical model that describes epidemics in a parsimonious and interpretable way a lofty goal. It is important to note, then, that the intention of this thesis is not to create an automatically-discovered model that immediately sets a standard for modelling practices of infectious diseases. We simply wish to present this method as a proof-of-concept, demonstrating the feasibility of data-driven models that have the potential to be both predictive and interpretive, reconciling these models with current theory.

To this end, this thesis will be structured as follows: The SINDy algorithm is described in Chapter 2, as well as the nuances of applying it to epidemiological data. The results are detailed in Chapter 3, demonstrating the ability of the algorithm to recover compartmental disease models from simulated data, as well as discover parsimonious models given epidemiological data from a range of disease types, each exhibiting different dynamic behaviour. Finally, Chapter 4 includes a discussion of the methods and results presented, as well as outlining limitations and future work.

Chapter 2

Methods

2.1 Sparse Identification of Nonlinear Dynamics (SINDy)

This work builds on the sparse regression methods outlined in Ref. [50]. Given the recent advances in both compressed sensing [67, 68, 69] and sparse regression [37, 46] it has become computationally feasible to extract system dynamics from large, multimodal datasets. These techniques rely heavily on the fact that many dynamical systems can be represented by governing equations that are sparse in the space of all possible functions. In this work we focus on dynamical systems that are given by a system of ordinary differential equations of the form

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}(t), t), \tag{2.1}$$

where $\boldsymbol{x}(t) = (x_1(t), x_2(t), ..., x_n(t))$ represents the state of the n-dimensional system at time t, and $\boldsymbol{f} = (f_1, f_2, ..., f_n)$ is the sparse set of functions that dictate the dynamics of the system.

It is assumed that the time series data is sampled at points $t_1, t_2, \ldots, t_m$ for both $\boldsymbol{x}$ and $\dot{\boldsymbol{x}}$, usually given as either data from simulations or empirical data from measurements. Depending on the system in question, numerical differentiation methods to approximate $\dot{\boldsymbol{x}}$ that are well-suited for the level of noise must be used. The method used in Ref. [50] is total variation regularization [70, 71] that works well on a noisy system when only the state variables are available. Alternatively, a discrete adaptation of SINDy may be used, where the response of the system $\boldsymbol{f}(\boldsymbol{x}_t, t)$ is x_{t+1}. Regardless, the time series data of the state variables and the response are represented by the matrices

$$\boldsymbol{X} = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \dots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \dots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \dots & x_n(t_m) \end{bmatrix}$$

$$\dot{\boldsymbol{X}} = \begin{bmatrix} \dot{x_1}(t_1) & \dot{x_2}(t_1) & \dots & \dot{x_n}(t_1) \\ \dot{x_1}(t_2) & \dot{x_2}(t_2) & \dots & \dot{x_n}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x_1}(t_m) & \dot{x_2}(t_m) & \dots & \dot{x_n}(t_m) \end{bmatrix}.$$

We then construct a library of linear and nonlinear candidate functions for the model, given prior knowledge of the system we wish to describe. Common choices for these functions are polynomial and trigonometric functions of the state variables, though other functions (e.g. exponential, rational) functions may be included as well. This function library is then evaluated at each time-step, generating the $m \times p$ matrix

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} 1 & \boldsymbol{X} & \boldsymbol{X}^{P_2} & \boldsymbol{X}^{P_3} & \dots & \sin(\boldsymbol{X}) & \cos(\boldsymbol{X}) & \sin(\boldsymbol{2X}) & \cos(\boldsymbol{2X}) & \dots \end{bmatrix}, \tag{2.2}$$

where $\boldsymbol{X}^{P_n}$ represents all possible polynomials of degree n that can be constructed by the state variables. Now, relying on the assumption that the derivative $\dot{\boldsymbol{X}}$ can be described by relatively few of the nonlinearities active in $\Theta(\boldsymbol{X})$, we may set up the sparse regression problem

$$\dot{\boldsymbol{X}} = \Theta(\boldsymbol{X})\Xi, \tag{2.3}$$

where $\Xi = (\xi_1, \xi_2, \dots, \xi_p)$ is a set of sparse coefficient vectors.

There are several current methods that have been developed to perform sparse regression. A common choice is the LASSO (least absolute and shrinkage operator) [37, 46], a regression method that promotes sparsity by applying an l_1 penalty on the norm of the coefficient vector.

However, this method does not scale well to large datasets. This thesis utilizes an iterative method developed by Brunton et. al., as described below:

1. Perform a least-squares regression on the relation in Eq. [2.3].
2. Set all terms in Ξ that are less (in absolute value) than some threshold λ to zero.
3. Create new library Θ', dropping functions that correspond to zero entries in Ξ.
4. Repeat steps 1-3 until equilibrium (i.e. no terms in Ξ are smaller in magnitude than λ), or some other stopping criteria is reached.

This yields the set of sparse vectors that provides an approximate solution to Eq. [2.3]. We can then reconstruct the kth row of the dynamical system by taking

$$\dot{\boldsymbol{x}}_k = \Theta(\boldsymbol{x}_k^T)\xi_k, \tag{2.4}$$

where $\Theta(\boldsymbol{x}_k^T)$ is the symbolic representations of the elements of $\boldsymbol{x}$.

Finally, combining all of the rows of the discovered dynamical system results in the system of equations

$$\dot{\boldsymbol{x}} = \Xi^T\Theta(\boldsymbol{x}^T)^T. \tag{2.5}$$

The code for this algorithm, along with several examples that demonstrate its application, can be found at Ref. [72]. The modified repository used for all computation done for this theis can be found at Ref. [73].

2.2 Applying SINDy to Epidemiological Systems

The application of data-driven model discovery methods to epidemiological systems presents a unique set of challenges. Firstly, incidence data is often subjected to noise at several levels, notably inconsistent reporting of disease cases from medical clinics [18, 74, 75]. In addition, the derivative data must be approximated using numerical methods, leading to another source of inaccuracy. Secondly, as presented in Chapter 1, most compartmental disease models depend on both the infected and the susceptible classes. However, temporal data of the seropositive individuals in a population would require extensive and invasive surveying and is not currently available for any demographic. Instead, several methods for approximating the susceptible class from the given incidence data are outlined in Section 2.2.2.

2.2.1 Data Preprocessing

Temporal data of disease incidence of various infections and time periods has been made available by numerous sources, often from governmental reporting programs. The three infectious diseases and the corresponding locations and time periods used for this study are measles in England and Wales from 1948-1967 (from Ref. [76]), varicella (chicken pox) in Ontario (Canada) from 1946-1967, and rubella in Ontario from 1946-1960 (both from [13]). These diseases and time periods were chosen as they exhibit contrasting dynamic behaviour, most notably in the period of the epidemic cycle. It is important to note that for each of these diseases, the time frame chosen is before the vaccines for the respective diseases became commonly available. Once this data was imported and both the time and case vectors were labelled, both the birth and population data (taken from [76, 77, 78, 79, 80]) were imported and interpolated to be given per week, the same scale as the disease data.

As outlined in Chapter 1, the compartmental models that motivate our development of these methods have two main classes: the susceptible population and the infectious population. The latter is referred to as the *prevalence* of the disease, defined by the number (or proportion) of infectious individuals at any given time. However, the data are most often given in the form of newly occurring cases, referred to as *incidence* data. Hence both the susceptible and the prevalence data must be recovered from the incidence data before the SINDy algorithm can be applied. The subsequent sections illustrate several methods for recovering each of these two time series.

2.2.2 Susceptible Reconstruction

The problem of estimating the susceptible class given prevalence data is an issue that has not been conclusively addressed, and at present there is no convention on what method provides the best approximation. We apply several of the current methods to the sources of data listed above and compare the resulting time series to known qualities of the dynamics of susceptible classes, from both epidemiological theory and compartmental models.

Perhaps the simplest method for the reconstruction of the susceptible class is to iterate the equation

$$S_{t+1} = S_t - \alpha C_{t,t+1} + B_{t,t+1}, \tag{2.6}$$

where S_t represents the number of susceptibles at the start of week t, $C_{t,t+1}$ and $B_{t,t+1}$ are the number of new cases and births respectively in week t, and α is the rate at which cases are reported (i.e. α^{-1} is the average proportion of all cases that are reported to the data collection agency) [18]. The idea behind this method is simple: each week the suceptible class grows by the number of new births into the population (in the absence of vaccination), and shrinks by the number of new infections. If the reporting rate α was well known, this relation would provide a good approximation. However, reporting varies significantly for different diseases and locations [75] as well as changing temporally [21]. It is also difficult to estimate explicitly, due to the lack of serological data available.

An extension of this method is derived in Ref. [21]. They assume the discrete relation

$$S_{t+1} = S_t - \alpha_t C_{t,t+1} + B_{t-d,t-d+1} + u_t, \tag{2.7}$$

where u describes the additive noise ($E(u) = 0, V(u) = \sigma_u^2$), and d represents a short delay to allow for the period of time between birth and susceptibility to the disease.

Now let Z_t describe the deviation from the mean $E(S) = \bar{S}$ at week t, i.e.

$$S_t = \bar{S} + Z_t. \tag{2.8}$$

By substituting Eq. [2.8] into Eq. [2.7] we see that Z_t also satisfies the relation

$$Z_{t+1} = Z_t - \alpha_t C_{t,t+1} + B_{t-d,t-d+1} + u. \tag{2.9}$$

Iterating this expression results in the relation

$$Z_t = Z_0 - \sum_{i=1}^{t} \alpha_i C_{i,i+1} + \sum_{i=1}^{t} B_{i-d,i-d+1} + \sum_{i=1}^{t} u_i \tag{2.10}$$

Finkenstadt and Grenfell in Ref. [21] use the simplifying notation

$$X_t = \sum_{i=1}^{t} C_{i,i+1}, \qquad Y_t = \sum_{i=1}^{t} B_{i-d,i-d+1}, \qquad U_t = \sum_{i=1}^{t} u_i, \qquad R_t = \sum_{i=1}^{t} (\alpha_i - \bar{\alpha}) C_i.$$

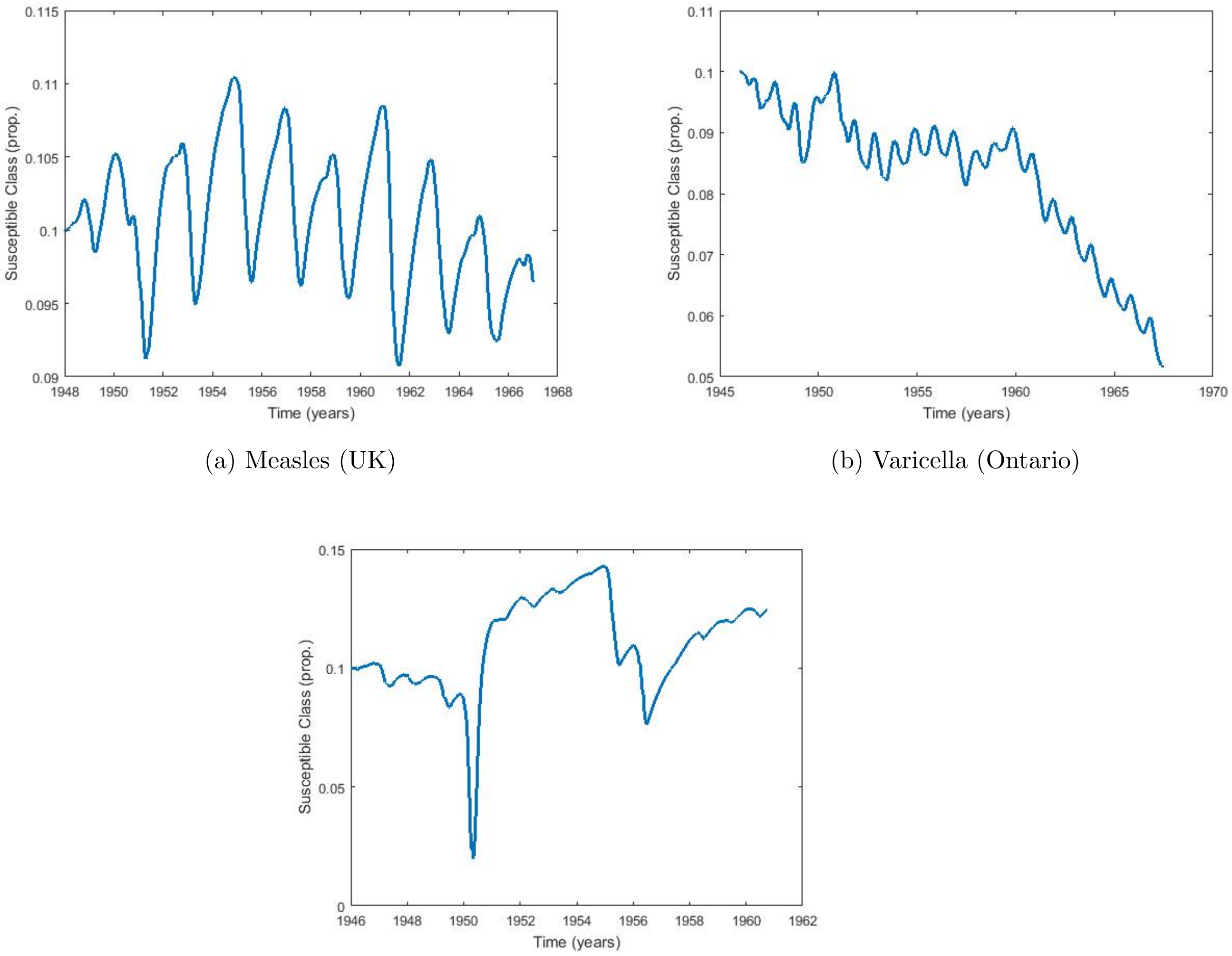
This simplifies Eq. [2.10] to

$$Z_t = Z_0 - \bar{\alpha} X_t + Y_t - R_t + U_t. \tag{2.11}$$

If it is assumed that the reporting rate is constant ($R_t \approx 0$) and noise is negligible ($U_t \approx 0$), this reduces to the linear relationship

$$Y_t = \bar{\alpha} X_t + (Z_t - Z_0). \tag{2.12}$$

Hence, applying a linear regression to the cumulative births (Y_t) against the cumulative cases (X_t) provides an estimate for the residuals $Z_t - Z_0$ and the average reporting rate $\bar{\alpha}$. Applying this reconstruction method yields susceptible classes for each of the datasets in Section 2.2.1, as shown in Figure 2.1.



(a) Measles (UK)

(b) Varicella (Ontario)

(c) Rubella (Ontario)

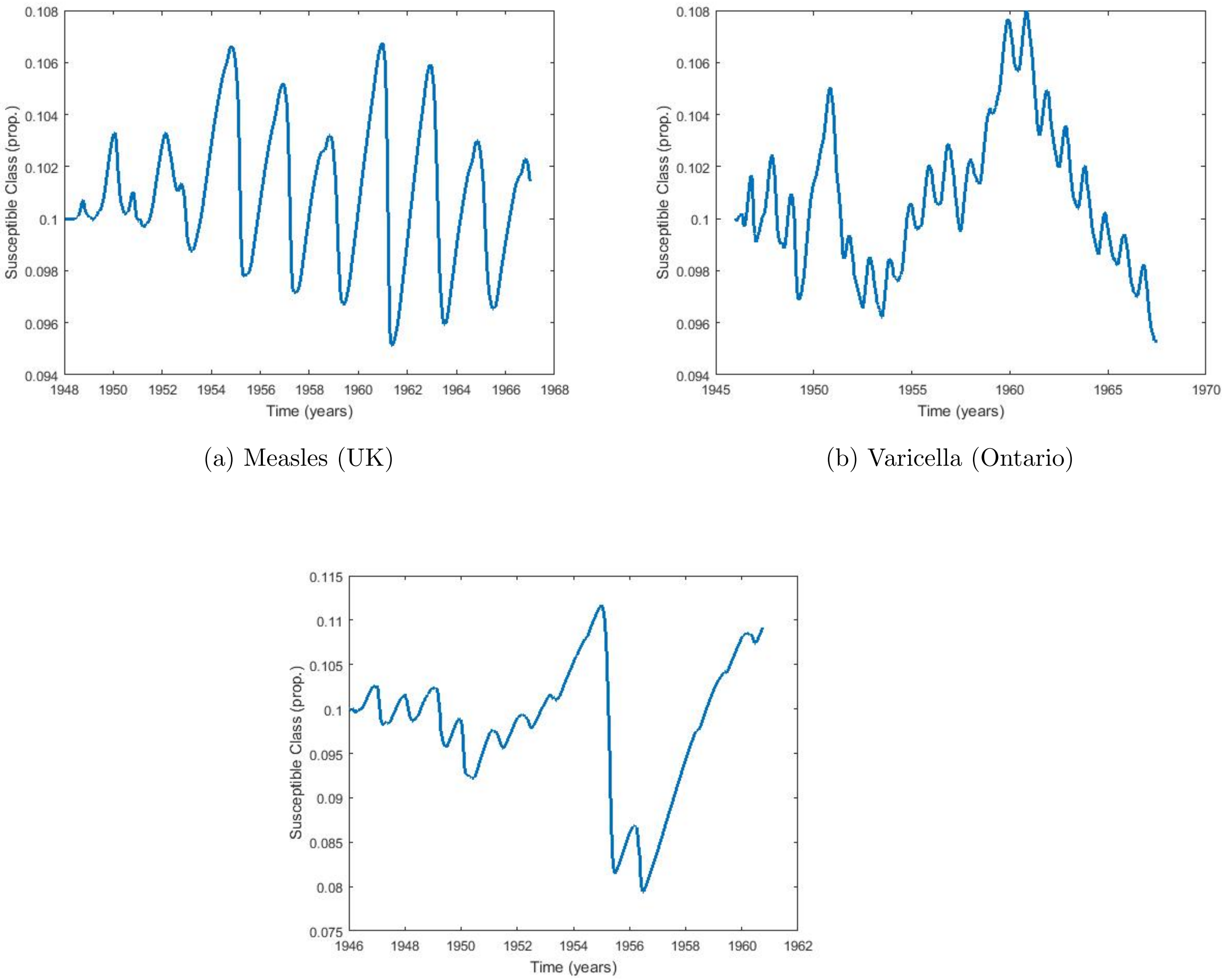
Figure 2.1: Suceptible reconstructions for measles (a), varicella (b), and rubella (c) using the global regression method.

From these figures it can be seen that each reconstruction (especially from the varicella and rubella case notification data) suffers from local shifts in the mean, caused by the assumption that the reporting rate is temporally invariant. Finkenstadt and Grenfell account for this by supposing the dominant fluctuations in Eq. 2.11 are caused by variation in the reporting rate α_t rather than in external noise (u_t). Eq. 2.11 can then be expressed as

$$Y_{t+1} = R_t - U_t Z_0 - (\alpha_{t+1} - \bar{\alpha}) X_t + \alpha_{t+1} X_{t+1} + Z_{t+1} - u_{t+1}. \tag{2.13}$$

Local linear regression techniques can then be applied to estimate both the reporting rate and the susceptible class. The method used in this work is the same as in Ref. [21] which is sensitive to the bandwith parameter, and must be tuned beforehand to minimize large-scale fluctuations from the global mean.

Utilizing this locally linear regression method, the susceptible time series for the measles, varicella and rubella data used previously in Figure 2.1 are now



(a) Measles (UK)

(b) Varicella (Ontario)

(c) Rubella (Ontario)

Figure 2.2: Suceptible reconstructions for measles (a), varicella (b), and rubella (c) using the locally linear regression method.

2.2.3 Incidence to Prevalence Conversion

As noted in Section 2.2.1, prevalence data is required to create models analogous to typical compartmental modes, but empirical data is usually presented in incidence form. Hence, before performing any model extraction, we must first convert the given incidence data into prevalence data.

Given temporal case data C_t, suppose that the duration of infection (D_i), the mean individual lifespan (L), and the proportion of people that will contract the disease in their lifetime (p) are known and constant. The average proportion of the population that is infected at any given time is then given by

$$\langle P_t \rangle = \frac{pD_i}{L} \tag{2.14}$$

From the relation

$$\frac{P_t}{\langle P_t \rangle} = \frac{C_t}{\langle C_t \rangle}$$

we then obtain

$$P_t = \frac{C_t p D_i}{\langle C_t \rangle L} \tag{2.15}$$

which is used to construct the prevalence (infectious) class given incidence data.

2.2.4 Weighted Thresholding

The engine that drives the SINDy model discovery algorithm is sparse regression, a statistical learning technique that performs feature selection while fitting the active terms to the data. The realization of this technique used in Ref. [50] is the iterated thresholding method, outlined in Section 2.1. The key parameter in this algorithm is λ, a chosen threshold below which coefficients (and their corresponding functions) are eliminated on any given iteration. In Ref. [50] and subsequent papers this parameter is taken as constant, though Ref. [55] analyses the effects of fitting λ using cross-validation. However, epidemiological data present an additional challenge, as the state variables are often orders of magnitude apart (see Figures 3.17a - 3.19b for examples of this). When evaluating a higher order function library using data on contrasting scales, high order functions of small state variables (such as I^3) have a much smaller column norm than larger state variables or functions with a smaller polynomial order. As a result, the iterated sparse regression algorithm can assign them large coefficients to account for this, which are much less likely to be eliminated by a fixed thresholding value.

To account for this, we introduce a threshold for each function in the library that is scaled according to the norm of the corresponding column. For each column k in the function library $\Theta(\boldsymbol{X})$, we construct the threshold

$$\lambda_w^{(k)} = \frac{\lambda_c}{|\Theta^{(k)}(\boldsymbol{X})|}, \tag{2.16}$$

where $\Theta^{(k)}(\boldsymbol{X})$ is the kth column in the function library, $|\cdot|$ is the l_2- norm, and λ_c is a constant threshold value. The algorithm in Section 2.1 is then performed in the same way, using this function-dependent sparsity knob instead. This is the technique utilized in the rest of this thesis, and any reference to a constant λ value is the λ_c parameter in Eq. 2.16.

2.2.5 Choice of a Functional Basis

Determining the correct basis of elementary functions is a key step when generating a model using SINDy, and the lack of a rigorous method to identify such a basis is one of its notable downfalls [50]. Nevertheless, most compartmental models in epidemiology have been constructed using a simple basis of polynomial and trigonometric functions, which is what we use in this analysis. Many compartmental disease models only use polynomial functions on the second degree or lower, so we commonly limit our function library to second or third order polynomials.

Depending on the nature of the system and the assumptions made, it becomes necessary to add several features to the function library. As discussed in Section 1.1.3, the dynamics of the prevalence of both measles and varicella are strongly dictated by a seasonal component. Hence, a new parameter β is constructed such that

$$\beta = \beta_0(1 + \beta_1 \cos(2\pi t/T - \phi)), \tag{2.17}$$

where T is the period of the seasonal oscillations (usually $1yr^{-1}$) and ϕ is the phase shift. This parameter is then multiplied by each of the p columns in Θ to create p new features in the function library.

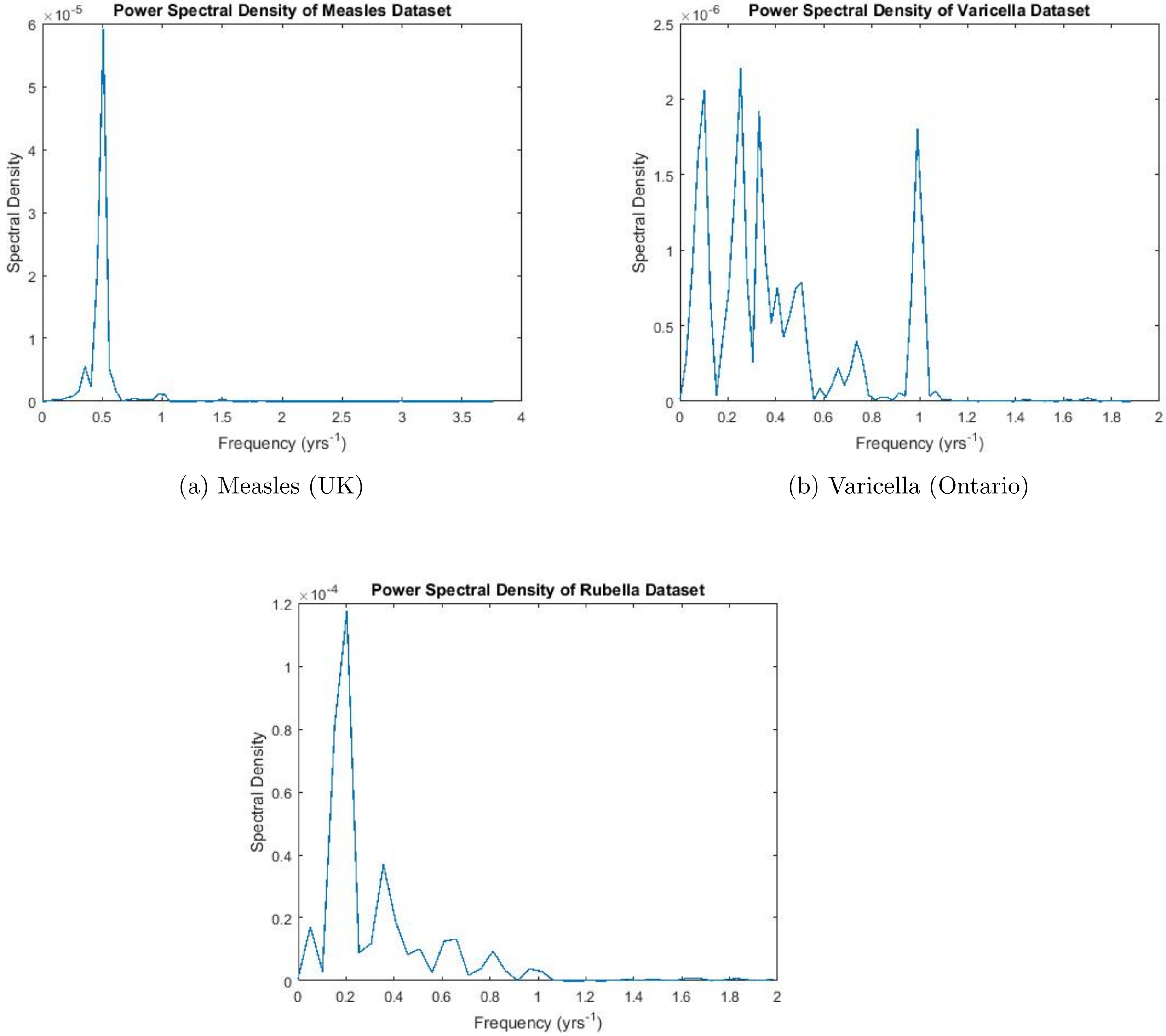
Given that the susceptible population is influenced heavily by the birth rate, the addition of a birth parameter is also beneficial. A functional form of the birth rate can be assumed and added to the library, but given that in the place and time period of this study the birth rate does not behave in a way that can be described by a linear or exponential function we choose to represent the birth rate in the function library by simply including a column of the empirical data $B(t)$ that gives the total number of births in week t. This data is already required to scale the state variables and the source for each location used in this thesis is given in Section 2.2.1.

2.2.6 Power Spectral Density

Perhaps the most characteristic feature of a pre-vaccination infectious disease is the frequency at which epidemics occur. The diseases studied in this thesis exhibit three distinct patterns: an annual cycle (varicella, Fig. 1.2(c)), a biennial cycle (measles, Fig. 1.2(a)) and a multi-annual cycle (rubella, Fig (1.2(e)). When using automated model discover from data it is important to ensure the model captures the underlying attractor and exhibits the same temporal pattern. However, a number of factors such as initial conditions or the phase of the seasonal forcing can give model instances that captures these patterns but are out-of-sync with the data, resulting in a poor evaluation by a model selection metric such as AIC or residual error.

Another method of comparing time series that captures the frequency of temporal oscillations is the power spectral density (PSD), which shows the relative strength of the various frequencies present within the data. Comparing the PSDs generated from time series of different models can identify which models have similar cycles regardless of whether the peaks are synchronized. Given situations where the overarching dynamic behaviour of the model is seen as more important the accuracy of a specific realization, this technique can be advantageous in the model selection process.

In order to compute the PSD of a given infectious time series the data was first trend-corrected, tapered with a split cosine bell and smoothed using a moving average [13, 81]. The PSD was then computed using the MATLAB function `periodogram` [82]. The results from each disease given in Section 2.2.1 are given in Figure 2.3.



(a) Measles (UK)

(b) Varicella (Ontario)

(c) Rubella (Ontario)

Figure 2.3: Power spectral density estimates for the prevalence time series of the measles (a), varicella (b), and rubella (c) datasets. These estimates show the underlying attractor(s) present within the data and can be used for qualitative time series analysis and comparison.

Chapter 3

Results

The main results from this thesis can be separated into two distinct sections: recovering model equations given simulated data from a known (SIR) system, and discovering model equations given empirical disease data. The former has been the focus of most research conducted using the SINDy algorithm [50, 55, 56, 59, 60, 61] and is motivated in part to developing understanding of the method with the intention of applying the algorithm to observable systems. With this motivation in mind, we apply the SINDy algorithm to continuous and discrete SIR models, both the basic model and then adding seasonal forcing and demographics. This then lends insight to the second section, which examines the models discovered by the algorithm when applied to empirical data from the three datasets mentioned in Section 2.2.1.

3.1 Model Rediscovery from Simulated Data

This work explores modelling disease outbreak using both the continuous and discrete temporal regimes. Both have benefits and drawbacks depending on the system in question and the problem the model attempts to solve. Additionally, when working in the context of using SINDy as a model discovery technique, the need to estimate the response of the dynamical system is an important consideration. In order to use a continuous time-scale when applying SINDy it is necessary to evaluate the derivative of the input data, which can yield noisy and unpredictable values when using empirical data. Hence, when using simulated data to emulate the model discovery process, it is beneficial to consider both the continuous and discrete time-scales to lend insight into what techniques might translate well when applying

SINDy to empirical data of disease dynamics.

As discussed in Chapter 1, the standard model of disease dynamics is the SIR compartmental model. In its most basic form, realizations of this system model single outbreaks, though with the additions of vital dynamics (birth and death rates) and seasonal forcing, more complex dynamics that describe long-term patterns of endemic infectious diseases can be achieved. Thus we seek to rediscover various forms of the SIR model given simulated data, to lend insight to how the SINDy algorithm behaves when presented with epidemiological time series.

3.1.1 Continuous Regime

The SIR Model

The model equations for the continuous SIR model are

$$\begin{aligned} S'(t) &= \nu - \beta S(t)I(t) - \mu S(t) \quad (3.1) \\ I'(t) &= \beta S(t)I(t) - \gamma I(t) - \mu I(t) \quad (3.2) \\ R'(t) &= \gamma I(t) - \mu R(t) \quad (3.3) \end{aligned}$$

where S, I, and R are state variables describing the proportion of the population that are susceptible, infected and recovered respectively, β is the transmission rate, γ^{-1} is the mean duration of infection, ν is the mean birth rate and μ is the mean death rate.

Firstly, it is relevant to note that R is a redundant state variable, meaning that the remaining state variables do not depend on it explicitly. This creates unnecessary complications in the model selection process, as noted in Ref. [55]. Therefore the equation for $R'(t)$ is omitted when applying SINDy to the simulated data.

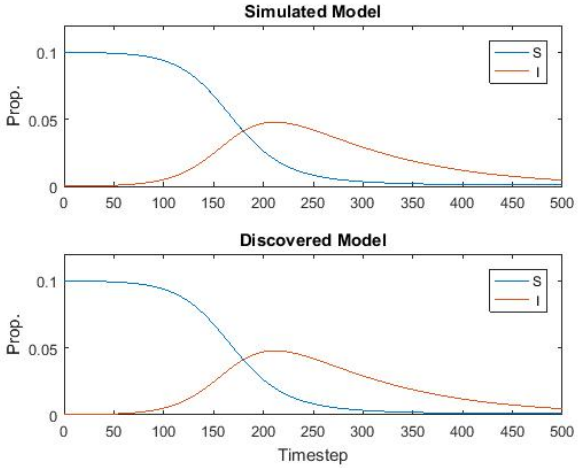
The model described by Eqs. 3.1 - 3.3 was simulated using the parameter values and initial conditions found in Tables 3.1 and 3.2, corresponding to simple and more complicated versions of the model (Cases I and II, respectively). The data for the derivative vector $x(t) = [S(t), I(t)]$ was also found through model simulation, rather than from numerical differentiation methods. Additive noise at various levels was introduced after simulation in each of the state variables. The function library was then compiled (taking polynomials up to the second order) and the SINDy sparse regression algorithm was run.

Case I: Constant transmission rate, no demographics

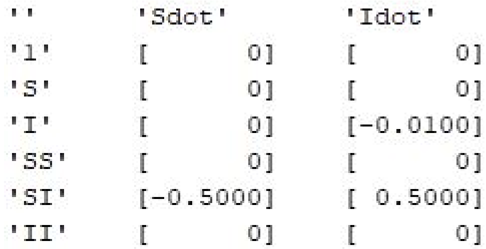
The transmission rate in the simulated model was assumed to be time-invariant (i.e. β is constant) and the birth and death rate were taken to be zero. The corresponding function library taken was

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} \mathbf{1} & \boldsymbol{S} & \boldsymbol{I} & \boldsymbol{S}^2 & \boldsymbol{I}^2 & \boldsymbol{SI} \end{bmatrix}$$

In Figures 3.1 and 3.2, the simulated model and the model discovered by SINDy are compared at varying levels of additive noise.



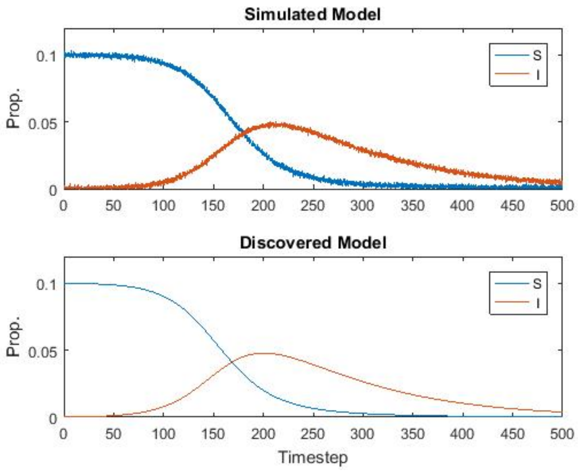
(a) Time-series



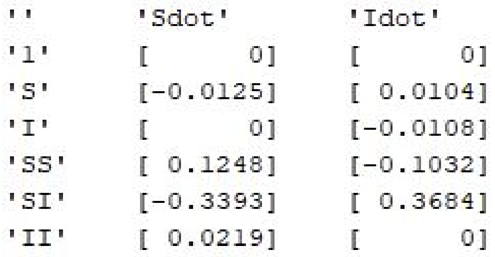
```
''        'Sdot'        'Idot'
'1'       [      0]     [      0]
'S'       [      0]     [      0]
'I'       [      0]     [-0.0100]
'SS'      [      0]     [      0]
'SI'      [-0.5000]     [ 0.5000]
'II'      [      0]     [      0]
```

(b) Model Coefficients

Figure 3.1: Comparison of the simulated SI model with no vital dynamics, time-invariant transmission rate, and additive noise of $\epsilon = 0.00001$ with the corresponding discovered model. In this example with low relative noise, SINDy successfully identifies the correct active terms of the system, as well as the magnitude of the corresponding coefficients. Parameters used are found in Table 3.1.



(a) Time-series



```
''      'Sdot'         'Idot'
'1'     [       0]     [       0]
'S'     [-0.0125]      [ 0.0104]
'I'     [       0]     [-0.0108]
'SS'    [ 0.1248]      [-0.1032]
'SI'    [-0.3393]      [ 0.3684]
'II'    [ 0.0219]      [       0]
```

(b) Model Coefficients

Figure 3.2: Comparison of the simulated SI model with no vital dynamics, time-invariant transmission rate, and additive noise of $\epsilon = 0.001$ with the corresponding discovered model. The correct terms are still present in the discovered model, but the coefficients are no longer accurate. Other terms have also been selected in an attempt to overfit the model to the noisy data. Parameters used are found in Table 3.1.

$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$	$\boldsymbol{\nu}$
0.5	0.01	0	0

Table 3.1: Model coefficients for the SIR model (Eqs. 3.1 - 3.3) displaced in Figures 3.1 and 3.2. Units for all parameters are wk^{-1}.

In this relatively simple model, SINDy was able to correctly identify the active terms and parameter values ($\beta = 0.5, \gamma = 0.01$) given a sufficiently low level of additive noise. As the noise level increases, terms that are not active in the simulated model are given nonzero coefficients, in an attempt to locally overfit the resulting model to the noise. However, the resulting SINDy model still exhibits the same noticeable behaviour as the noisy simulated model.

Case II: Seasonal transmission rate, constant demographics

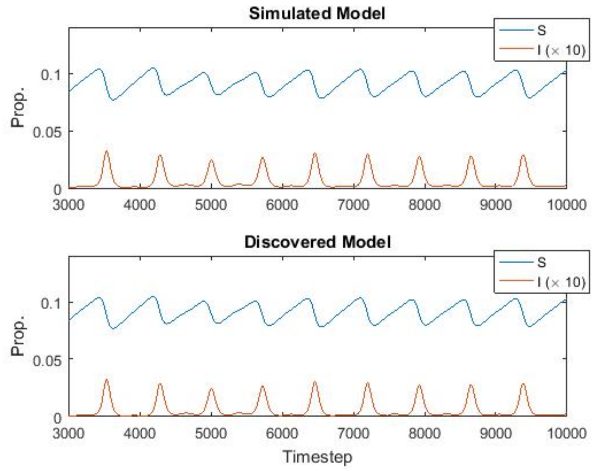
In this case a seasonally-varying transmission rate was introduced, of the form:

$$\beta = \beta_0(1 + \beta_1 \cos(2\pi t/T - \phi)), \tag{3.4}$$

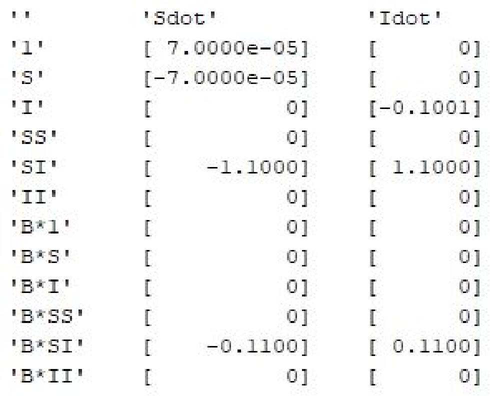
where $T = 1yr^{-1}$ is the period of the oscillations, and ϕ is an arbitrary phase shift corresponding with the yearly peak of the force of infection. When the model coefficients are given in figures describing SINDy-discovered models, this parameter is represented by B. In addition, a constant and equal (but nonzero) birth and death rate were included (parameters ν and μ in Eqs. 3.1 - 3.3). These additions give a function library of

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} \mathbf{1} & \boldsymbol{S} & \boldsymbol{I} & \boldsymbol{S^2} & \boldsymbol{I^2} & \boldsymbol{SI} & \boldsymbol{\beta} & \boldsymbol{\beta S} & \boldsymbol{\beta I} & \boldsymbol{\beta S^2} & \boldsymbol{\beta I^2} & \boldsymbol{\beta SI} \end{bmatrix}$$

In Figures 3.3 and 3.4, the simulated model and the model discovered by SINDy are compared at varying levels of additive noise.



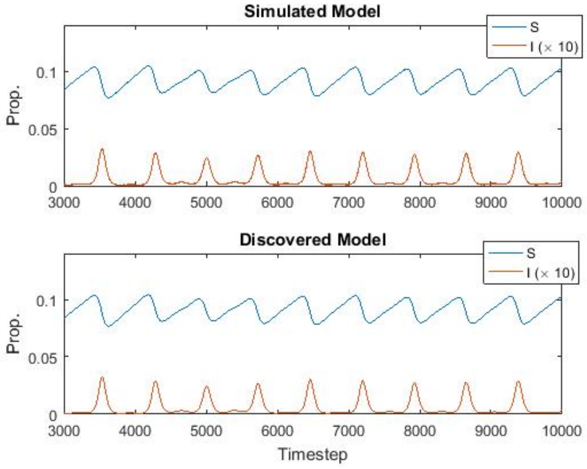
(a) Time-series



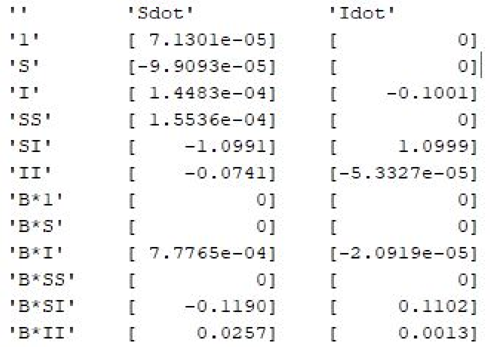
```
''          'Sdot'              'Idot'
'1'         [ 7.0000e-05]       [       0]
'S'         [-7.0000e-05]       [       0]
'I'         [           0]      [-0.1001]
'SS'        [           0]      [       0]
'SI'        [     -1.1000]      [ 1.1000]
'II'        [           0]      [       0]
'B*1'       [           0]      [       0]
'B*S'       [           0]      [       0]
'B*I'       [           0]      [       0]
'B*SS'      [           0]      [       0]
'B*SI'      [     -0.1100]      [ 0.1100]
'B*II'      [           0]      [       0]
```

(b) Model Coefficients

Figure 3.3: Comparison of the simulated SI model with vital dynamics, seasonal forcing, and additive noise of $\epsilon = 1 \times 10^{-8}$ with the corresponding discovered model. In this example with low relative noise, SINDy successfully identifies the correct active terms of the system, as well as the magnitude of the corresponding coefficients. Note that the infectious time series in both plots have been scaled by a factor of 10 to improve readability. Parameters used are found in Table 3.2.



(a) Time-series



```
''         'Sdot'           'Idot'
'1'        [ 7.1301e-05]    [          0]
'S'        [-9.9093e-05]    [          0]
'I'        [ 1.4483e-04]    [    -0.1001]
'SS'       [ 1.5536e-04]    [          0]
'SI'       [    -1.0991]    [     1.0999]
'II'       [    -0.0741]    [-5.3327e-05]
'B*1'      [          0]    [          0]
'B*S'      [          0]    [          0]
'B*I'      [ 7.7765e-04]    [-2.0919e-05]
'B*SS'     [          0]    [          0]
'B*SI'     [    -0.1190]    [     0.1102]
'B*II'     [     0.0257]    [     0.0013]
```

(b) Model Coefficients

Figure 3.4: Comparison of the simulated SI model with vital dynamics, seasonal forcing, and additive noise of $\epsilon = 0.00001$ with the corresponding discovered model. The correct terms are still present in the discovered model, but the coefficients are no longer accurate. Other terms have also been selected in an attempt to overfit the model to the noisy data. Note that the infectious time series in both plots have been scaled by a factor of 10 to improve readability. Parameters used are found in Table 3.2.

$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$	$\boldsymbol{\nu}$
1.1	0.1	0.1	0.00007	0.00007

Table 3.2: Model coefficients for the SIR model simulated above. Units for all parameters are wk^{-1}.

The inclusion of both a seasonally-varying transmission rate and nonzero birth and death rates yields a model that more accurately represents observed epidemiological systems, namely the existence of seasonal oscillations in both the susceptible and infectious classes. Even with these more complex dynamics and expanded functional library, SINDy was able to accurately identify the active terms and parameter values, though once again these results were sensitive to the level of additive noise.

3.1.2 Discrete Regime

The SIR Model

Analogous to the continuous case, the equations for the discrete SIR model are

$$
\begin{aligned}
S_{t+1} &= S_t + \nu - \beta S_t I_t - \mu S_t \qquad (3.5)\\
I_{t+1} &= I_t + \beta S_t I_t - \gamma I_t - \mu I_t \qquad (3.6)\\
R_{t+1} &= R_t + \gamma I_t - \mu R_t \qquad (3.7)
\end{aligned}
$$

where t is now a discrete independent variable and the parameters have the same meaning as in the previous section. As in the continuous case, we may treat the state variable R as redundant and exclude it from simulations.

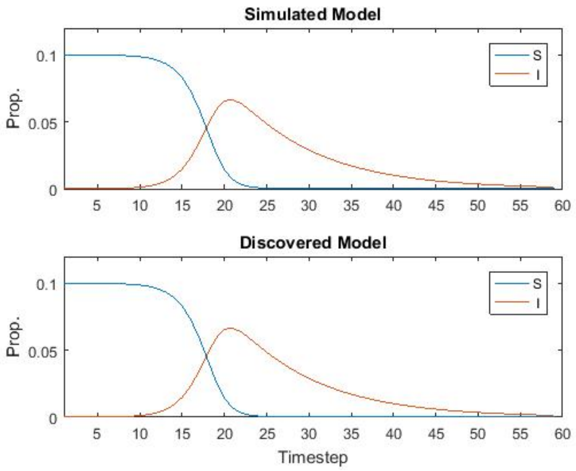
Discrete simulations were again run using both the simple and more complicated versions of the model given by Eqs. 3.5 - 3.7 (taking the parameter values and initial conditions found in Tables 3.3 and 3.4). The response vector used in the sparse regression is now $x_{t+1} = [S_{t+1}, I_{t+1}]$ and is also found directly through model simulation. Additive noise at various levels was introduced after simulation in each of the state variables. The function library was then compiled (taking polynomials up to the second order) and the SINDy sparse regression algorithm was run.

Case I: Constant transmission rate, no demographics

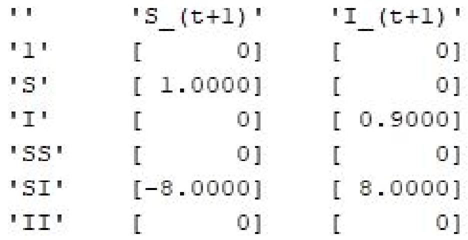
The transmission rate in the simulated model was assumed to be constant (i.e. β is time-invariant) and the birth and death rate were taken to be zero. The corresponding function library taken was

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} \mathbf{1} & \boldsymbol{S} & \boldsymbol{I} & \boldsymbol{S}^2 & \boldsymbol{I}^2 & \boldsymbol{SI} \end{bmatrix}$$

In Figures 3.5 and 3.6, the simulated model and the model discovered by SINDy are compared at varying levels of additive noise.



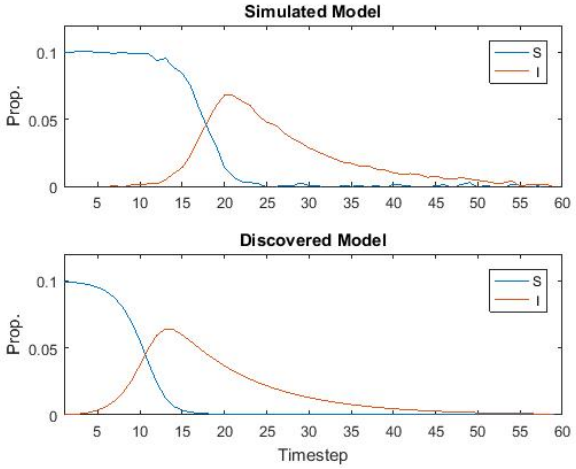
(a) Time-series



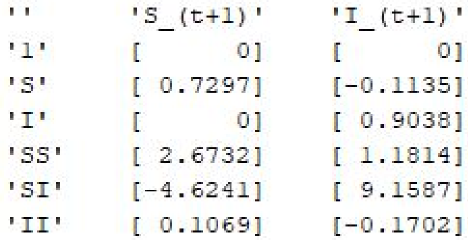
```
''       'S_(t+1)'    'I_(t+1)'
'1'      [      0]    [      0]
'S'      [ 1.0000]    [      0]
'I'      [      0]    [ 0.9000]
'SS'     [      0]    [      0]
'SI'     [-8.0000]    [ 8.0000]
'II'     [      0]    [      0]
```

(b) Model Coefficients

Figure 3.5: Comparison of the simulated SI model with no vital dynamics, time-invariant transmission rate, and additive noise of $\epsilon = 0.0000001$ with the corresponding discovered model. In this example with low relative noise, SINDy successfully identifies the correct active terms of the system, as well as the magnitude of the corresponding coefficients. Parameters used are found in Table 3.3.



(a) Time-series



```
''       'S_(t+1)'    'I_(t+1)'
'1'      [      0]    [      0]
'S'      [ 0.7297]    [-0.1135]
'I'      [      0]    [ 0.9038]
'SS'     [ 2.6732]    [ 1.1814]
'SI'     [-4.6241]    [ 9.1587]
'II'     [ 0.1069]    [-0.1702]
```

(b) Model Coefficients

Figure 3.6: Comparison of the simulated SI model with no vital dynamics, time-invariant transmission rate, and additive noise of $\epsilon = 0.0001$ with the corresponding discovered model. The correct terms are still present in the discovered model, but the coefficients are no longer accurate. Other terms have also been selected in an attempt to overfit the model to the noisy data. Parameters used are found in Table 3.3.

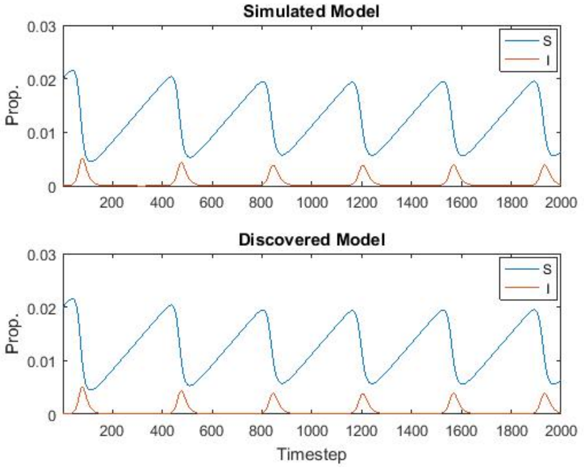
$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$	$\boldsymbol{\nu}$
8.0	0.1	0	0

Table 3.3: Model coefficients for the SIR model simulated above. Units for all parameters are wk^{-1}.

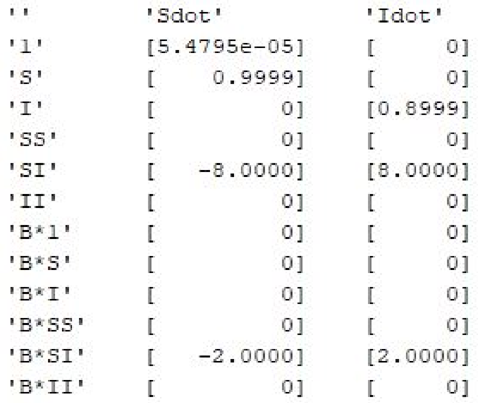
Again, when using the simplest form of the SIR model, SINDy can successfully recover the model coefficients given a library of first and second order polynomials. Increasing the noise again leads to the inclusion of active nonlinearities that are not active in the model, in an attempt to fit to the noise, but the effects of these terms appear negligible in the resulting time series.

Case II: Seasonal transmission rate, constant demographics

The same functional form for the transmission rate and function library as in the continuous case were used to introduce seasonality to the simulated model. In Figures 3.3 and 3.4, the simulated model and the model discovered by SINDy are compared at varying levels of additive noise.



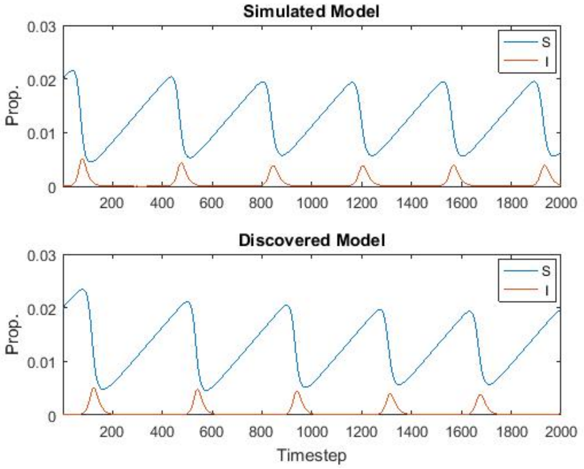
(a) Time-series



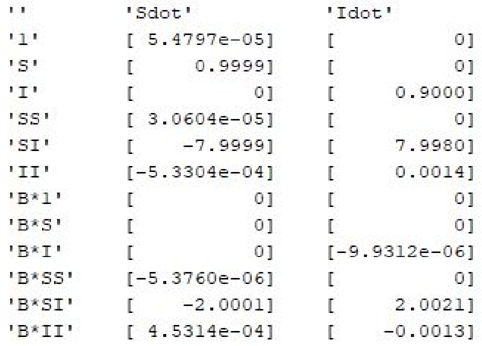
```
''          'Sdot'             'Idot'
'1'         [5.4795e-05]       [       0]
'S'         [     0.9999]      [       0]
'I'         [          0]      [0.8999]
'SS'        [          0]      [       0]
'SI'        [    -8.0000]      [8.0000]
'II'        [          0]      [       0]
'B*1'       [          0]      [       0]
'B*S'       [          0]      [       0]
'B*I'       [          0]      [       0]
'B*SS'      [          0]      [       0]
'B*SI'      [    -2.0000]      [2.0000]
'B*II'      [          0]      [       0]
```

(b) Model Coefficients

Figure 3.7: Comparison of the simulated SI model with vital dynamics, seasonal forcing, and additive noise of $\epsilon = 1 \times 10^{-10}$ with the corresponding discovered model. In this example with low relative noise, SINDy successfully identifies the correct active terms of the system, as well as the magnitude of the corresponding coefficients. Parameters used are found in Table 3.4.



(a) Time-series



```
''          'Sdot'              'Idot'
'1'         [ 5.4797e-05]       [           0]
'S'         [     0.9999]       [           0]
'I'         [          0]       [      0.9000]
'SS'        [ 3.0604e-05]       [           0]
'SI'        [    -7.9999]       [      7.9980]
'II'        [-5.3304e-04]       [      0.0014]
'B*1'       [          0]       [           0]
'B*S'       [          0]       [           0]
'B*I'       [          0]       [-9.9312e-06]
'B*SS'      [-5.3760e-06]       [           0]
'B*SI'      [    -2.0001]       [      2.0021]
'B*II'      [ 4.5314e-04]       [     -0.0013]
```

(b) Model Coefficients

Figure 3.8: Comparison of the simulated SI model with vital dynamics, seasonal forcing, and additive noise of $\epsilon = 1 \times 10^{-7}$ with the corresponding discovered model. The correct terms are still present in the discovered model, but the coefficients are no longer accurate. Other terms have also been selected in an attempt to overfit the model to the noisy data. Parameters used are found in Table 3.4.

$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$	$\boldsymbol{\nu}$
8	0.25	0.1	5.4795×10^{-5}	5.4795×10^{-5}

Table 3.4: Model coefficients for the SIR model simulated above. Units for all parameters are wk^{-1}.

Even with the inclusion of a seasonal transmission rate and demographics, SINDy was still able to correctly recover the coefficients of the simulated model. It should be noted, however, that the noise level required for this accurate identification is significantly lower than in Case I, indicating that the inclusion of more complex dynamics impacts the tolerance of the algorithm to random variation.

The results in this section demonstrate the ability of the SINDy algorithm to correctly identify the dynamics of models commonly used to analyse and make predictions about real-world disease systems. This suggests that given empirical data with similar dynamics, SINDy may be able to recover analogous models that can either confirm or expand upon current theory. The subsequent section describes our attempt to apply this algorithm to case notification data from several infectious diseases and details the resulting models.

3.2 Model Discovery from Empirical Data

Previously we have discussed the ability of SINDy to rediscover dynamic disease models given simulated realizations of these systems. The motivation behind this was to lend insight into the next topic of discussion, which is the discovery of dynamical systems that describe empirical data of disease prevalence. This is, of course, a much more interesting and influential area of research; the ability to either confirm the compartmental models currently used or discover new models that describe these systems in more accurate detail could be of great use to the field of mathematical epidemiology.

3.2.1 Overview of Features

Given the complex nature of the epidemiological systems and the structure of current modelling practices there were a number of crucial choices that were found to impact the accuracy of the discovered model. Some of these factors related directly to features of the epidemiological data, while others apply more generally to statistical modelling methods. Below we list each of these factors along with a brief discussion of the impact that the choice has on the identification algorithm.

- **Type of Disease:** There were three diseases and corresponding locations and time periods selected: measles in the UK from 1948-1967, varicella in Ontario (Canada), from 1946-1967, and rubella in Ontario from 1946-60. Each provides a different example of an attractor class: measles is biennial, varicella is annual, and rubella is multiennial.

- **Continuous vs. Discrete Time Scale:** As discussed in Chapter 1, disease dynamics can be modelled using both the continuous and discrete time models. Due to the necessity of numerically computing derivative data, operating in the continuous regime can lead to noisy response variables and make it difficult for the sparse regression algorithm to obtain a global minimum. Conversely, due to the weekly or bi-weekly structure of the empirical data, the discrete SINDy framework seems to lend itself naturally to this problem. Any systems following hereafter will use the discrete time scale.

- **Polynomial Order:** When selecting the function library, an important consideration is the largest order of polynomials that will be included. Selecting polynomials of at most 2nd order gives models most comparable to the current compartmental models, whereas selecting polynomials of at most 3rd order allows SINDy to capture extra features of the data, but sometimes at the cost of overfitting. Both function libraries will be considered and contrasted, though only the 2nd order libaries will be contrasted with fitted compartmental models.

- **Initial Susceptible Value (S_0):** As the nature of the susceptible class (and, by extension, the infectious class) depends heavily on its initial value this parameter plays an important role in the resulting dynamics. Notably, the susceptible population scales inversely with the basic reproductive ratio $\mathcal{R}_0$ [63], and as the initial value impacts the temporal average of the susceptible class it is important that S_0 is selected carefully to ensure accurate replication of dynamics. Additionally, as data on the susceptible class is not explicitly available, the initial value is not known for any of the analysed datasets. As a result, we most often make the choice to treat S_0 as a varying parameter and consider the resulting discovered model across an appropriate range.

- **Threshold Value (λ):** As discussed in the Methods chapter, at each sparse regression iteration of the SINDy algorithm this parameter acts as a cutoff value, dictating that the algorithm remove any functions corresponding to coefficients smaller than it. Usually a wide range of threshold values are selected, then either the optimal model will be chosen or each model across the range will be considered and contrasted.

Another important factor when considering the quality of the discovered model is the level of sparsity present within the coefficient matrix. To quantify this we introduce the concept of a *sparsity ratio*, the ratio of coefficients of value 0 to total possible functions, i.e.

$$r = 1 - \frac{||\Xi||_0}{||\Theta||_0},$$

where Ξ is the set of coefficient vectors and Θ is the collection of library functions.

3.2.2 Algorithm Demonstration

For each disease dataset (see Section 2.2.1), the adapted SINDy algorithm was applied to each point of a $S_0 - \lambda$ parameter grid, where S_0 is the initial susceptible proportion and λ is the cutoff value used in SINDy's thresholding algorithm. The sensitivity of the algorithm to these parameters is discussed further in Section 3.2.3. In addition, realizations of these models are also sensitive to the phase shift in the seasonal forcing (the parameter ϕ in Eq. 1.5). However, this parameter cannot be fit using the sparse regression algorithm as it is not a coefficient of a function in the library. To account for this, at each point of the parameter sweep a range was selected to uniformly sample the phase shift (using a step size of $\Delta\phi = 0.5$wk across 52 weeks) and the model with the best AIC score was selected to capture any shifts in dynamics caused by phase changes.

First we demonstrate the effectiveness of the SINDy algorithm by applying it to each of the three datasets, exploring the results using libraries including up to 2nd order polynomials and 3rd order polynomials. Figures 3.9-3.14 show the most parsimonious discovered models (the models with the lowest AIC score after the parameter sweep) plotted against the data, as well as the model coefficients. The AIC scores and further analysis of the effects of the parameters are presented in Section 3.2.3.

When using up to 2nd order polynomials, the function library used was

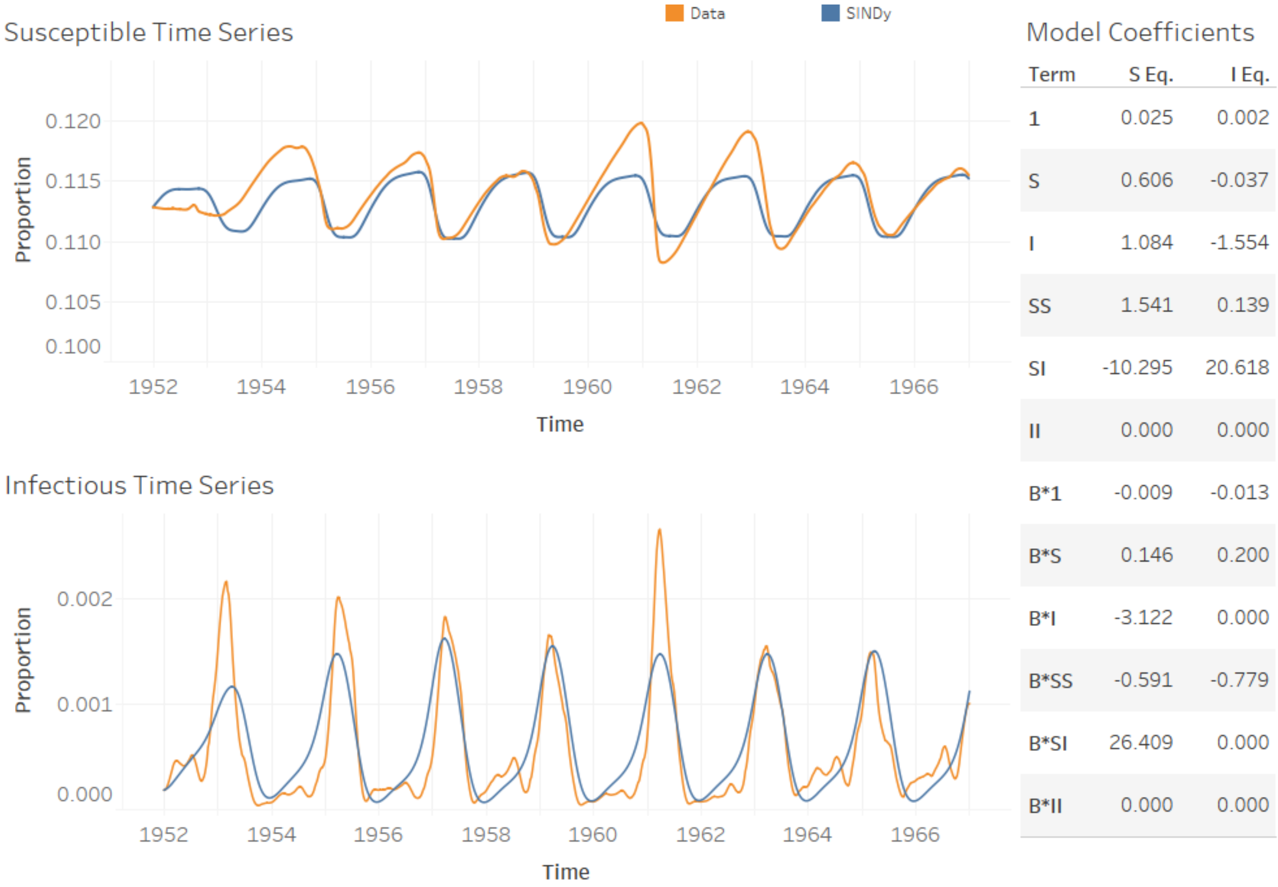
$$\Theta(\boldsymbol{X}) = \begin{bmatrix}\mathbf{1} & \boldsymbol{S} & \boldsymbol{I} & \boldsymbol{S^2} & \boldsymbol{I^2} & \boldsymbol{SI} & \boldsymbol{\beta} & \boldsymbol{\beta S} & \boldsymbol{\beta I} & \boldsymbol{\beta S^2} & \boldsymbol{\beta I^2} & \boldsymbol{\beta SI}\end{bmatrix},$$

and when using up to 3rd order polynomials, the function library used was

$$\begin{aligned}\Theta(\boldsymbol{X}) = [\mathbf{1} \quad & \boldsymbol{S} \quad \boldsymbol{I} \quad \boldsymbol{S^2} \quad \boldsymbol{I^2} \quad \boldsymbol{SI} \quad \boldsymbol{S^3} \quad \boldsymbol{S^2 I} \quad \boldsymbol{SI^2} \quad \boldsymbol{I^3} \\ & \boldsymbol{\beta} \quad \boldsymbol{\beta S} \quad \boldsymbol{\beta I} \quad \boldsymbol{\beta S^2} \quad \boldsymbol{\beta I^2} \quad \boldsymbol{\beta SI} \quad \boldsymbol{\beta S^3} \quad \boldsymbol{\beta S^2 I} \quad \boldsymbol{\beta SI^2} \quad \boldsymbol{\beta I^3}],\end{aligned}$$

where β is the seasonally-varying transmission rate given in Eq. 2.17. When the model coefficients are given in figures describing SINDy-discovered models, this parameter is represented by B.

Measles (UK), 2nd Order Polynomials

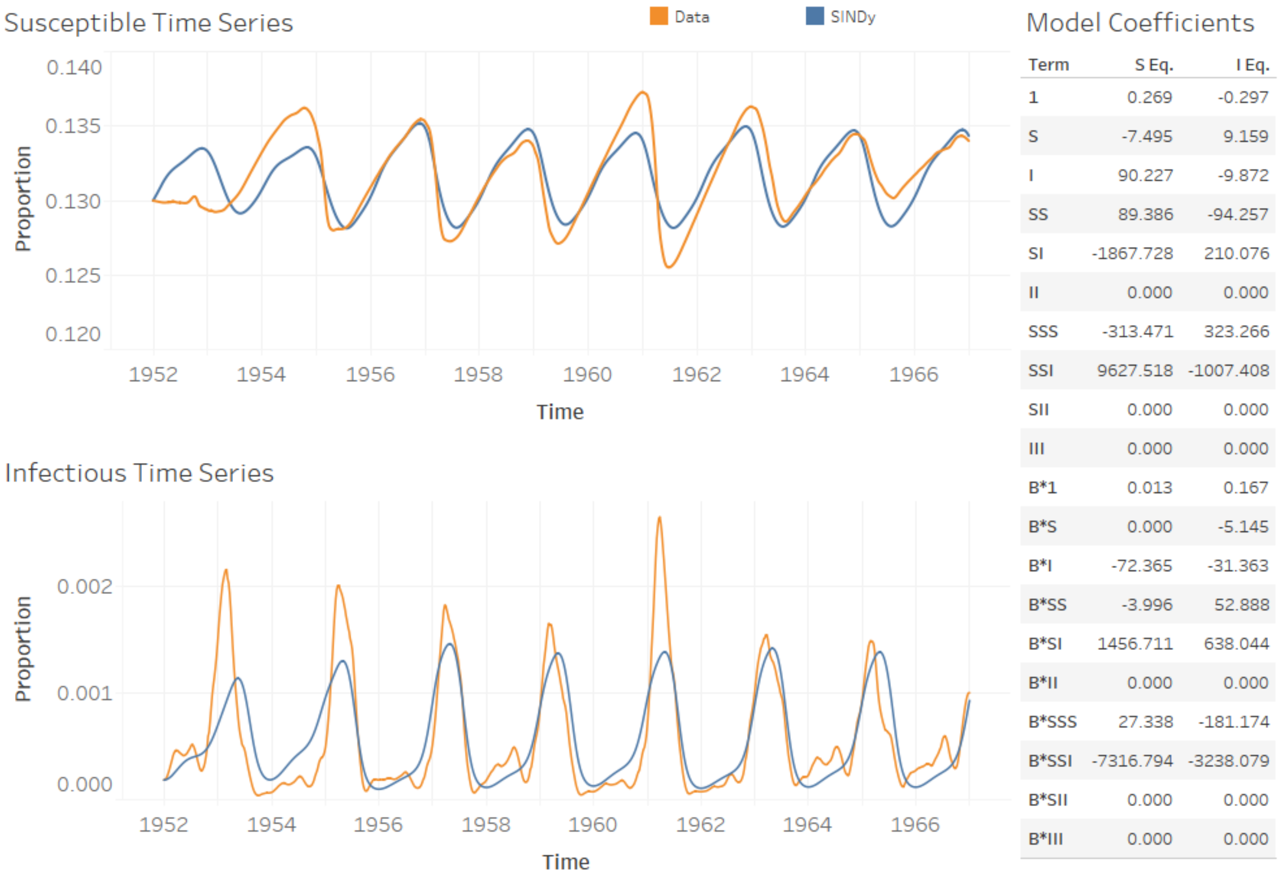


Model Coefficients

Term	S Eq.	I Eq.
1	0.025	0.002
S	0.606	-0.037
I	1.084	-1.554
SS	1.541	0.139
SI	-10.295	20.618
II	0.000	0.000
B*1	-0.009	-0.013
B*S	0.146	0.200
B*I	-3.122	0.000
B*SS	-0.591	-0.779
B*SI	26.409	0.000
B*II	0.000	0.000

Figure 3.9: Comparison between measles data and the best SINDy-discovered model using a function library of polynomials up to 2nd order. The discovered model accurately replicates the biennium present in the data in both the susceptible and infection classes. It also identifies a strong dependence on the SI and βSI cross terms, the driving terms behind the mass action incidence mechanism present in the SIR model. The sparse regression resulted in the exclusion of six terms, giving a sparsity ratio of 0.25.

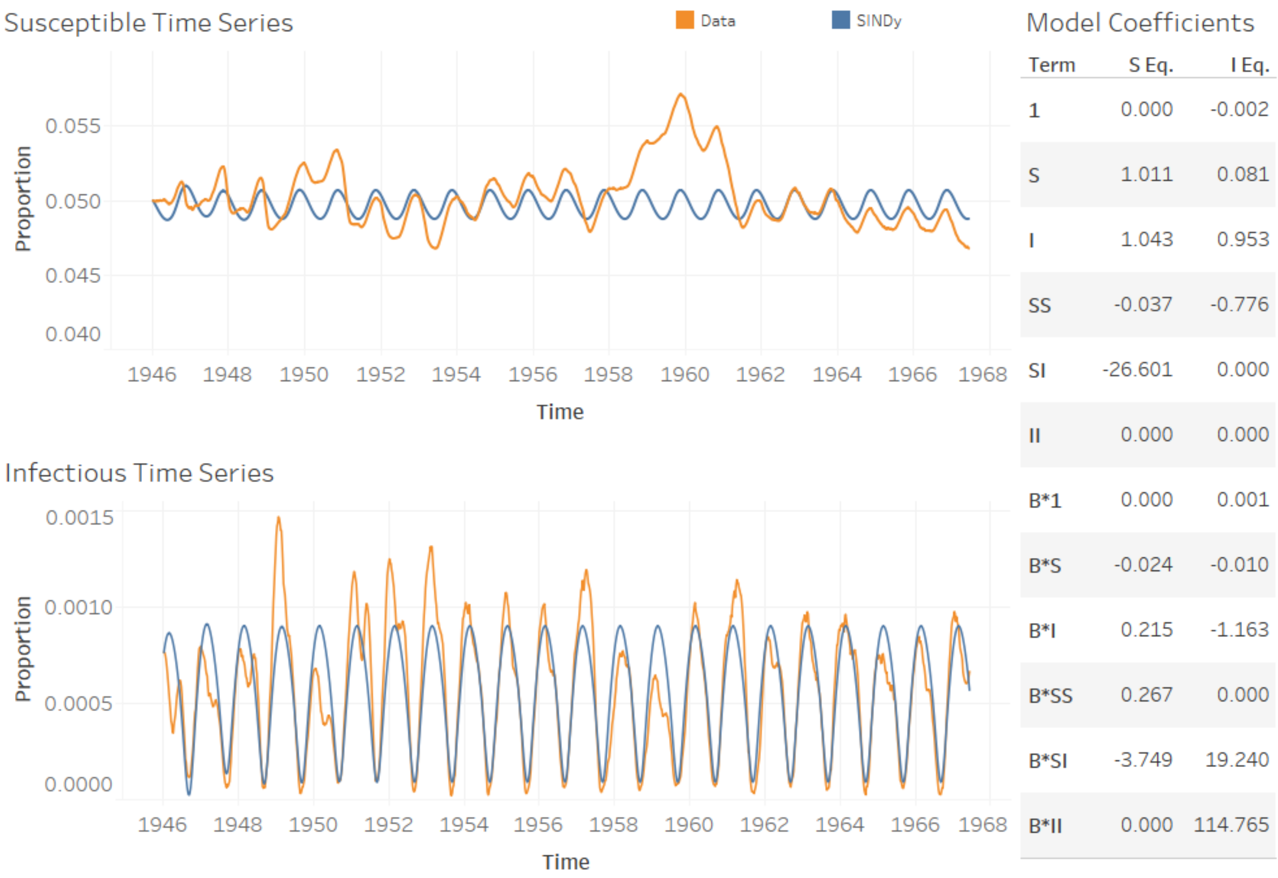
Measles (UK), 3rd Order Polynomials



Term	S Eq.	I Eq.
1	0.269	-0.297
S	-7.495	9.159
I	90.227	-9.872
SS	89.386	-94.257
SI	-1867.728	210.076
II	0.000	0.000
SSS	-313.471	323.266
SSI	9627.518	-1007.408
SII	0.000	0.000
III	0.000	0.000
B*1	0.013	0.167
B*S	0.000	-5.145
B*I	-72.365	-31.363
B*SS	-3.996	52.888
B*SI	1456.711	638.044
B*II	0.000	0.000
B*SSS	27.338	-181.174
B*SSI	-7316.794	-3238.079
B*SII	0.000	0.000
B*III	0.000	0.000

Figure 3.10: Comparison between measles data and the best SINDy-discovered model using a function library of polynomials up to 3rd order. As in the case above, the discovered model accurately replicates the biennium present in the data in both the susceptible and infection classes. It also again identifies a strong dependence on the SI cross term, as well as the S^2I and (to a lesser extent) the S^3 terms. The sparse regression resulted in the exclusion of thirteen terms, giving a sparsity ratio of 0.325.

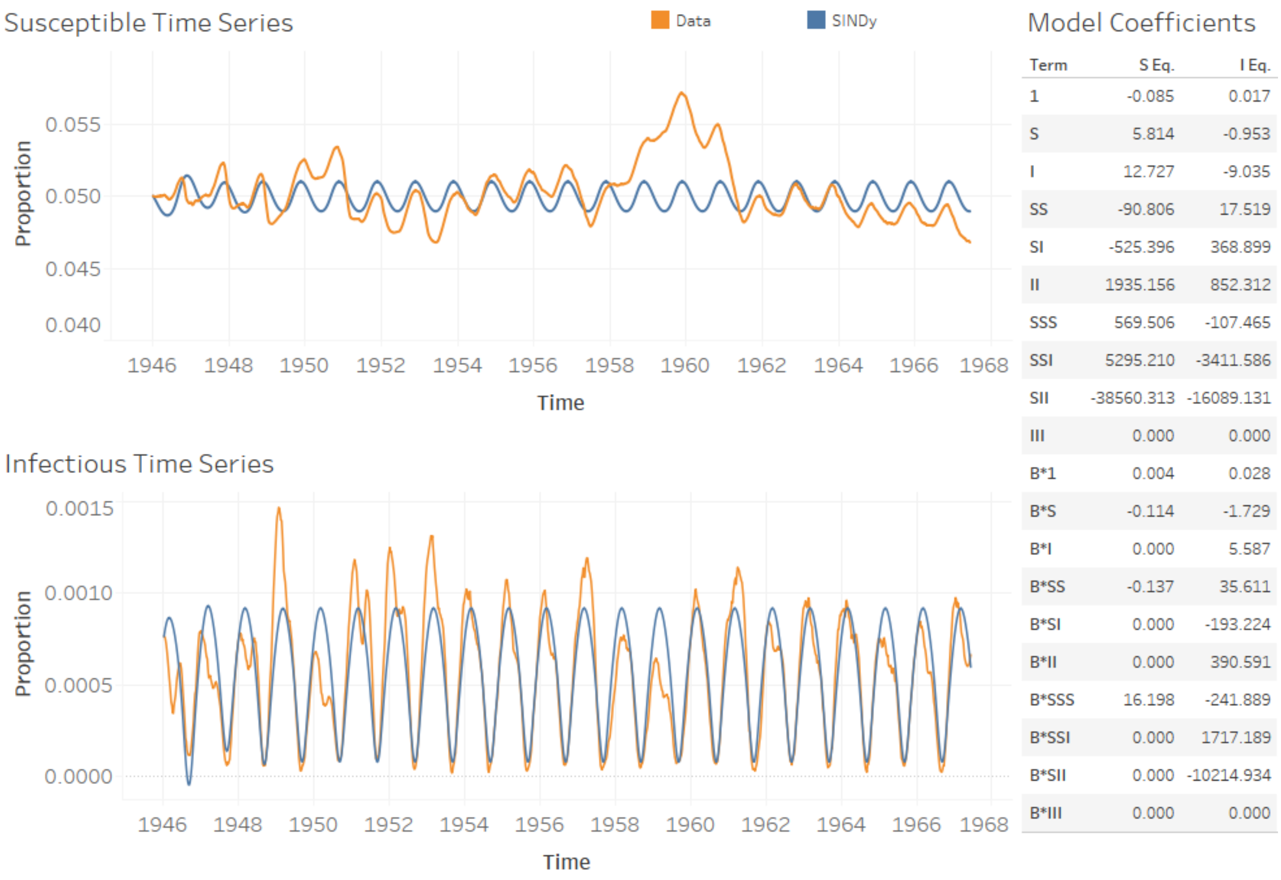
Varicella (Ontario), 2nd Order Polynomials



Term	S Eq.	I Eq.
1	0.000	-0.002
S	1.011	0.081
I	1.043	0.953
SS	-0.037	-0.776
SI	-26.601	0.000
II	0.000	0.000
B*1	0.000	0.001
B*S	-0.024	-0.010
B*I	0.215	-1.163
B*SS	0.267	0.000
B*SI	-3.749	19.240
B*II	0.000	114.765

Figure 3.11: Comparison between varicella data and the best SINDy-discovered model using a function library of polynomials up to 2nd order. The discovered model accurately replicates the annual cycle present in the data in both the susceptible and infection classes. As in the measles case, it also identifies a strong dependence on the mass action incidence term in both the S and I equations. Note also that the coefficient of S and I in their respective equations are close to 1, as expected in discrete disease models. The sparse regression resulted in the exclusion of six terms, giving a sparsity ratio of 0.25.

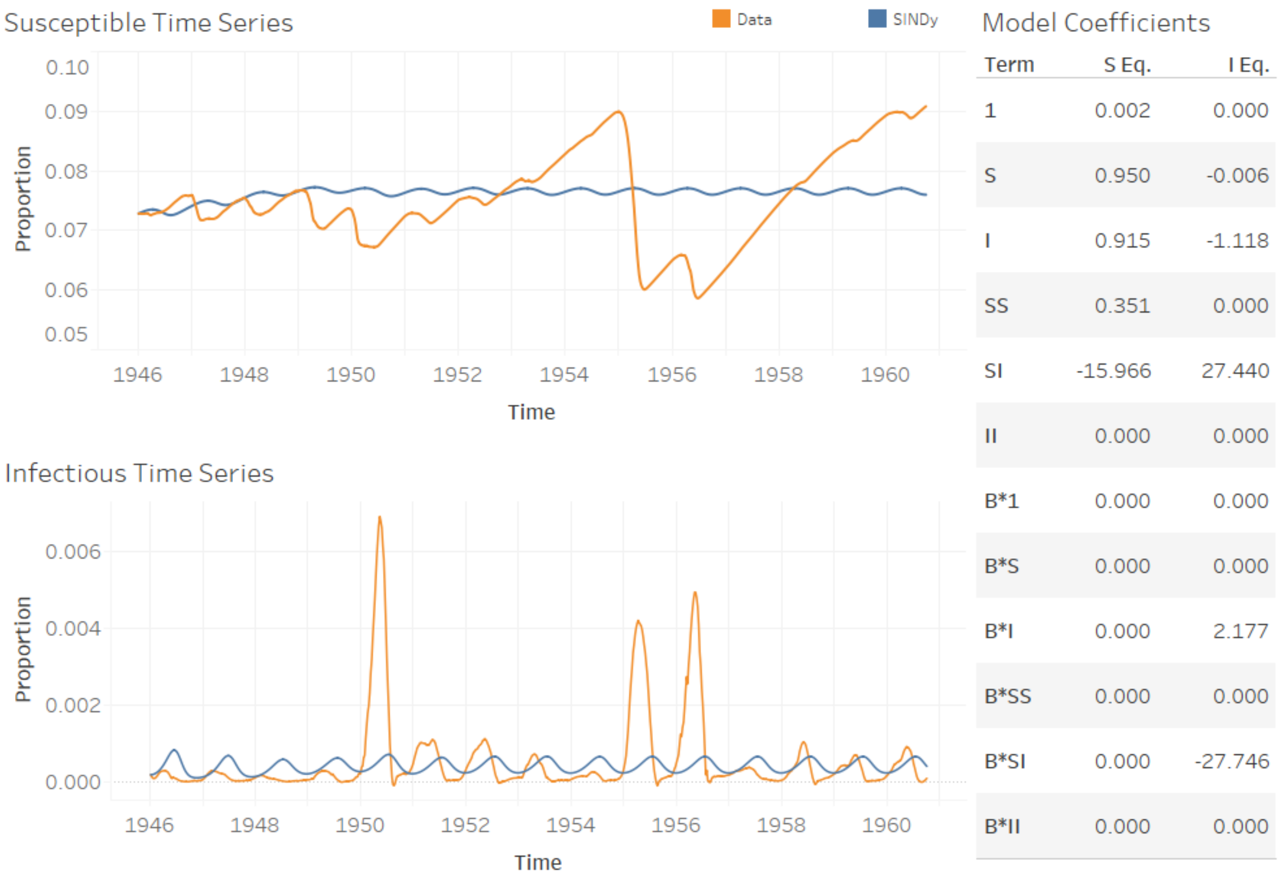
Varicella (Ontario), 3rd Order Polynomials



Term	S Eq.	I Eq.
1	-0.085	0.017
S	5.814	-0.953
I	12.727	-9.035
SS	-90.806	17.519
SI	-525.396	368.899
II	1935.156	852.312
SSS	569.506	-107.465
SSI	5295.210	-3411.586
SII	-38560.313	-16089.131
III	0.000	0.000
B*1	0.004	0.028
B*S	-0.114	-1.729
B*I	0.000	5.587
B*SS	-0.137	35.611
B*SI	0.000	-193.224
B*II	0.000	390.591
B*SSS	16.198	-241.889
B*SSI	0.000	1717.189
B*SII	0.000	-10214.934
B*III	0.000	0.000

Figure 3.12: Comparison between varicella data and the best SINDy-discovered model using a function library of polynomials up to 3rd order. Again the discovered model accurately replicates the annual cycle present in the data in both the susceptible and infection classes. The dependence on the mass action incidence term is again noticeable, though the S^2I, SI^2 and S^3 terms have dominant coefficients as well. The sparse regression resulted in the exclusion of nine terms, giving a sparsity ratio of 0.225.

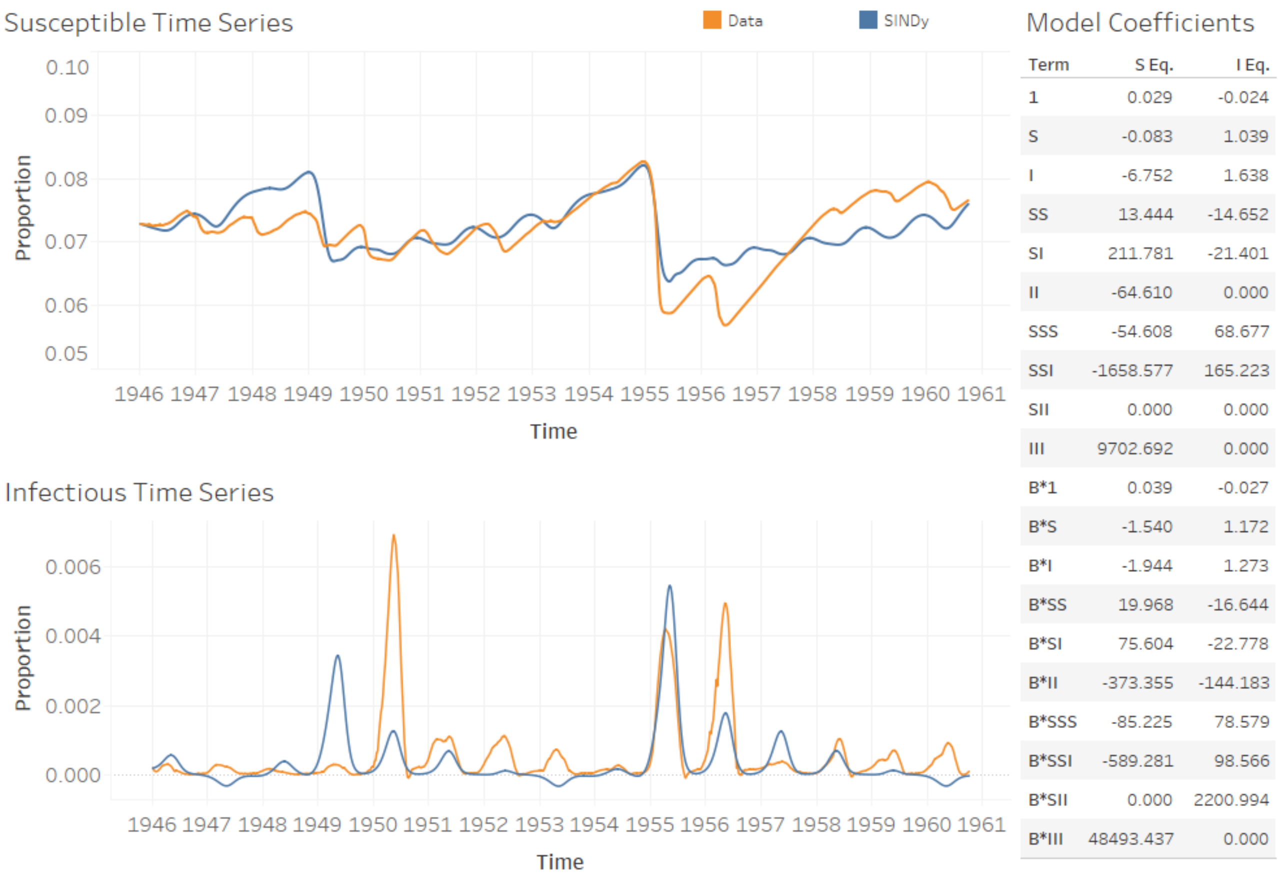
Rubella (Ontario), 2nd Order Polynomials



Term	S Eq.	I Eq.
1	0.002	0.000
S	0.950	-0.006
I	0.915	-1.118
SS	0.351	0.000
SI	-15.966	27.440
II	0.000	0.000
B*1	0.000	0.000
B*S	0.000	0.000
B*I	0.000	2.177
B*SS	0.000	0.000
B*SI	0.000	-27.746
B*II	0.000	0.000

Figure 3.13: Comparison between rubella data and the best SINDy-discovered model using a function library of polynomials up to 2nd order. The algorithm was unable to discover a model that exhibited the multiennial cycle observed in the data, instead returning an annual cycle. Despite this, strong dependence on the mass action incidence term is again present. The sparse regression resulted in the exclusion of fourteen terms, giving a sparsity ratio of 0.7.

Rubella (Ontario), 3rd Order Polynomials



Term	S Eq.	I Eq.
1	0.029	-0.024
S	-0.083	1.039
I	-6.752	1.638
SS	13.444	-14.652
SI	211.781	-21.401
II	-64.610	0.000
SSS	-54.608	68.677
SSI	-1658.577	165.223
SII	0.000	0.000
III	9702.692	0.000
B*1	0.039	-0.027
B*S	-1.540	1.172
B*I	-1.944	1.273
B*SS	19.968	-16.644
B*SI	75.604	-22.778
B*II	-373.355	-144.183
B*SSS	-85.225	78.579
B*SSI	-589.281	98.566
B*SII	0.000	2200.994
B*III	48493.437	0.000

Figure 3.14: Comparison between rubella data and the best SINDy-discovered model using a function library of polynomials up to 3rd order. The discovered model successfully recovers a multiennial cycle in the prevalence time series, similar to the one present in the data. It also replicates the temporal fluctuations from the mean present in the susceptible reconstruction. The dependence on the mass action incidence term again exists, though perhaps not as strong as in models recovered from the other diseases. The sparse regression resulted in the exclusion of six terms, giving a sparsity ratio of 0.15.

The time series plots in Figures 3.9-3.14 demonstrate the effectiveness of the SINDy algorithm at recovering a sparse model that closely mimics the dynamics of the empirical disease data for both the susceptible and infectious classes. Specifically, in the case of a biennium (Figures 3.9 and 3.10) and of an annual cycle (Figures 3.11 and 3.12) a second order library of polynomials was sufficient for SINDy to identify the attractor class of the underlying dynamics of the system. The algorithm also successfully identified the phase shift of the seasonal transmission rate, correctly syncing the peaks of the model with those of the data. In the case of the multi-annual rubella (Figures 3.13 and 3.14), a third order library was required to produce the appropriate cycle, despite it being shown previously that a second order library is sufficient to obtain a multi-annual attractor [11]. However, the inability of the model selection criteria to identify a model that exhibits a multiennial attractor is most likely caused by a misalignment of major peaks, which results in a poor residual error. In this case, comparison of power spectral density estimates can be more effective at selecting models with similar dynamics, as discussed in Section 2.2.6. The results when using this technique are presented in section 3.2.5.

The promotion of sparsity within the model is also noticeable within the model coefficients, where dominant terms are clearly present. Specifically, the algorithm assigned dominant coefficients to terms corresponding with bilinear incidence (SI and βSI), a driving mechanism in the standard SIR model. Additionally, when a third order library was included, a strong dependence was put on the S^2I and SI^2 terms (and their corresponding seasonal terms), which may indicate a more general nonlinear incidence mechanism is present in the underlying system. It has been shown in Refs. [83, 84] that an incidence function of the form S^pI^q (where $p, q > 0$) may more adequately represent some endemic cycles, a form that is present when a 3rd order polynomial library is included. Impacts of this result are discussed further in Section 4.3.

It is also important to note that the algorithm has sensitivity to both the initial susceptible value and the thresholding cutoff value, which creates a segue to both of the subsequent sections: a connection with compartmental modelling techniques and the need to balance the goodness of fit of the recovered model with relative sparsity in the selection of active terms.

3.2.3 Balancing Sparsity with Goodness of Fit

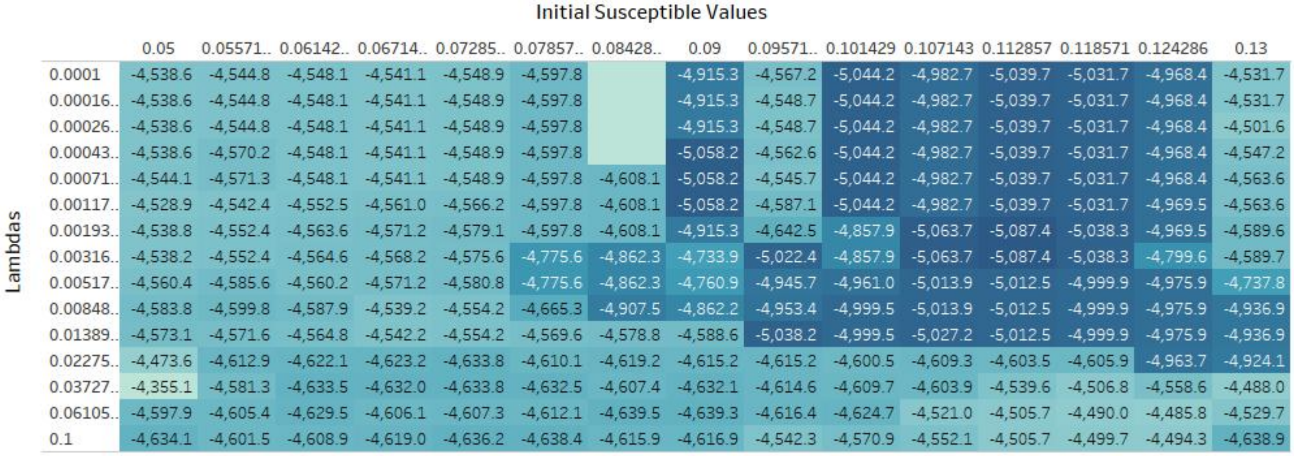
Perhaps the most important parameter of the SINDy algorithm is the sparsity knob λ. This value defines the threshold below which coefficients (and their corresponding features) are removed from the library at each iteration. Varying this parameter has a dramatic effect on the sparsity and resulting behaviour and predictive value of the discovered model. If the threshold is set too low then the generated model will include most (or all) of the functions in the library, resulting in overfitting. Conversely, if the threshold is set too high then features required to emulate the dynamics of the system may be removed, resulting in a model that is too simplistic. AIC scores (see Section 1.2.3) provide a measure of relative quality between models and is used for model selection when sparsity is desirable. When applying SINDy to empirical data, then, it is possible to automate the model selection process by running the algorithm over a chosen range of the thresholding parameter and choosing the model that gives the lowest AIC value.

In the context of epidemiological data used in this work, another key parameter is the initial susceptible value (S_0). There is no feature of the raw case notification data that would suggest a starting point for the susceptible time series, and empirical data on the susceptible class is rarely available for a given demographic. Theory dictates [63] that the average susceptible value can be approximated as

$$\bar{S} \approx \frac{1}{\mathcal{R}_0}$$

which gives an estimate for the initial value, as the basic reproductive ratio is known for the infectious diseases in question. However, this estimate is not without uncertainty, and the dynamics of the system are strongly influenced by the initial condition. Thus it is necessary to investigate the SINDy-discovered models across a biologically relevant range of initial susceptible proportions.

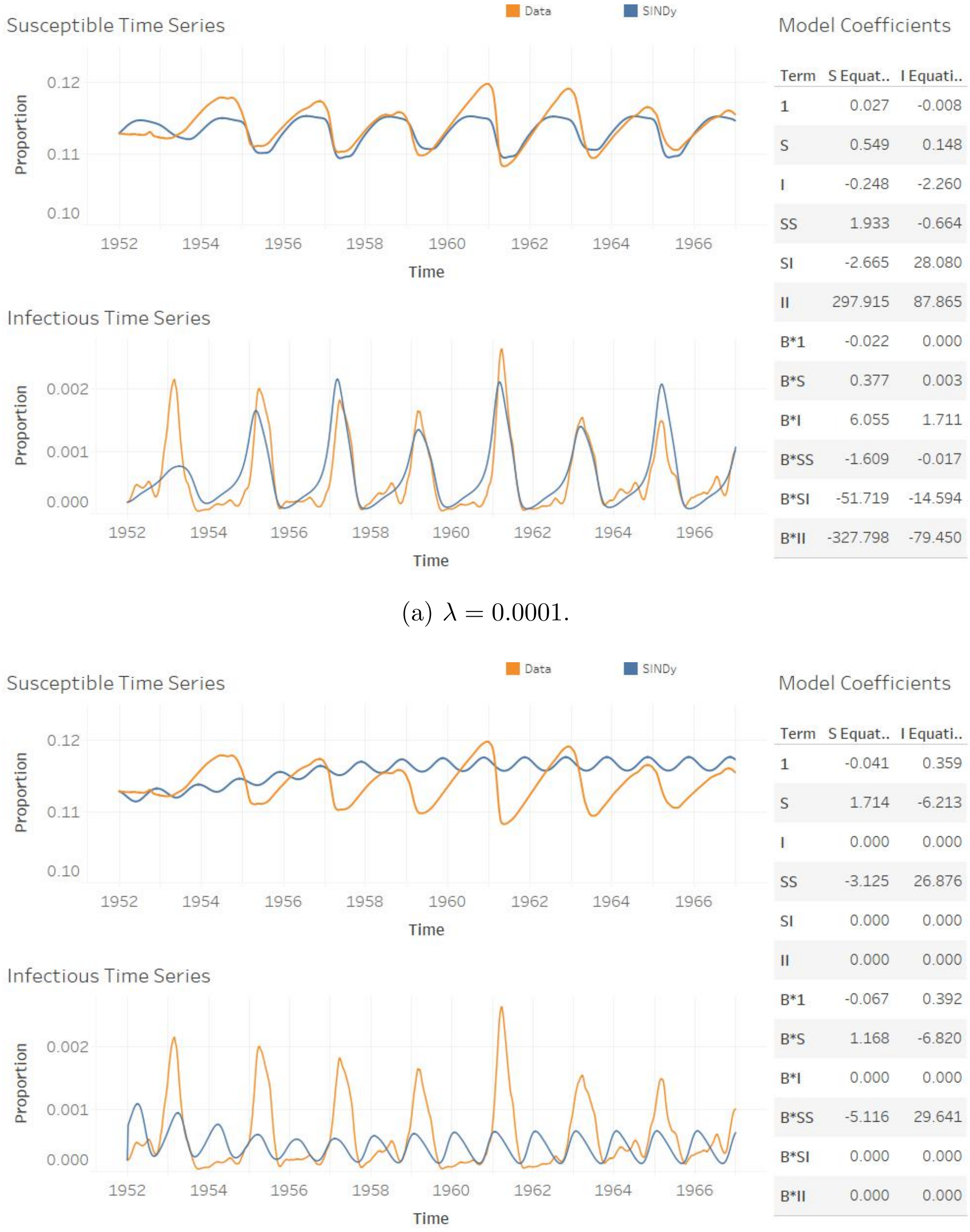
For each of the three diseases, a range of initial susceptible and threshold values were selected. The susceptible class was reconstructed (using the locally linear regression method described in Section 2.2.2) and the SINDy algorithm was applied to the data using each possible pair (S_0, λ) from a parameter grid created by linear sampling from the respective ranges. The AIC values comparing the discovered models to the empirical data were recorded, and the values for the UK measles dataset are presented in Figure 3.15 (the results for the other two datasets are found in Appendix A).



Lambdas	Initial Susceptible Values														
	0.05	0.05571..	0.06142..	0.06714..	0.07285..	0.07857..	0.08428..	0.09	0.09571..	0.101429	0.107143	0.112857	0.118571	0.124286	0.13
0.0001	-4,538.6	-4,544.8	-4,548.1	-4,541.1	-4,548.9	-4,597.8		-4,915.3	-4,567.2	-5,044.2	-4,982.7	-5,039.7	-5,031.7	-4,968.4	-4,531.7
0.00016..	-4,538.6	-4,544.8	-4,548.1	-4,541.1	-4,548.9	-4,597.8		-4,915.3	-4,548.7	-5,044.2	-4,982.7	-5,039.7	-5,031.7	-4,968.4	-4,531.7
0.00026..	-4,538.6	-4,544.8	-4,548.1	-4,541.1	-4,548.9	-4,597.8		-4,915.3	-4,548.7	-5,044.2	-4,982.7	-5,039.7	-5,031.7	-4,968.4	-4,501.6
0.00043..	-4,538.6	-4,570.2	-4,548.1	-4,541.1	-4,548.9	-4,597.8		-5,058.2	-4,562.6	-5,044.2	-4,982.7	-5,039.7	-5,031.7	-4,968.4	-4,547.2
0.00071..	-4,544.1	-4,571.3	-4,548.1	-4,541.1	-4,548.9	-4,597.8	-4,608.1	-5,058.2	-4,545.7	-5,044.2	-4,982.7	-5,039.7	-5,031.7	-4,968.4	-4,563.6
0.00117..	-4,528.9	-4,542.4	-4,552.5	-4,561.0	-4,566.2	-4,597.8	-4,608.1	-5,058.2	-4,587.1	-5,044.2	-4,982.7	-5,039.7	-5,031.7	-4,969.5	-4,563.6
0.00193..	-4,538.8	-4,552.4	-4,563.6	-4,571.2	-4,579.1	-4,597.8	-4,608.1	-4,915.3	-4,642.5	-4,857.9	-5,063.7	-5,087.4	-5,038.3	-4,969.5	-4,589.6
0.00316..	-4,538.2	-4,552.4	-4,564.6	-4,568.2	-4,575.6	-4,775.6	-4,862.3	-4,733.9	-5,022.4	-4,857.9	-5,063.7	-5,087.4	-5,038.3	-4,799.6	-4,589.7
0.00517..	-4,560.4	-4,585.6	-4,560.2	-4,571.2	-4,580.8	-4,775.6	-4,862.3	-4,760.9	-4,945.7	-4,961.0	-5,013.9	-5,012.5	-4,999.9	-4,975.9	-4,737.8
0.00848..	-4,583.8	-4,599.8	-4,587.9	-4,539.2	-4,554.2	-4,665.3	-4,907.5	-4,862.2	-4,953.4	-4,999.5	-5,013.9	-5,012.5	-4,999.9	-4,975.9	-4,936.9
0.01389..	-4,573.1	-4,571.6	-4,564.8	-4,542.2	-4,554.2	-4,569.6	-4,578.8	-4,588.6	-5,038.2	-4,999.5	-5,027.2	-5,012.5	-4,999.9	-4,975.9	-4,936.9
0.02275..	-4,473.6	-4,612.9	-4,622.1	-4,623.2	-4,633.8	-4,610.1	-4,619.2	-4,615.2	-4,615.2	-4,600.5	-4,609.3	-4,603.5	-4,605.9	-4,963.7	-4,924.1
0.03727..	-4,355.1	-4,581.3	-4,633.5	-4,632.0	-4,633.8	-4,632.5	-4,607.4	-4,632.1	-4,614.6	-4,609.7	-4,603.9	-4,539.6	-4,506.8	-4,558.6	-4,488.0
0.06105..	-4,597.9	-4,605.4	-4,629.5	-4,606.1	-4,607.3	-4,612.1	-4,639.5	-4,639.3	-4,616.4	-4,624.7	-4,521.0	-4,505.7	-4,490.0	-4,485.8	-4,529.7
0.1	-4,634.1	-4,601.5	-4,608.9	-4,619.0	-4,636.2	-4,638.4	-4,615.9	-4,616.9	-4,542.3	-4,570.9	-4,552.1	-4,505.7	-4,499.7	-4,494.3	-4,638.9

Figure 3.15: The AIC values for SINDy models across a range of both initial susceptible and threshold values, utilizing the measles dataset and a 2nd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model. Cells without value refer to a model which, when simulated, resulted in a diverging time series.

These values show that there exists an optimal region in the $S_0 - \lambda$ plane that should be chosen to ensure the SINDy algorithm can generate regularized, accurate models from empirical disease data. The model selected by the AIC scores as the most parsimonious used an initial susceptible value of $S_0 = 0.11286$ and a sparsity knob of $\lambda = 0.00517$. This is the model shown in Figure 3.9 and is an example of a model that balances a well-fitting time series with a sparse set of active terms. Using the same dataset, function library, and initial susceptible value, a much lower sparsity knob ($\lambda = 0.0001$) was taken and the resulting model is shown in Figure 3.16a. As expected, the model is overfit, resulting in a time series that is too closely fit to the random fluctuations in peak height as well as a model that has all possible functions active. Conversely, if the sparsity knob is set much higher ($\lambda = 0.1$) the discovered model no longer exhibits the biennium that the data does, but rather an annual attractor that does not match the dynamics of the system (Figure 3.16b).



Model Coefficients

Term	S Equat..	I Equati..
1	0.027	-0.008
S	0.549	0.148
I	-0.248	-2.260
SS	1.933	-0.664
SI	-2.665	28.080
II	297.915	87.865
B*1	-0.022	0.000
B*S	0.377	0.003
B*I	6.055	1.711
B*SS	-1.609	-0.017
B*SI	-51.719	-14.594
B*II	-327.798	-79.450

(a) $\lambda = 0.0001$.

Model Coefficients

Term	S Equat..	I Equati..
1	-0.041	0.359
S	1.714	-6.213
I	0.000	0.000
SS	-3.125	26.876
SI	0.000	0.000
II	0.000	0.000
B*1	-0.067	0.392
B*S	1.168	-6.820
B*I	0.000	0.000
B*SS	-5.116	29.641
B*SI	0.000	0.000
B*II	0.000	0.000

(b) $\lambda = 1$.

Figure 3.16: Resulting time series and coefficients from SINDy-discovered models of the measles dataset (with a 2nd order polynomial library) using extreme sparsity thresholds. Using a relatively small threshold ($\lambda = 0.0001$) results in a good fit and accurate recovery of attractor class, but a very low number of non-active terms, which is an indicator of an overfit model (a). Conversely, using a relatively large threshold ($\lambda = 1$) gives a sparse model, but at the cost of a good fit and recovery of attractor class (b).

3.2.4 Comparison with Compartmental Models

The basis for many modern mathematical models in epidemiology lies in compartmental modelling, primarily stemming from the seminal work of Kermack and McKendrick [5]. The theory behind these modelling techniques is discussed in depth in Chapter 1. For the purposes of this section, recall that perhaps the most fundamental compartmental model used to model periodically-occurring infectious diseases is the seasonally-forced SIR model, which (given a discrete timescale) takes the form

$$\begin{aligned} S_{t+1} &= S_t + \nu - \beta_t S_t I_t - \mu S_t \\ I_{t+1} &= I_t + \beta_t S_t I_t - \gamma I_t - \mu I_t \\ R_{t+1} &= R_t + \gamma I_t - \mu R_t \end{aligned}$$

where β_t is the discrete analogue of the seasonally-varying transmission rate given in Eq. 1.5, given by

$$\beta_t = \beta_0(1 + \beta_1 \text{cos}(2\pi t/T - \phi)).$$

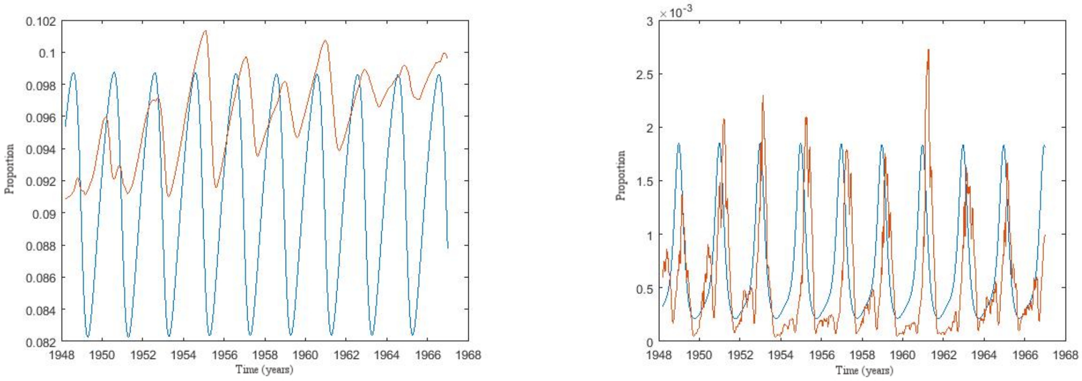
Several prominent features exist within this model, the most important of which is the existence of the mass action incidence term ($\beta_t S_t I_t$) in both the S_{t+1} and I_{t+1} equations. This term represents the theoretical mechanism by which susceptible individuals are transferred to the infectious class, and the weight that this term is given heavily influences the dynamics of the model. Besides this term, the only other active terms present are linear.

It is of interest, then, to compare the data-driven models discovered by the SINDy algorithm to the theoretically-derived SIR model presented above. In order to do this, we must first obtain estimates for the parameters in the SIR model by fitting it to the empirical data. This can be done by simulating the model across a wide range of linearly-spaced parameter values and selecting the model which minimizes the sum of squares error between the simulated model and the observed data. Baseline values for the parameters were taken from Refs. [11, 63]. For simplicity a closed system was assumed, implying that the birth and death rates were equal ($\nu = \mu$). Also, recall that given basic reproductive ratio $\mathcal{R}_0$ and recovery rate γ, the mean transmission rate is completely determined by $\beta_0 = \mathcal{R}_0 \cdot \gamma$. The parameters that were varied, with corresponding ranges and step sizes, are presented in Table 3.5:

Symbol	$\boldsymbol{\mathcal{R}_0}$	$\boldsymbol{\gamma}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\mu}$	ϕ
Description	Basic rep. ratio	Recovery rate	Forcing amplitude	Birth/death rate	Forcing phase
Range	6 - 16	0.55 - 1.25	0.05 - 0.35	3×10^{-4} - 6×10^{-4}	0-51.5
Step Size	0.5	0.05	0.025	5×10^{-5}	0.5

Table 3.5: Parameters, ranges, and parameter step sizes used for fitting discrete SIR model (Eqs. 3.5 - 3.7) to empirical data for each of the three disease datasets, using a timestep of $\Delta t = 1$week.

Simulations of the discrete SIR model were run at each point in the parameter plane, beginning 150 years prior to the temporal range of the data to eliminate the effects of transients and the impact of the initial conditions, which were fixed at $(S_0, I_0) = (0.1, 5 \times 10^{-5})$. The resulting models with the minimal sum of squares error when compared with the data were selected and plotted in Figures 3.17a - 3.19b.



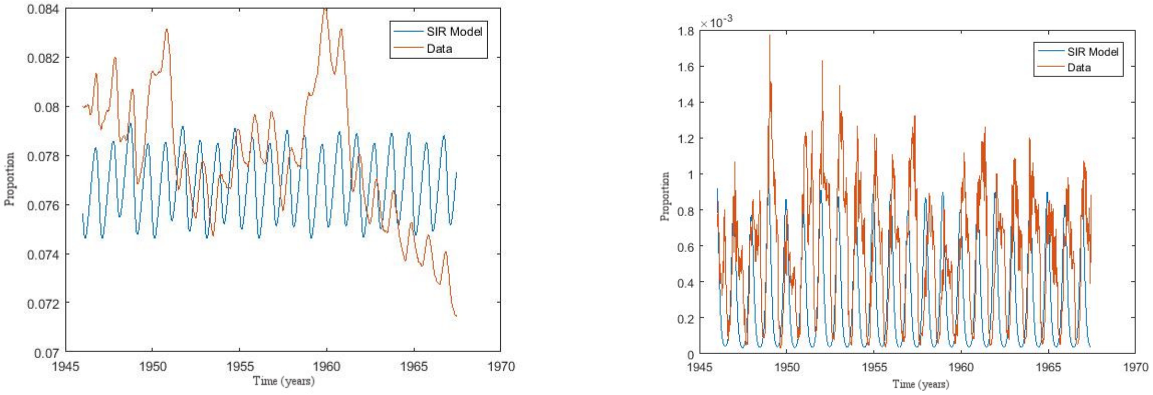
(a) Susceptibles

(b) Infectious

$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$
7.7	0.03	0.7	0.0005

Figure 3.17: Comparison between the empirical data (orange) and a fitted SIR model (blue) for the measles dataset. Parameters used when simulating the discrete SIR model (defined by Eqs. 1.9 - 1.11, using a time step of $\Delta t = 1$week) were obtained by sweeping across the grid defined by the ranges and parameter step sizes given in Table 3.5, selecting parameters that gave the model that minimized residual error.

.



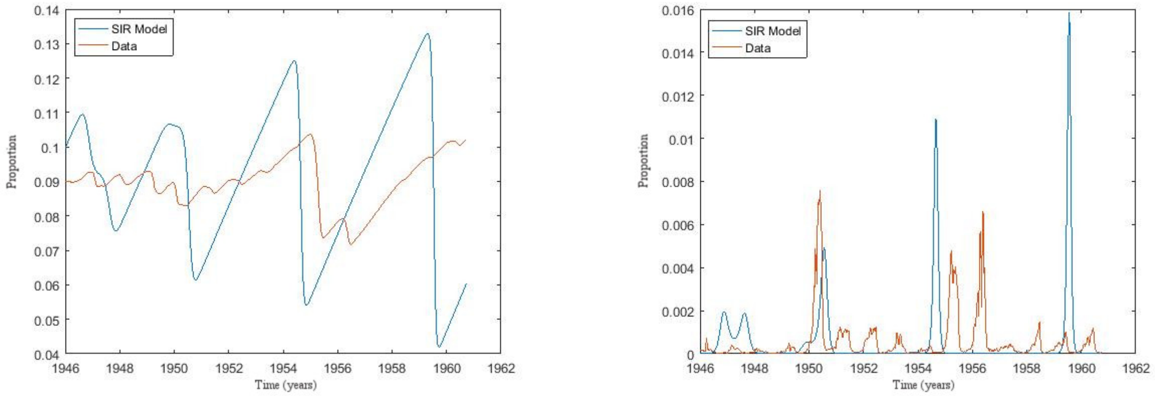
(a) Susceptibles

(b) Infectious

$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$
7.8	0.3	0.6	0.0002

Figure 3.18: Comparison between the empirical data (orange) and a fitted SIR model (blue) for the varicella dataset. Parameters used when simulating the discrete SIR model (defined by Eqs. 1.9 - 1.11, using a time step of $\Delta t = 1$week) were obtained by sweeping across the grid defined by the ranges and parameter step sizes given in Table 3.5, selecting parameters that gave the model that minimized residual error.

.



(a) Susceptibles

(b) Infectious

$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma}$	$\boldsymbol{\mu}$
7.7	0.15	0.7	0.0004

Figure 3.19: Comparison between the empirical data (orange) and a fitted SIR model (blue) for the rubella dataset. Parameters used when simulating the discrete SIR model (defined by Eqs. 1.9 - 1.11, using a time step of $\Delta t = 1$week) were obtained by sweeping across the grid defined by the ranges and parameter step sizes given in Table 3.5, selecting parameters that gave the model that minimized residual error.

.

Once the models were fit, the most-parsimonious 2nd order SINDy models (Figures 3.9, 3.11, and 3.13) were selected and compared to the corresponding models. Bar chart figures were constructed for each of the datasets that contrast the magnitude and sign of the coefficients.

Measles

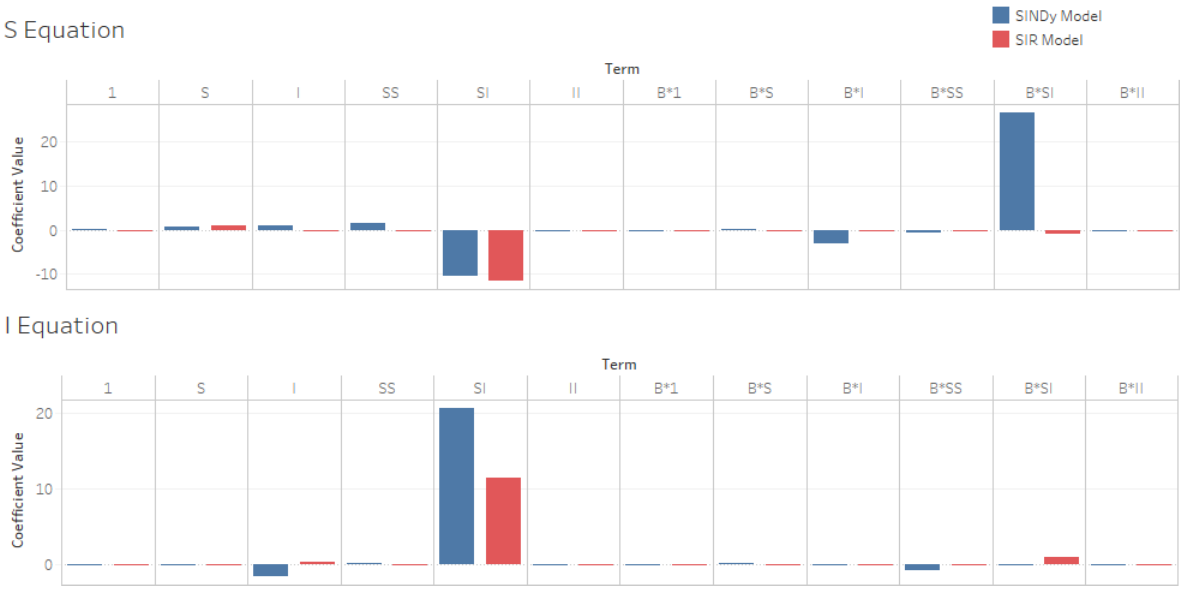


Figure 3.20: Comparison of coefficients between SINDy-discovered (using a function library of 1st and 2nd order polynomials) and fitted SIR model for the measles dataset.

Varicella

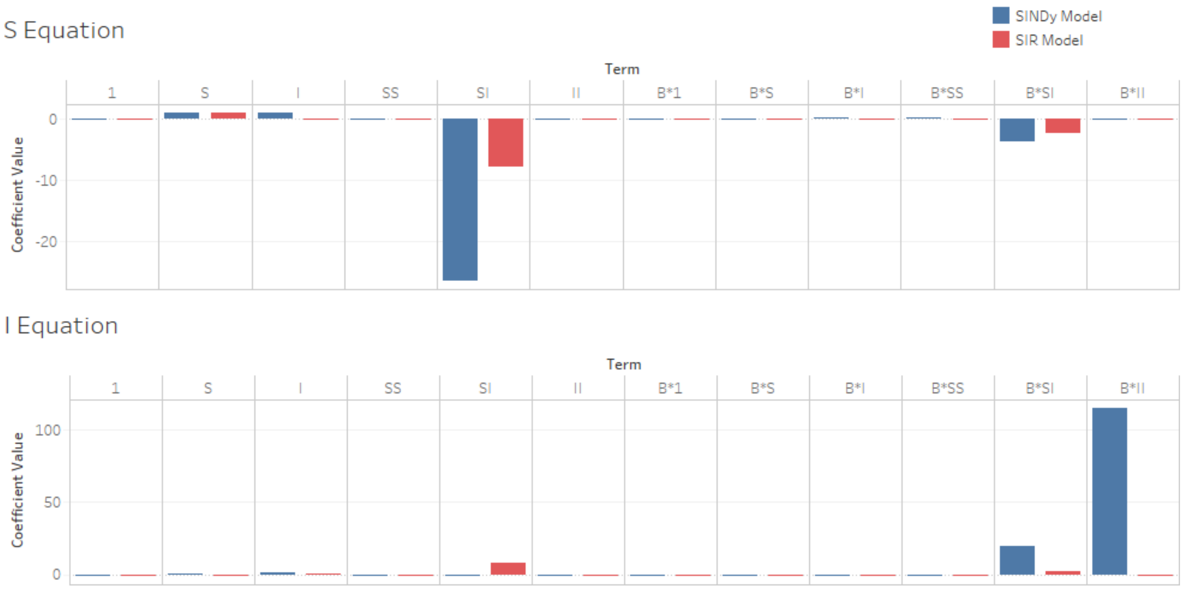


Figure 3.21: Comparison of coefficients between SINDy-discovered (using a function library of 1st and 2nd order polynomials) and fitted SIR model for the varicella dataset.

Rubella

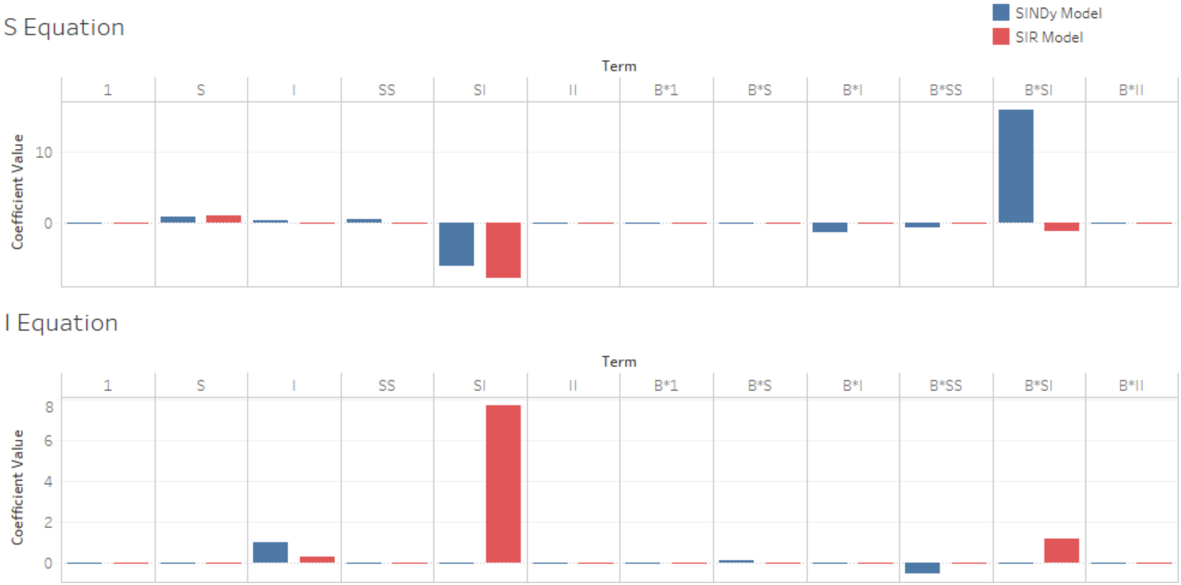


Figure 3.22: Comparison of coefficients between SINDy-discovered (using a function library of 1st and 2nd order polynomials) and fitted SIR model for the rubella dataset.

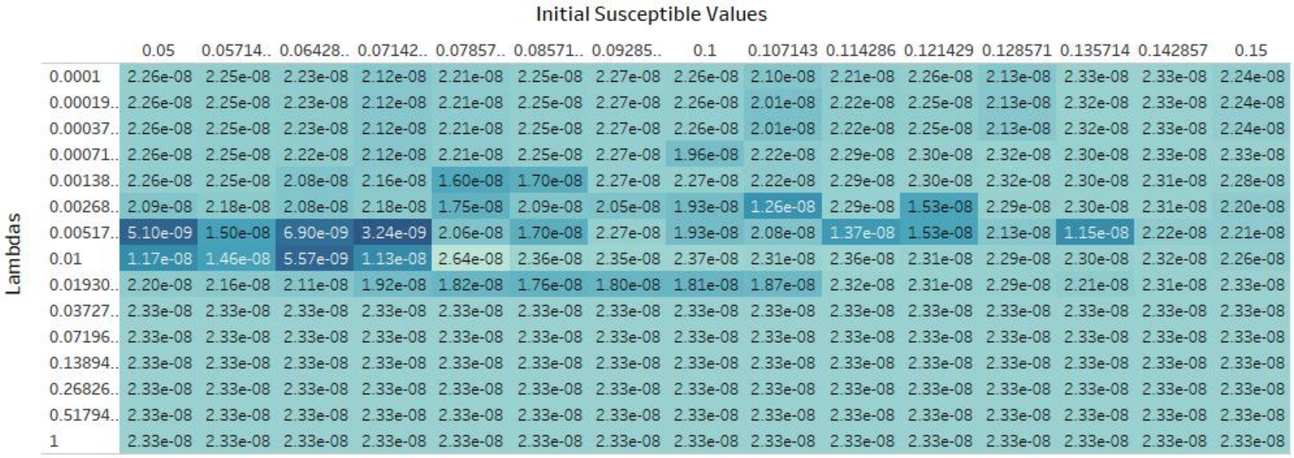
Perhaps the most noticeable feature in this comparison is the striking similarities in the SI terms for both the data-driven and theoretical models, most notably in the model of the measles data but to a lesser extent in the model of the varicella data. This demonstrates the effectiveness of the SINDy algorithm in capturing the theoretical principle of mass action incidence, a driving component of most epidemiological models. The SINDy models also exhibit dependence on the corresponding linear terms for each of the S and I equations, though not always with the expected magnitude and sign. Additionally, several other features are noticeably different than the theoretically-dictated model, indicating that drawing direct biological conclusions from any given term at present is not feasible. However, the algorithm's success recovering a 2nd order model that captures both the observed disease dynamics (Figures 3.9, 3.11, and 3.13) and prominent features of current theoretical models shows promise that this data-driven technique could lend insight to model creation and selection practices.

3.2.5 Using Power Spectral Density for Model Selection

Model selection using AIC score, while adept at selecting parsimonious models that fit the data well, can fail when qualitative features of the underlying system are viewed as more important that accuracy of fit. When modelling disease dynamics, recovery of the attractor class present within the data is a key concern, and in certain cases more so than a precise fit to the specific data. This is exemplified when using AIC as a model selection metric to identify the most parsimonious SINDy recovered model based on the rubella dataset (Figures 3.13 - 3.14). When using a second order library, the model with the lowest AIC score does not recover the multiennial pattern present in the data. However, given the right set of initial parameters (initial susceptible value, sparsity knob, and seasonal transmission phase) SINDy can identify a model that recovers these dynamics, though perhaps not with correctly aligned major outbreaks. This will result in large residual error which gives a higher AIC score, despite the model being qualitatively better. Thus in cases where identifying the frequency and magnitude of outbreaks (peaks) is more important than exactly fitting the data, another model selection criteria is necessary, one that quantitatively compares these elements.

Power spectral density (discussed in Section 2.2.6) is useful for this exact purpose, by providing a quantitative way to compare the attractor present within a discovered model to that present within the data. Determining the fit (through residual error or similar means) of the spectral density plot computed from a simulated model to that computed from the empirical data can be used to identify which model provides the best fit of attractor class. However, sparsity is still to be valued within these models, which is why the AIC score was initially used for model selection. This leads to a model selection process of computing the AIC score of the spectral densities of the model and the data, which will promote a parsimonious model that attempts to match the qualitative features present in the data.

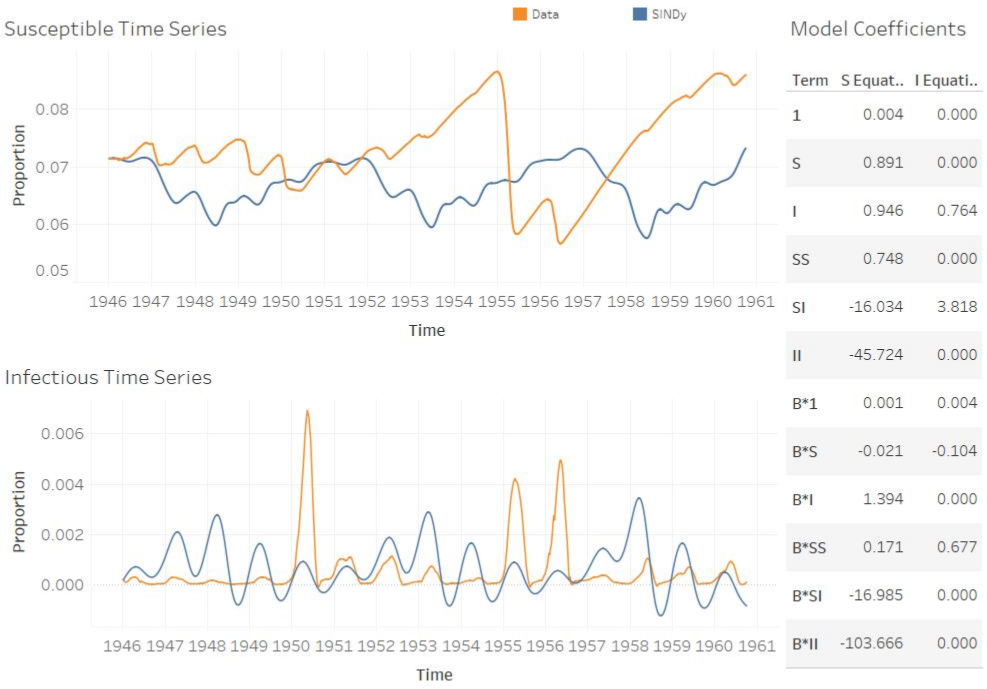
SINDy models were computed using a similar parameter sweep as when using the standard AIC selection method (see Section 3.2.3) except in this instance the power spectral density of the infectious time series of each model was computed and the AIC score between each spectral density and that of the corresponding data was found. These values for models discovered from the rubella dataset using a function library of up to 2nd order polynomials are shown in Figure 3.23.



	Initial Susceptible Values														
Lambdas	0.05	0.05714..	0.06428..	0.07142..	0.07857..	0.08571..	0.09285..	0.1	0.107143	0.114286	0.121429	0.128571	0.135714	0.142857	0.15
0.0001	2.26e-08	2.25e-08	2.23e-08	2.12e-08	2.21e-08	2.25e-08	2.27e-08	2.26e-08	2.10e-08	2.21e-08	2.26e-08	2.13e-08	2.33e-08	2.33e-08	2.24e-08
0.00019..	2.26e-08	2.25e-08	2.23e-08	2.12e-08	2.21e-08	2.25e-08	2.27e-08	2.26e-08	2.01e-08	2.22e-08	2.25e-08	2.13e-08	2.32e-08	2.33e-08	2.24e-08
0.00037..	2.26e-08	2.25e-08	2.23e-08	2.12e-08	2.21e-08	2.25e-08	2.27e-08	2.26e-08	2.01e-08	2.22e-08	2.25e-08	2.13e-08	2.32e-08	2.33e-08	2.24e-08
0.00071..	2.26e-08	2.25e-08	2.22e-08	2.12e-08	2.21e-08	2.25e-08	2.27e-08	1.96e-08	2.22e-08	2.29e-08	2.30e-08	2.32e-08	2.30e-08	2.33e-08	2.33e-08
0.00138..	2.26e-08	2.25e-08	2.08e-08	2.16e-08	1.60e-08	1.70e-08	2.27e-08	2.27e-08	2.22e-08	2.29e-08	2.30e-08	2.32e-08	2.30e-08	2.31e-08	2.28e-08
0.00268..	2.09e-08	2.18e-08	2.08e-08	2.18e-08	1.75e-08	2.09e-08	2.05e-08	1.93e-08	1.26e-08	2.29e-08	1.53e-08	2.29e-08	2.30e-08	2.31e-08	2.20e-08
0.00517..	5.10e-09	1.50e-08	6.90e-09	3.24e-09	2.06e-08	1.70e-08	2.27e-08	1.93e-08	2.08e-08	1.37e-08	1.53e-08	2.13e-08	1.15e-08	2.22e-08	2.21e-08
0.01	1.17e-08	1.46e-08	5.57e-09	1.13e-08	2.64e-08	2.36e-08	2.35e-08	2.37e-08	2.31e-08	2.36e-08	2.31e-08	2.29e-08	2.30e-08	2.32e-08	2.26e-08
0.01930..	2.20e-08	2.16e-08	2.11e-08	1.92e-08	1.82e-08	1.76e-08	1.80e-08	1.81e-08	1.87e-08	2.32e-08	2.31e-08	2.29e-08	2.21e-08	2.31e-08	2.33e-08
0.03727..	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08
0.07196..	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08
0.13894..	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08
0.26826..	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08
0.51794..	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08
1	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08	2.33e-08

Figure 3.23: The AIC values for SINDy models generated from power spectral density estimates of the infectious time series across a range of both initial susceptible and threshold values, using the rubella dataset and a 2nd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model.

The most parsimonious model as determined by this model selection criterion is shown in Figure 3.24. The corresponding power spectral density estimates of the infectious time series of both the simulated model and the data are shown in Figure 3.25. Despite qualitative improvement from the model selected using a 2nd order library and the standard AIC method (Figure 3.13), the multiennial pattern is still lacking in the resulting infectious time series. However, given inspection of other ”candidate” models with low AIC values from Figure 3.23, the model presented in Figure 3.26 was discovered, which does recover a multiennial pattern similar to the one present in the data.



Model Coefficients

Term	S Equat..	I Equati..
1	0.004	0.000
S	0.891	0.000
I	0.946	0.764
SS	0.748	0.000
SI	-16.034	3.818
II	-45.724	0.000
B*1	0.001	0.004
B*S	-0.021	-0.104
B*I	1.394	0.000
B*SS	0.171	0.677
B*SI	-16.985	0.000
B*II	-103.666	0.000

Figure 3.24: Comparison between rubella data and the best SINDy-discovered model using a function library of polynomials up to 2nd order and spectral density estimates for model selection.

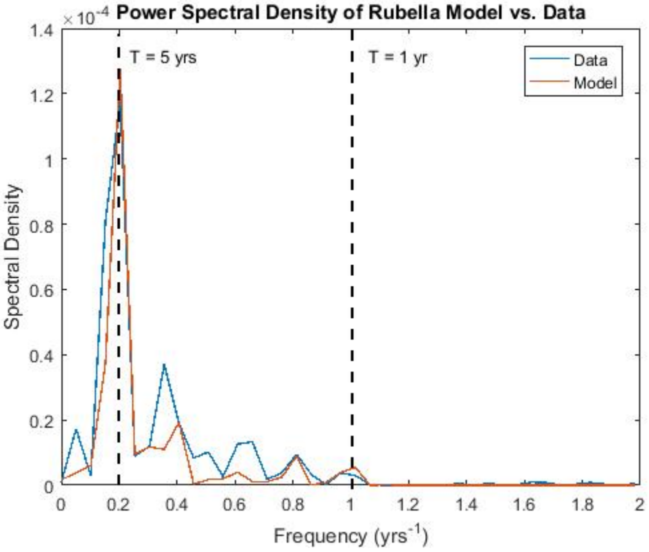
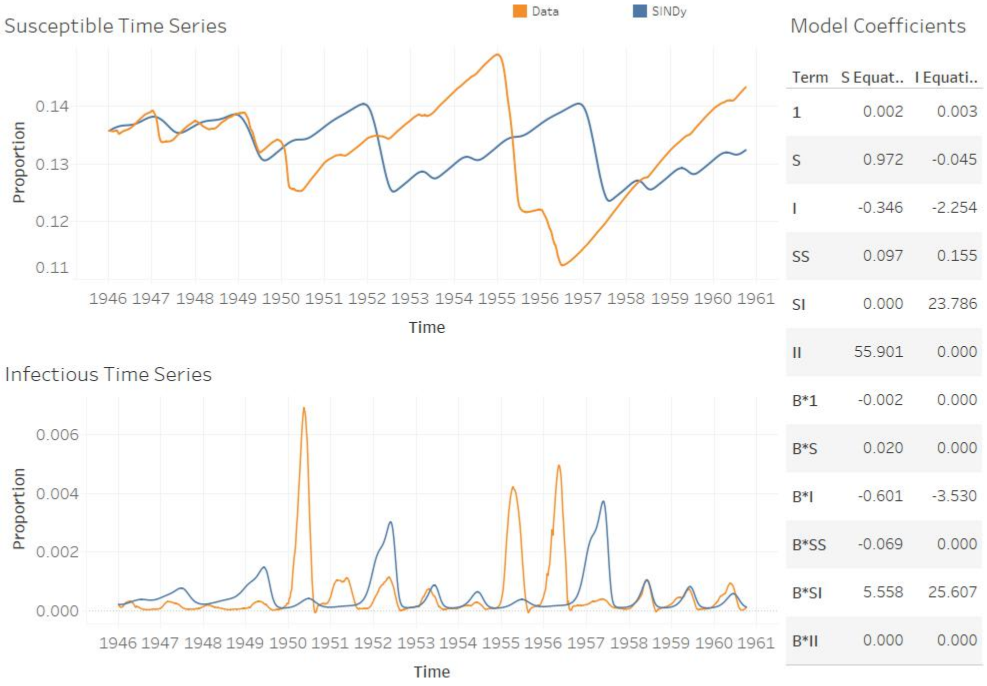


Figure 3.25: Comparison of power spectral density estimates of rubella data and the most parsimonious SINDy-discovered model. The peaks corresponding to the most notable attractors present within the data (with periods of 1 year and 5 years) are noted with the dashed lines.



Term	S Equat..	I Equati..
1	0.002	0.003
S	0.972	-0.045
I	-0.346	-2.254
SS	0.097	0.155
SI	0.000	23.786
II	55.901	0.000
B*1	-0.002	0.000
B*S	0.020	0.000
B*I	-0.601	-3.530
B*SS	-0.069	0.000
B*SI	5.558	25.607
B*II	0.000	0.000

Figure 3.26: Comparison between rubella data and a selected parsimonious SINDy-discovered model using a function library of polynomials up to 2nd order and spectral density estimates for model selection.

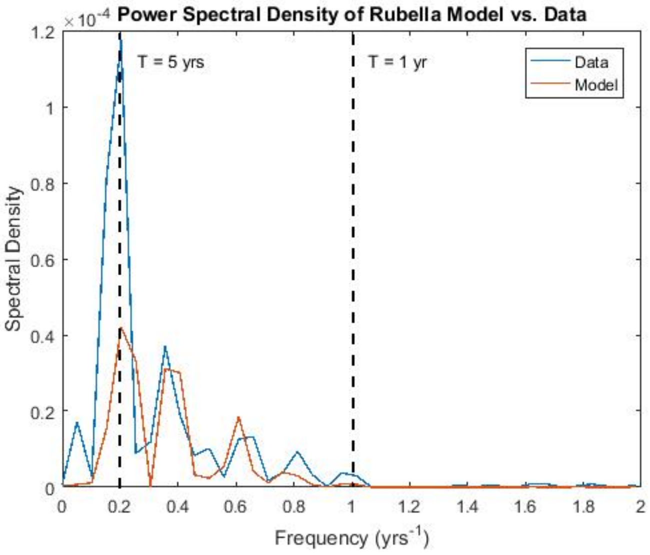


Figure 3.27: Comparison of power spectral density estimates of rubella data and a selected SINDy-discovered model. The peaks corresponding to the most notable attractors present within the data (with periods of 1 year and 5 years) are noted with the dashed lines.

Chapter 4

Discussion and Future Work

4.1 Conclusions

Applying the SINDy algorithm to epidemiological data yielded models that mimicked the dynamic behaviour of the data while still exhibiting sparsity through regularization. The models discovered from both the measles and varicella datasets successfully recovered the attractor class present in the underlying dynamics of the system using a polynomial library of no more than 2nd order, whereas the model discovered from the rubella dataset required a polynomial library of no more than third order to replicate the multiennial cycle. In the case of measles and varicella, use of the AIC metric was also used to find a parsimonious model that correctly identified the phase of the seasonal transmission rate.

These models also compare favourably to their theoretical counterpart, the seasonally-forced SIR model with demographics. The discovered models all demonstrated significant dependence on the mass action incidence terms (SI and βSI), which provide the driving mechanism behind current compartmental models of epidemiological systems. While certain features of the discovered model remain difficult to reconcile with theoretical modelling, these similarities indicate the potential of sparse modelling techniques to inductively verify the current modelling practices in mathematical epidemiology.

4.2 Limitations

One notable limitation when working with epidemiological systems is the availability and quality of data. Since the introduction of vaccination in the 1970s the temporal behaviour of endemic infections changed dramatically [18, 22, 85, 86] and data from the two eras are thus inconsistent with each other. In addition, as noted in Refs. [18, 21, 74, 75], case reporting is inefficient and inconsistent, leading to several necessary assumptions to simplify the process of susceptible reconstruction. Given that empirical susceptible data is not available on the necessary scale, the results of this reconstruction are difficult to verify other than with a heuristic comparison to compartmental models. As this reconstruction comprises one of the two state variables of the system, the nature of the reconstruction will strongly affect the resulting statistical model.

The results generated by applying SINDy to simulated disease data (Section 3.1) demonstrate that another limitation when applying this technique is sensitivity to noisy systems. In Figures 3.1 - 3.8 it is shown that increasing the level of additive noise in the system reduces the ability of the algorithm to accurately recover the correct sparse basis, instead including terms that are not present in the original model. While these figures also show that this increase in noise does not drastically impact the resulting time series, these erroneous coefficients detract from the sparsity of the model and will impact its use. We hypothesize that noise is also the cause of unexpected terms in models discovered from empirical data. Given this sensitivity, the results in Section 3.2.4 are more encouraging as the high level of noise in disease data makes it surprising that SINDy was able to recover features that are expected in a compartmental disease model.

Another limitation, cited by Ref. [50] as the largest challenge when using the SINDy approach, is the proper choice of functional basis, and certainly this issue exists when identifying epidemiological systems. The choice of time-varying transmission function and polynomial order, as well as the inclusion of non-constant demographic variables and other necessary features may be limiting the discovery of a more parsimonious and interpretable model. The state variables themselves are also defined to facilite theoretical mechanisms and may not be the appropriate choice to recover the dynamics of the data.

4.3 Future Work

There still remains much room to develop further techniques that assist in applying sparse identification methods to epidemiological data. A more exhaustive literature review of current disease modelling practices would aid in determining a functional basis that could successfully capture a sparse model using subset selection methods. There exist many modern adaptations on compartmental modelling of infectious diseases [87, 88] which incorporate functions that extend beyond a simple polynomial basis constructed from state variables. Specifically, it is noted in Section 3.2.2 that there is a strong dependance on special cases of a general nonlinear incidence (a transmission function of the form $S^p I^q$, where $p, q > 0$). Extending the function library to include non-integer values for p and q may allow a more accurate representation of the transmission mechanism and result in a more parsimonious discovered model. In addition, the choice of a sinusoidal function for the transmission rate may not be appropriate [11] and while a term-time alternative was tested in our research, further analysis in this area is necessary.

Considering a coupled behaviour-disease system, one that incorporates human decision to vaccinate by including vaccination proportion as a state variable as well as infection dynamics, could also be of interest. Given the influence of social pressure on vaccination rates, the complex interplay between human behaviour and disease prevalence can lead to interesting dynamics that are relevant for current governmental decisions. Models are currently being developed that capture this interplay [89, 90], an excellent review is provided by [91]. Given the availability of vaccination rate data, sparse identification methods could assist the model discovery process of these coupled behaviour-disease models by providing a data-driven approach.

Bibliography

[1] John Graunt. *Natural and Political Observations Made upon the Bills of Mortality.* 1661.

[2] Daniel Bernoulli. "Essai d'une nouvelle analyse de la mortalite causee par la petite verole." *Mem. Math. Phys. Acad. Roy. Sci., Paris* (1766).

[3] William Hamer. *Epidemic Diseases in England: The Evidence of Variability and Persistency of Type.* 1906.

[4] Hans Heesterbeek. *Ecological paradigms lost: routes of theory change.* 2005.

[5] William Kermack and Anderson McKendrick. "A contribution to the mathematical theory of epidemics". *Proc. R. Soc. Lond. A* 115.772 (1927).

[6] Fred Brauer. "Compartmental models in epidemiology". In: *Mathematical epidemiology.* Springer, 2008, pp. 19–79.

[7] Leah Edelstein-Keshet. *Mathematical models in biology.* Vol. 46. Siam, 1988.

[8] Linda J. S. Allen et al. *Mathematical epidemiology.* Vol. 1945. Springer, 2008.

[9] Péter Érdi and János Tóth. *Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models.* Manchester University Press, 1989.

[10] Wayne P. London and James A. Yorke. "Recurrent Outbreaks of Measles, Chickenpox, and Mumps". *American Journal of Epidemiology* 98.6 (1978).

[11] David J. D. Earn et al. "A Simple Model for Complex Dynamical Transitions in Epidemics". *Science* 287.667 (2000).

[12] Ben M. Bolker and Bryan T. Grenfell. "Chaos and biological complexity in measles dynamics". *Proceedings of the Royal Society of London B: Biological Sciences* 251.1330 (1993), pp. 75–81.

[13] Chris T. Bauch and David J. D. Earn. “Transients and attractors in epidemics”. *Proc. R. Soc. Lond. B* 270 (2003), pp. 1573–1578.

[14] Herbert W. Hethcote. “Asymptotic behavior and stability in epidemic models”. In: *Mathematical Problems in Biology.* Springer, 1974, pp. 83–92.

[15] Danilo Diedrichs, Paul Isihara, and Doeke Buursma. “The Schedule Effect: can recurrent peak infections be reduced without vaccines, quarantines or school closings?” (Dec. 2013).

[16] Sonia Altizer et al. “Seasonality and the dynamics of infectious diseases”. *Ecol. Lett.* 9 (2006).

[17] Herbert W. Hethcote and Simon A. Levin. “Periodicity in epidemiological models”. *Biomathematics* 18 (1989).

[18] Paul E. M. Fine and Jacqueline A. Clarkson. “Measles in England and Wales - I: An Analysis of Factors Underlying Seasonal Patterns”. *International Journal of Epidemiology* 11.1 (1982).

[19] Paul E. M. Fine and Jacqueline A. Clarkson. “Measles in England and Wales - II: The Impact of the Measles Vaccination Programme on the Distribution of Immunity in the Population”. *International Journal of Epidemiology* 11.1 (1982).

[20] Georgiy V. Bobashev et al. “Reconstructing Susceptible and Recruitment Dynamics from Measles Epidemic Data”. *Mathematical Population Studies* 8.1 (2000).

[21] Bärbel F. Finkenstädt and Bryan T. Grenfell. “Time series modelling of childhood diseases: a dynamical systems approach”. *Appl. Statist.* 49 (2000), pp. 187–205.

[22] Dieter Schenzle. “An Age-Structured Model of Pre- and Post-Vaccination Measles Transmission”. *Mathematical Medicine and Biology: A Journal of the IMA* 1.2 (1984), pp. 169–191.

[23] Herbert E. Soper. “The Interpretation of Periodicity in Disease Prevalence”. *Journal of the Royal Statistical Society* 92.1 (1929), pp. 34–73.

[24] William M. Schaffer and Mark Kot. “Nearly one dimensional dynamics in an epidemic”. *Journal of Theoretical Biology* 112 (1985), pp. 403–427.

[25] Lars F. Olsen, G. L. Truty, and William M. Schaffer. “Oscillations and chaos in epidemics: A nonlinear dynamic study of six childhood diseases in Copenhagen, Denmark”. *Theoretical Population Biology* 33.3 (1988), pp. 344–370.

[26] Stephen P. Ellner, B. A. Bailey, and Georgiy V. Bobashev. “Noise and Nonlinearity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling”. *The American Naturalist* 151.5 (1998), pp. 425–440.

[27] Aleksandr Mikhailovich Lyapunov. “The general problem of the stability of motion”. *International Journal of Control* 55.3 (1992), pp. 531–534.

[28] Andrei Korobeinikov and Graeme C Wake. “Lyapunov functions and global stability for SIR, SIRS, and SIS epidemiological models”. *Applied Mathematics Letters* 15.8 (2002), pp. 955–960.

[29] Francis C. Moon. “Chaotic vibrations: an introduction for applied scientists and engineers”. *Research supported by NSF, USAF, US Navy, US Army, and IBM. New York, Wiley-Interscience, 1987, 322 p.* (1987).

[30] Steven H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering.* CRC Press, 2018.

[31] Santo Banerjee, Lamberto Rondoni, and Mala Mitra. “Applications of Chaos and Nonlinear Dynamics in Science and Engineering-Vol. 4”. *Applications of Chaos and Nonlinear Dynamics in Science and Engineering* 4 (2015).

[32] Hermann Schichl. “Models and History of Modeling”. In: *Modeling Languages in Mathematical Optimization.* 2004. Chap. 2, pp. 25–39.

[33] Michael I. Jordan and Tom Mitchell. “Machine learning: Trends, perspectives, and prospects”. *Science* 349.6245 (2015), pp. 255–260.

[34] Alawi A. Alsheikh-Ali et al. “Public Availability of Published Research Data in High-Impact Journals”. *PLOS One* 6.9 (Sept. 2011), pp. 1–4.

[35] Saint John Walker. “Big Data: A Revolution That Will Transform How We Live, Work, and Think”. *International Journal of Advertising* 33.1 (2014), pp. 181–183.

[36] David W. Bates et al. “Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients”. *Health Affairs* 33.7 (2014).

[37] Gareth James et al. *An introduction to statistical learning.* New York: Springer, 2013.

[38] Edward N. Lorenz. “Deterministic Nonperiodic Flow”. *Journal of the Atmospheric Sciences* 20 (1963), pp. 130–141.

[39] John Guckenheimer. “From data to dynamical systems”. *Nonlinearity* 27 (2014), pp. 41–50.

[40] Norman Packard et. al. “Geometry from a Time Series”. *Phys. Rev. Lett.* 45 (9 Sept. 1980), pp. 712–716.

[41] Floris Takens. “Detecting strange attractors in turbulence”. In: *Dynamical systems and turbulence, Warwick 1980.* Springer, 1981, pp. 366–381.

[42] Wenxu Wang, Ying-Cheng Lai, and Celso Grebogi. “Data Based Identification and Prediction of Nonlinear and Complex Dynamical Systems”. *Physics Reports* 644 (2016), pp. 1–76.

[43] Josh Bongard and Hod Lipson. “Automated reverse engineering of nonlinear dynamical systems”. *Proceedings of the National Academy of Sciences* 104.24 (2007), pp. 9943–9948.

[44] Michael Schmidt and Hod Lipson. “Distilling free-form natural laws from experimental data”. *science* 324.5923 (2009), pp. 81–85.

[45] John R. Koza. “Genetic programming as a means for programming computers by natural selection”. *Statistics and Computing* 4.2 (June 1994), pp. 87–112.

[46] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288.

[47] Arturo Rosenblueth and Norbert Wiener. “The Role of Models in Science”. *Philosophy of Science* 12.4 (1945), pp. 316–321.

[48] George H John, Ron Kohavi, and Karl Pfleger. "Irrelevant features and the subset selection problem". In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 121–129.

[49] Merlise Clyde, Giovanni Parmigiani, and Brani Vidakovic. "Multiple shrinkage and subset selection in wavelets". *Biometrika* 85.2 (1998), pp. 391–401.

[50] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". *Proceedings of the National Academy of Sciences* 113.15 (2016), pp. 3932–3937.

[51] Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media, 2003.

[52] Gerda Claeskens, Nils Lid Hjort, et al. "Model selection and model averaging". *Cambridge Books* (2008).

[53] Mark Woodward. *Epidemiology: study design and data analysis.* CRC press, 2013.

[54] Hirotogu Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Breakthroughs in statistics.* Springer, 1992, pp. 610–624.

[55] Niall M. Mangan et al. "Model selection for dynamical systems via sparse regression and information criteria". *Proc. R. Soc. A* 473.2204 (2017), p. 20170009.

[56] Samuel H. Rudy et al. "Data-driven discovery of partial differential equations". *Science Advances* 3.4 (2017), e1602614.

[57] Giang Tran and Rachel Ward. "Exact recovery of chaotic systems from highly corrupted data". *Multiscale Modeling & Simulation* 15.3 (2017), pp. 1108–1129.

[58] Eurika Kaiser, J. Nathan Kutz, and Steven L. Brunton. "Sparse identification of nonlinear dynamics for model predictive control in the low-data limit". *arXiv preprint arXiv:1711.05501* (2017).

[59] Yosef El Sayed M., Richard Semaan, and Rolf Radespiel. “Sparse Modeling of the Lift Gains of a High-Lift Configuration with Periodic Coanda Blowing”. In: *2018 AIAA Aerospace Sciences Meeting*. 2018, p. 1054.

[60] Magnus Dam. “Topological bifurcations of coherent structures and dimension reduction of plasma convection models”. PhD thesis. DTU Compute, 2018.

[61] Niall M. Mangan et al. “Inferring biological networks by sparse identification of nonlinear dynamics”. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* 2.1 (2016), pp. 52–63.

[62] Markus Quade et al. “Sparse identification of nonlinear dynamics for rapid model recovery”. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.6 (2018), p. 063116.

[63] Roy M. Anderson and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1992.

[64] Amit Huppert and G. Katriel. “Mathematical modelling and prediction in infectious disease epidemiology”. *Clinical Microbiology and Infection* 19.11 (2013), pp. 999–1005.

[65] Fred Brauer. “Mathematical epidemiology: Past, present, and future”. *Infectious Disease Modelling* 2.2 (2017), pp. 113–127.

[66] Jeff Bartlett, James Devinney, and Eric Pudlowski. “Mathematical modeling of the 2014/2015 Ebola epidemic in West Africa”. *SIAM Undergraduate Research Online* 9 (2016), pp. 87–102.

[67] David L Donoho. “Compressed sensing”. *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[68] Emmanuel J. Candès and Michael B. Wakin. “An introduction to compressive sampling”. *IEEE signal processing magazine* 25.2 (2008), pp. 21–30.

[69] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. “Compressive sampling and dynamic mode decomposition”. *arXiv preprint arXiv:1312.5186* (2013).

[70] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms". *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.

[71] Rick Chartrand. "Numerical differentiation of noisy, nonsmooth data". *ISRN Applied Mathematics* 2011 (2011).

[72] Steve Brunton et al. URL: `faculty.washington.edu/sbrunton/sparsedynamics.zip`.

[73] Jonathan H. Horrocks and Steve Brunton. URL: `https://github.com/jonathanhorrocks/SINDy-data`.

[74] Susan F. Davis et. al. "Reporting efficiency during a measles outbreak in New York City, 1991." *American journal of public health* 83.7 (1993), pp. 1011–1015.

[75] Timothy J. Doyle, M. Kathleen Glynn, and Samuel L. Groseclose. "Completeness of notifiable infectious disease reporting in the United States: an analytical literature review". *American journal of epidemiology* 155.9 (2002), pp. 866–874.

[76] Ben Bolker. *Infectious disease data.* URL: `https://ms.mcmaster.ca/~bolker/measdata.html`.

[77] GB Historical GIS / University of Portsmouth. *England Dep through time.* URL: `http://www.visionofbritain.org.uk/unit/10061325/cube/TOT_POP`.

[78] *200 years of the Census in Wales.* URL: `https://web.archive.org/web/20090319202324/http://www.statistics.gov.uk/census2001/bicentenary/pdfs/wales.pdf`.

[79] Statistics Canada. URL: `https://www150.statcan.gc.ca/cansim/results/cansim-0530001-eng-2134590597138961162.csv`.

[80] Statistics Canada. URL: `https://www150.statcan.gc.ca/n1/pub/11-516-x/sectiona/4147436-eng.htm#1`.

[81] Maurice Bertram Priestley. "Spectral analysis and time series" (1981).

[82] *Periodogram power spectral density estimate.* URL: `https://www.mathworks.com/help/signal/ref/periodogram.html`.

[83] Wei-min Liu, Herbert W. Hethcote, and Simon A. Levin. "Dynamical behavior of epidemiological models with nonlinear incidence rates". *Journal of Mathematical Biology* 25.4 (1987), pp. 359–380.

[84] Andrei Korobeinikov and Philip K. Maini. "A Lyapunov function and global properties for SIR and SEIR epidemiological models with nonlinear incidence". *Mathematical Biosciences and Engineering* 1.1 (2004), pp. 57–60.

[85] Roy M. Anderson, Brian Grenfell, and Robert M. May. "Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis". *Epidemiology & Infection* 93.3 (1984), pp. 587–608.

[86] Hélène Broutin et al. "Epidemiological impact of vaccination on the dynamics of two childhood diseases in rural Senegal". *Microbes and Infection* 7.4 (2005), pp. 593–599.

[87] Junkichi Satsuma et al. "Extending the SIR epidemic model". *Physica A: Statistical Mechanics and its Applications* 336.3-4 (2004), pp. 369–375.

[88] Connell McCluskey. "Complete global stability for an SIR epidemic model with delay—distributed or discrete". *Nonlinear Analysis: Real World Applications* 11.1 (2010), pp. 55–59.

[89] Chris T. Bauch. "Imitation dynamics predict vaccinating behaviour". *Proceedings of the Royal Society of London B: Biological Sciences* 272.1573 (2005), pp. 1669–1675.

[90] Tamer Oraby, Vivek Thampi, and Chris T. Bauch. "The influence of social norms on the dynamics of vaccinating behaviour for paediatric infectious diseases". *Proc. R. Soc. B* 281.1780 (2014), p. 20133172.

[91] Zhen Wang et al. "Coupled disease–behavior dynamics on complex networks: A review". *Physics of life reviews* 15 (2015), pp. 1–29.

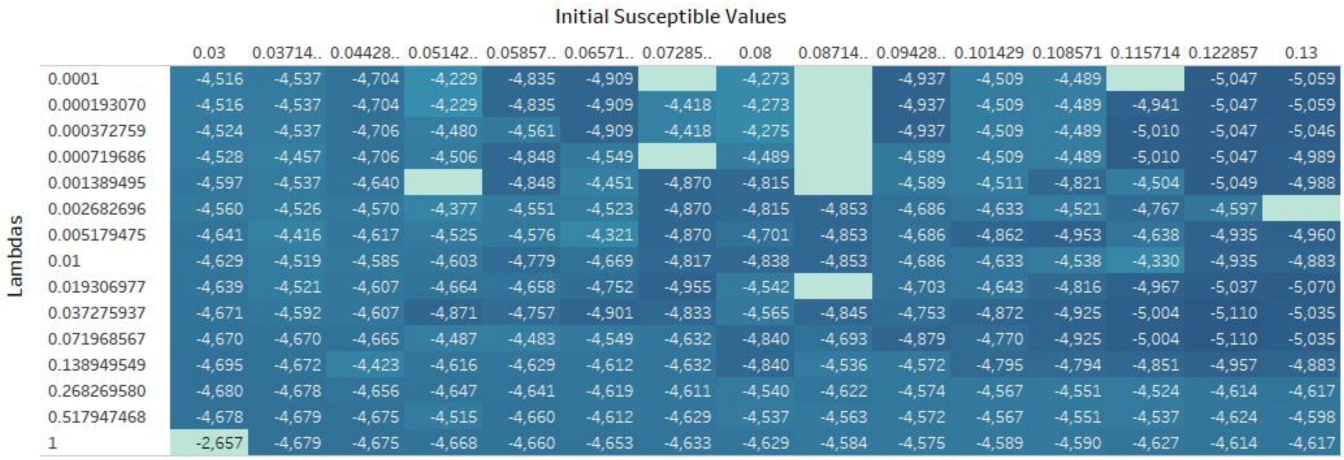
Appendices

Appendix A

Supplementary Figures

A.1 Varying Sparsity Threshold

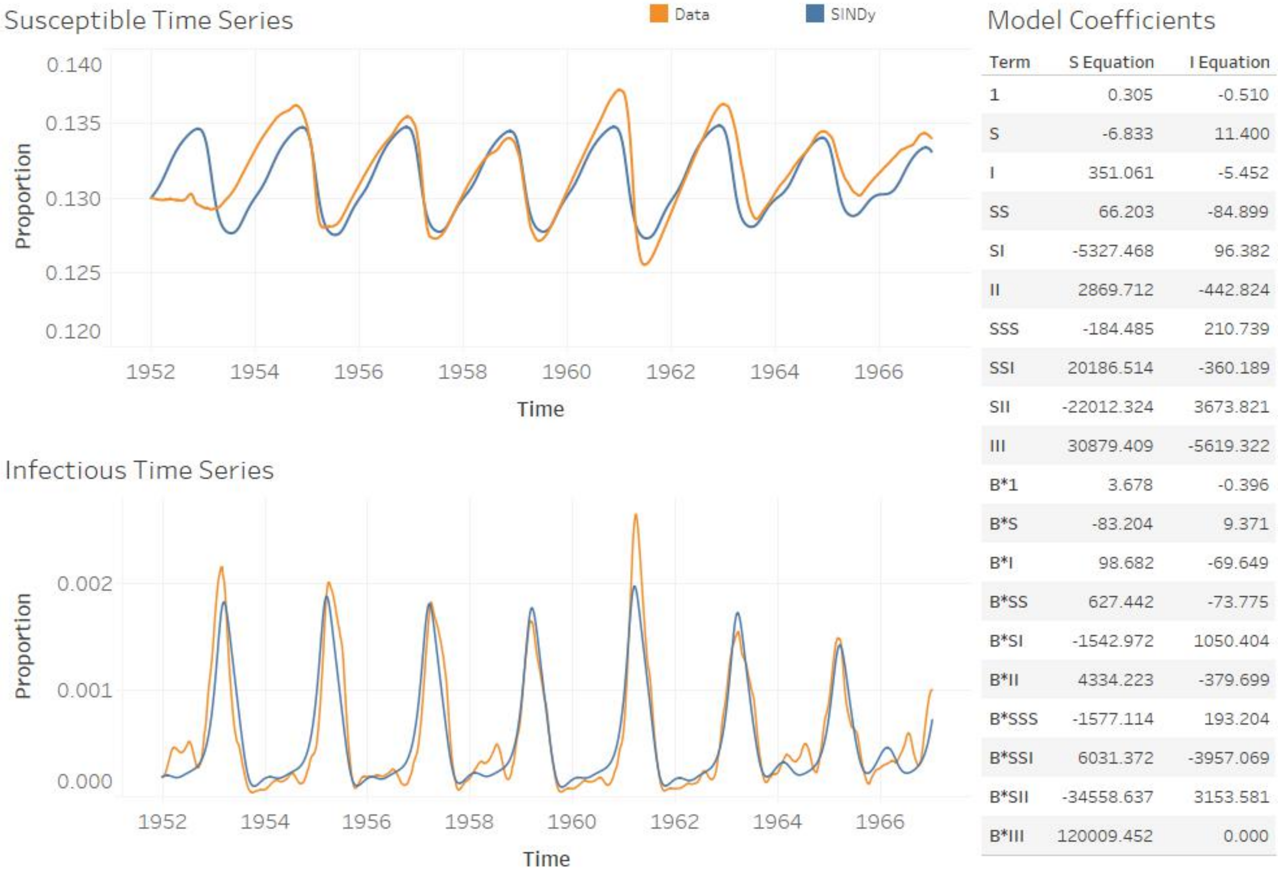
Measles (Third Order Library), AIC Values



	Initial Susceptible Values														
Lambdas	0.03	0.03714..	0.04428..	0.05142..	0.05857..	0.06571..	0.07285..	0.08	0.08714..	0.09428..	0.101429	0.108571	0.115714	0.122857	0.13
0.0001	-4,516	-4,537	-4,704	-4,229	-4,835	-4,909		-4,273		-4,937	-4,509	-4,489		-5,047	-5,059
0.000193070	-4,516	-4,537	-4,704	-4,229	-4,835	-4,909	-4,418	-4,273		-4,937	-4,509	-4,489	-4,941	-5,047	-5,059
0.000372759	-4,524	-4,537	-4,706	-4,480	-4,561	-4,909	-4,418	-4,275		-4,937	-4,509	-4,489	-5,010	-5,047	-5,046
0.000719686	-4,528	-4,457	-4,706	-4,506	-4,848	-4,549		-4,489		-4,589	-4,509	-4,489	-5,010	-5,047	-4,989
0.001389495	-4,597	-4,537	-4,640		-4,848	-4,451	-4,870	-4,815		-4,589	-4,511	-4,821	-4,504	-5,049	-4,988
0.002682696	-4,560	-4,526	-4,570	-4,377	-4,551	-4,523	-4,870	-4,815	-4,853	-4,686	-4,633	-4,521	-4,767	-4,597	
0.005179475	-4,641	-4,416	-4,617	-4,525	-4,576	-4,321	-4,870	-4,701	-4,853	-4,686	-4,862	-4,953	-4,638	-4,935	-4,960
0.01	-4,629	-4,519	-4,585	-4,603	-4,779	-4,669	-4,817	-4,838	-4,853	-4,686	-4,633	-4,538	-4,330	-4,935	-4,883
0.019306977	-4,639	-4,521	-4,607	-4,664	-4,658	-4,752	-4,955	-4,542		-4,703	-4,643	-4,816	-4,967	-5,037	-5,070
0.037275937	-4,671	-4,592	-4,607	-4,871	-4,757	-4,901	-4,833	-4,565	-4,845	-4,753	-4,872	-4,925	-5,004	-5,110	-5,035
0.071968567	-4,670	-4,670	-4,665	-4,487	-4,483	-4,549	-4,632	-4,840	-4,693	-4,879	-4,770	-4,925	-5,004	-5,110	-5,035
0.138949549	-4,695	-4,672	-4,423	-4,616	-4,629	-4,612	-4,632	-4,840	-4,536	-4,572	-4,795	-4,794	-4,851	-4,957	-4,883
0.268269580	-4,680	-4,678	-4,656	-4,647	-4,641	-4,619	-4,611	-4,540	-4,622	-4,574	-4,567	-4,551	-4,524	-4,614	-4,617
0.517947468	-4,678	-4,679	-4,675	-4,515	-4,660	-4,612	-4,629	-4,537	-4,563	-4,572	-4,567	-4,551	-4,537	-4,624	-4,598
1	-2,657	-4,679	-4,675	-4,668	-4,660	-4,653	-4,633	-4,629	-4,584	-4,575	-4,589	-4,590	-4,627	-4,614	-4,617

Figure A.1: The AIC values for SINDy models across a range of both initial susceptible and threshold values, utilizing the measles dataset and a 3rd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model. Cells without value refer to a model which, when simulated, resulted in a diverging time series.

Measles (Third Order Library), Low Threshold

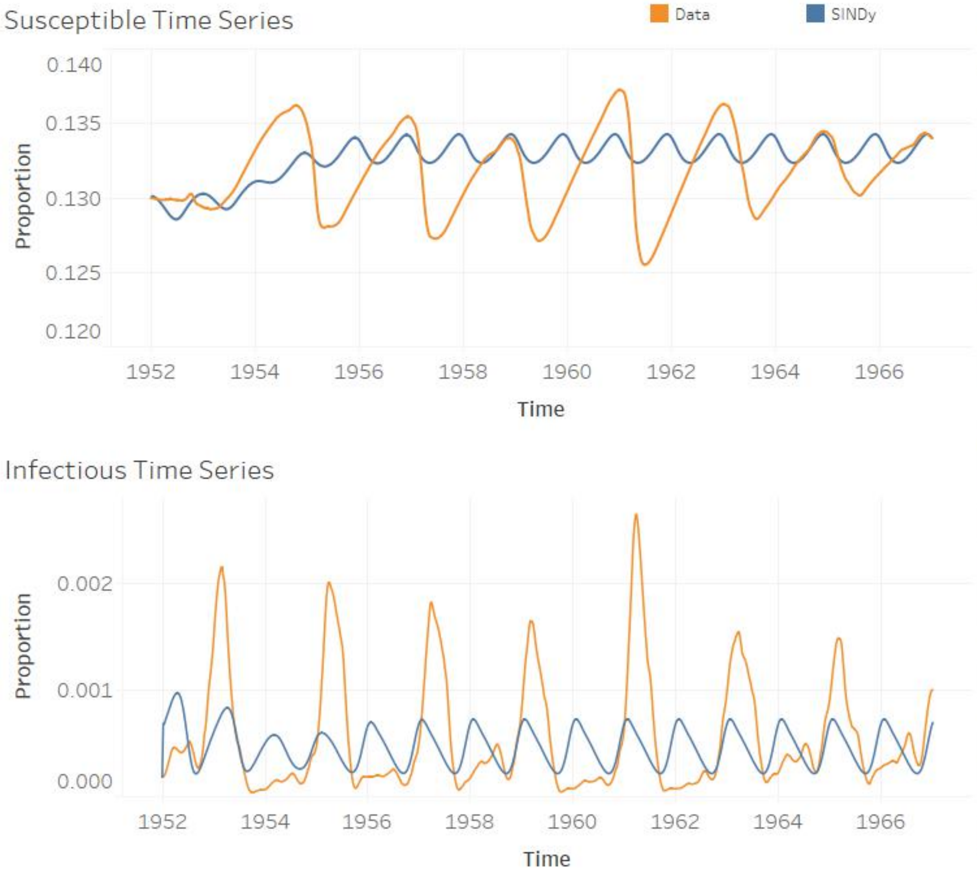


Model Coefficients

Term	S Equation	I Equation
1	0.305	-0.510
S	-6.833	11.400
I	351.061	-5.452
SS	66.203	-84.899
SI	-5327.468	96.382
II	2869.712	-442.824
SSS	-184.485	210.739
SSI	20186.514	-360.189
SII	-22012.324	3673.821
III	30879.409	-5619.322
B*1	3.678	-0.396
B*S	-83.204	9.371
B*I	98.682	-69.649
B*SS	627.442	-73.775
B*SI	-1542.972	1050.404
B*II	4334.223	-379.699
B*SSS	-1577.114	193.204
B*SSI	6031.372	-3957.069
B*SII	-34558.637	3153.581
B*III	120009.452	0.000

Figure A.2: Resulting time series and coefficients from SINDy-discovered models of the measles dataset (with a 3rd order polynomial library) using a relatively low threshold ($\lambda = 0.0001$). The resulting model exhibits a good fit and accurate recovery of attractor class, but a very low number of non-active terms, which is an indicator of an overfit model.

Measles (Third Order Library), High Threshold

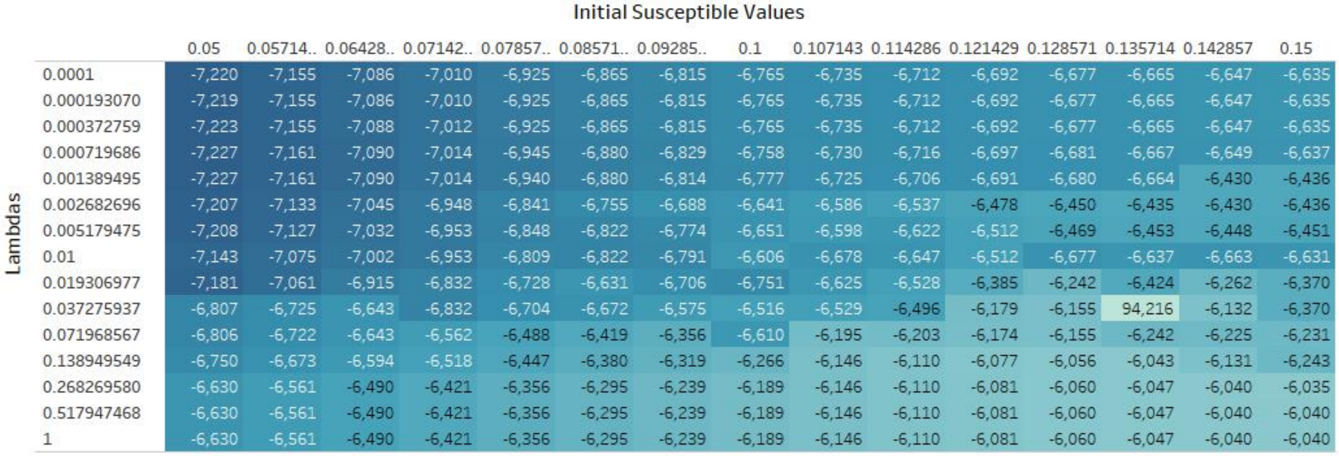


Model Coefficients

Term	S Equation	I Equation
1	-2.199	9.658
S	50.156	-216.002
I	213.931	0.000
SS	-366.080	1609.843
SI	-3249.802	0.000
II	0.000	0.000
SSS	908.250	-3998.005
SSI	12323.463	0.000
SII	0.000	0.000
III	0.000	0.000
B*1	3.136	-7.118
B*S	-70.539	156.556
B*I	-93.657	0.000
B*SS	528.734	-1146.285
B*SI	1393.328	0.000
B*II	0.000	0.000
B*SSS	-1320.754	2793.806
B*SSI	-5174.140	0.000
B*SII	0.000	0.000
B*III	0.000	0.000

Figure A.3: Resulting time series and coefficients from SINDy-discovered models of the measles dataset (with a 3rd order polynomial library) using a relatively high threshold ($\lambda = 1$). The resulting model exhibits sparsity, but at the cost of a good fit and recovery of attractor class.

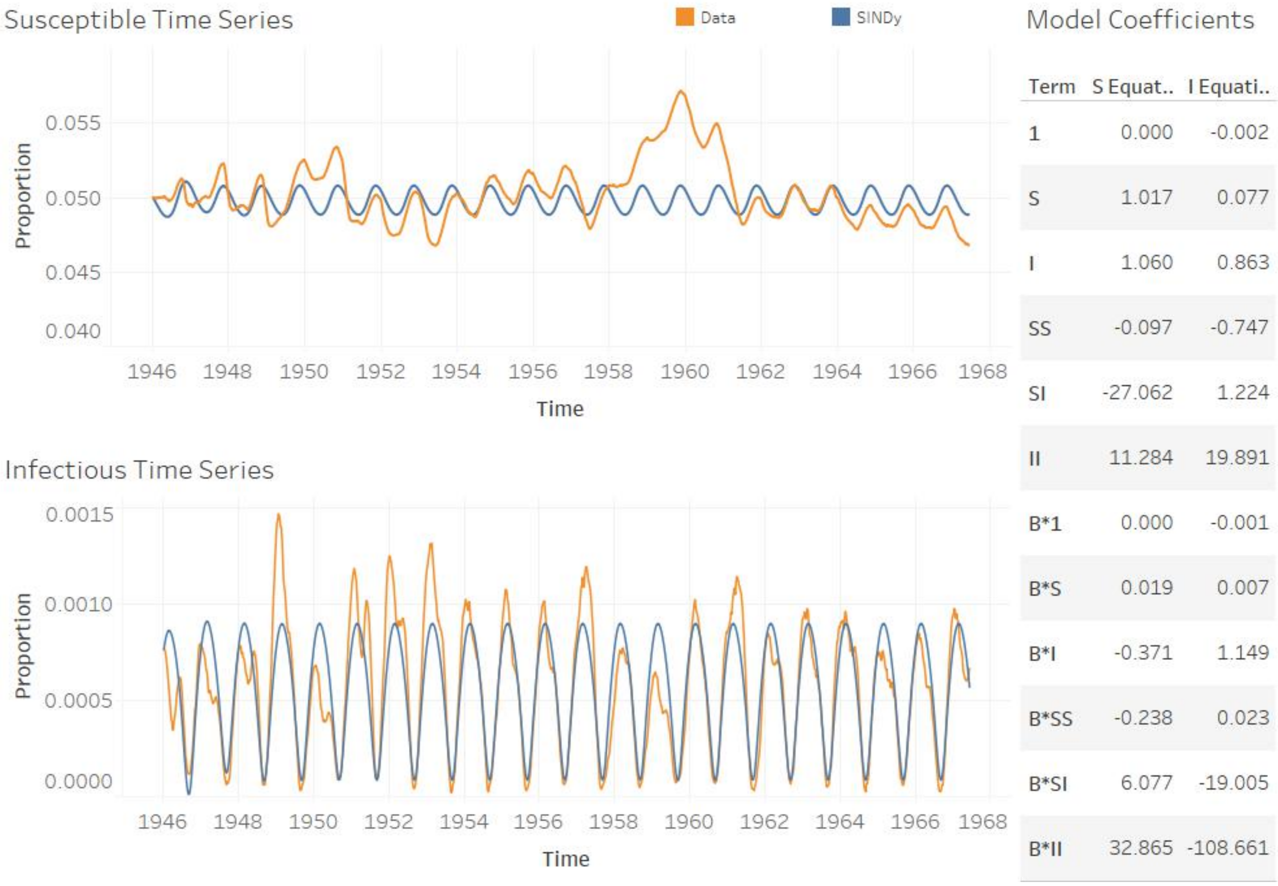
Varicella (Second Order Library), AIC Values



	Initial Susceptible Values														
Lambdas	0.05	0.05714..	0.06428..	0.07142..	0.07857..	0.08571..	0.09285..	0.1	0.107143	0.114286	0.121429	0.128571	0.135714	0.142857	0.15
0.0001	-7,220	-7,155	-7,086	-7,010	-6,925	-6,865	-6,815	-6,765	-6,735	-6,712	-6,692	-6,677	-6,665	-6,647	-6,635
0.000193070	-7,219	-7,155	-7,086	-7,010	-6,925	-6,865	-6,815	-6,765	-6,735	-6,712	-6,692	-6,677	-6,665	-6,647	-6,635
0.000372759	-7,223	-7,155	-7,088	-7,012	-6,925	-6,865	-6,815	-6,765	-6,735	-6,712	-6,692	-6,677	-6,665	-6,647	-6,635
0.000719686	-7,227	-7,161	-7,090	-7,014	-6,945	-6,880	-6,829	-6,758	-6,730	-6,716	-6,697	-6,681	-6,667	-6,649	-6,637
0.001389495	-7,227	-7,161	-7,090	-7,014	-6,940	-6,880	-6,814	-6,777	-6,725	-6,706	-6,691	-6,680	-6,664	-6,430	-6,436
0.002682696	-7,207	-7,133	-7,045	-6,948	-6,841	-6,755	-6,688	-6,641	-6,586	-6,537	-6,478	-6,450	-6,435	-6,430	-6,436
0.005179475	-7,208	-7,127	-7,032	-6,953	-6,848	-6,822	-6,774	-6,651	-6,598	-6,622	-6,512	-6,469	-6,453	-6,448	-6,451
0.01	-7,143	-7,075	-7,002	-6,953	-6,809	-6,822	-6,791	-6,606	-6,678	-6,647	-6,512	-6,677	-6,637	-6,663	-6,631
0.019306977	-7,181	-7,061	-6,915	-6,832	-6,728	-6,631	-6,706	-6,751	-6,625	-6,528	-6,385	-6,242	-6,424	-6,262	-6,370
0.037275937	-6,807	-6,725	-6,643	-6,832	-6,704	-6,672	-6,575	-6,516	-6,529	-6,496	-6,179	-6,155	94,216	-6,132	-6,370
0.071968567	-6,806	-6,722	-6,643	-6,562	-6,488	-6,419	-6,356	-6,610	-6,195	-6,203	-6,174	-6,155	-6,242	-6,225	-6,231
0.138949549	-6,750	-6,673	-6,594	-6,518	-6,447	-6,380	-6,319	-6,266	-6,146	-6,110	-6,077	-6,056	-6,043	-6,131	-6,243
0.268269580	-6,630	-6,561	-6,490	-6,421	-6,356	-6,295	-6,239	-6,189	-6,146	-6,110	-6,081	-6,060	-6,047	-6,040	-6,035
0.517947468	-6,630	-6,561	-6,490	-6,421	-6,356	-6,295	-6,239	-6,189	-6,146	-6,110	-6,081	-6,060	-6,047	-6,040	-6,040
1	-6,630	-6,561	-6,490	-6,421	-6,356	-6,295	-6,239	-6,189	-6,146	-6,110	-6,081	-6,060	-6,047	-6,040	-6,040

Figure A.4: The AIC values for SINDy models across a range of both initial susceptible and threshold values, utilizing the varicella dataset and a 2nd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model. Cells without value refer to a model which, when simulated, resulted in a diverging time series.

Varicella (Second Order Library), Low Threshold

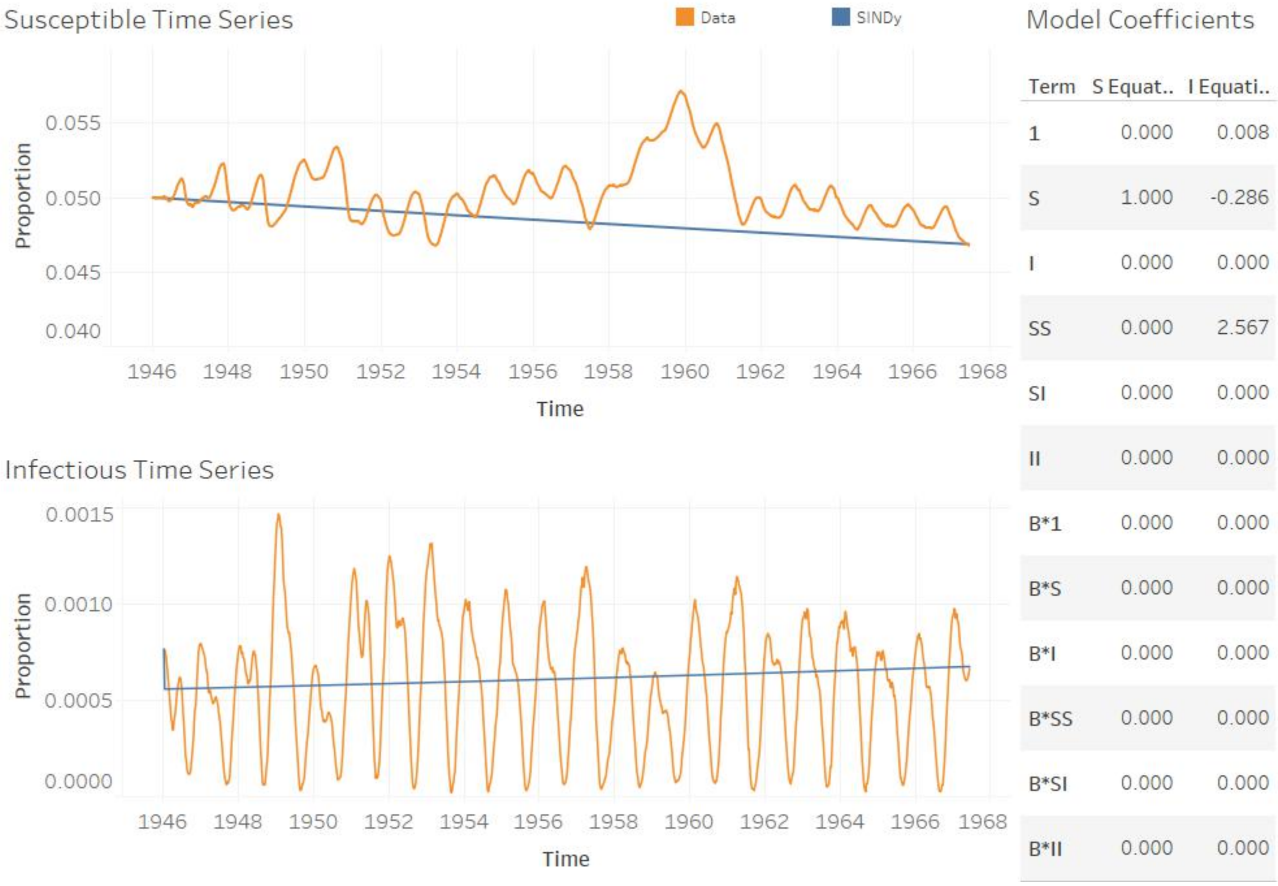


Model Coefficients

Term	S Equat..	I Equati..
1	0.000	-0.002
S	1.017	0.077
I	1.060	0.863
SS	-0.097	-0.747
SI	-27.062	1.224
II	11.284	19.891
B*1	0.000	-0.001
B*S	0.019	0.007
B*I	-0.371	1.149
B*SS	-0.238	0.023
B*SI	6.077	-19.005
B*II	32.865	-108.661

Figure A.5: Resulting time series and coefficients from SINDy-discovered models of the varicella dataset (with a 2nd order polynomial library) using a relatively low threshold ($\lambda = 0.0001$). The resulting model exhibits a good fit and accurate recovery of attractor class, but a very low number of non-active terms, which is an indicator of an overfit model.

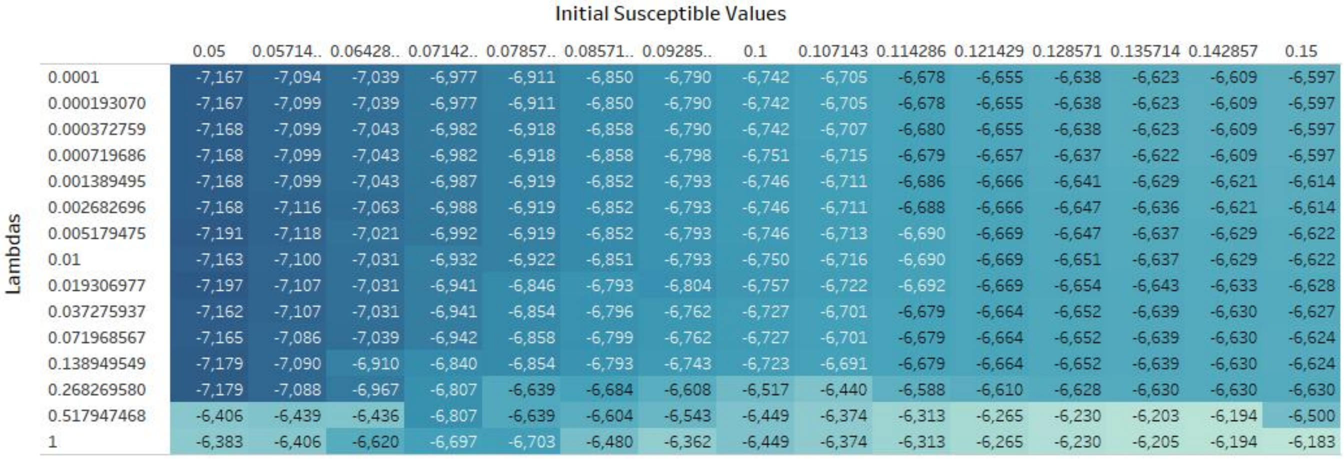
Varicella (Second Order Library), High Threshold



Term	S Equat..	I Equati..
1	0.000	0.008
S	1.000	-0.286
I	0.000	0.000
SS	0.000	2.567
SI	0.000	0.000
II	0.000	0.000
B*1	0.000	0.000
B*S	0.000	0.000
B*I	0.000	0.000
B*SS	0.000	0.000
B*SI	0.000	0.000
B*II	0.000	0.000

Figure A.6: Resulting time series and coefficients from SINDy-discovered models of the varicella dataset (with a 2nd order polynomial library) using a relatively high threshold ($\lambda = 1$). The resulting model exhibits a high level of sparsity, but the linear fit does not accurately represent the dynamics of the system whatsoever.

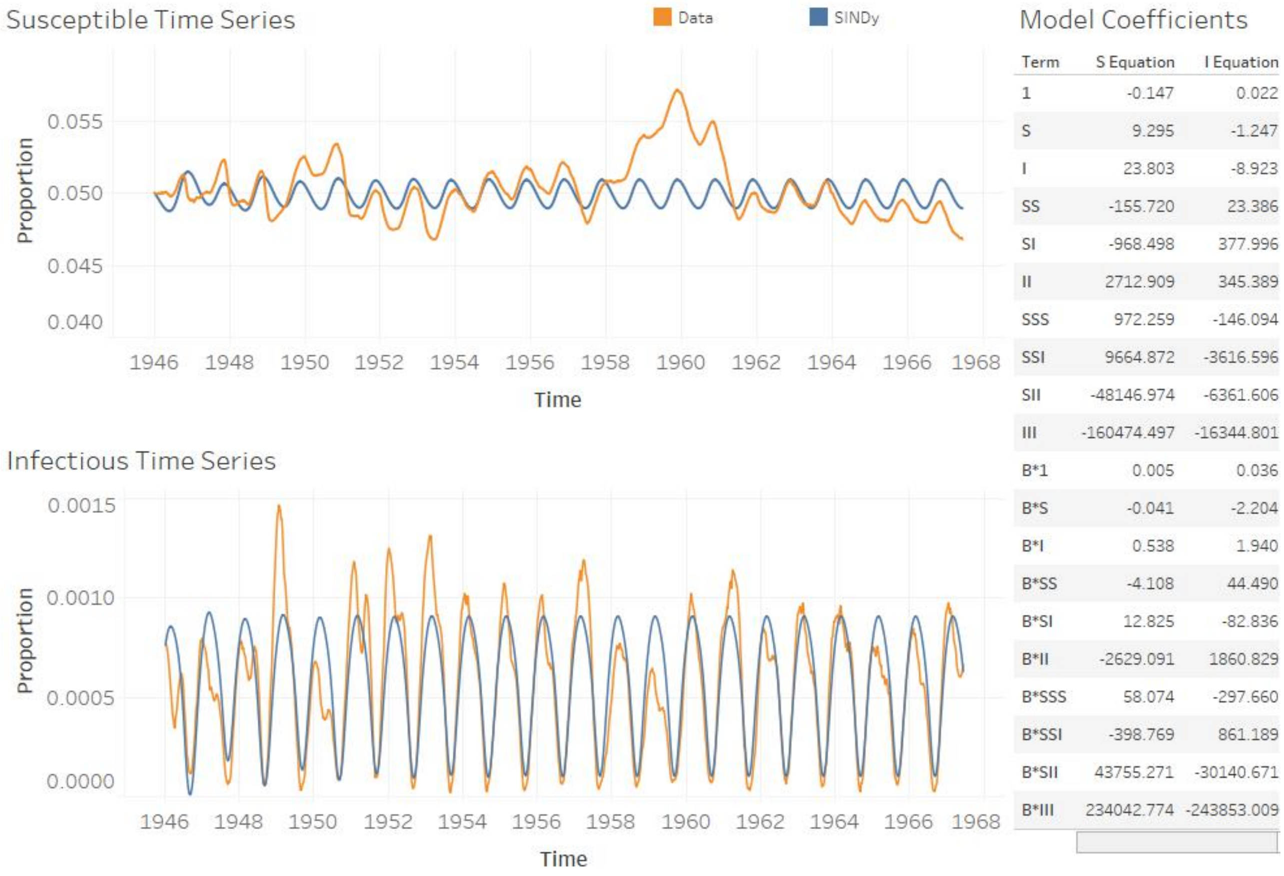
Varicella (Third Order Library), AIC Values



	Initial Susceptible Values														
Lambdas	0.05	0.05714..	0.06428..	0.07142..	0.07857..	0.08571..	0.09285..	0.1	0.107143	0.114286	0.121429	0.128571	0.135714	0.142857	0.15
0.0001	-7,167	-7,094	-7,039	-6,977	-6,911	-6,850	-6,790	-6,742	-6,705	-6,678	-6,655	-6,638	-6,623	-6,609	-6,597
0.000193070	-7,167	-7,099	-7,039	-6,977	-6,911	-6,850	-6,790	-6,742	-6,705	-6,678	-6,655	-6,638	-6,623	-6,609	-6,597
0.000372759	-7,168	-7,099	-7,043	-6,982	-6,918	-6,858	-6,790	-6,742	-6,707	-6,680	-6,655	-6,638	-6,623	-6,609	-6,597
0.000719686	-7,168	-7,099	-7,043	-6,982	-6,918	-6,858	-6,798	-6,751	-6,715	-6,679	-6,657	-6,637	-6,622	-6,609	-6,597
0.001389495	-7,168	-7,099	-7,043	-6,987	-6,919	-6,852	-6,793	-6,746	-6,711	-6,686	-6,666	-6,641	-6,629	-6,621	-6,614
0.002682696	-7,168	-7,116	-7,063	-6,988	-6,919	-6,852	-6,793	-6,746	-6,711	-6,688	-6,666	-6,647	-6,636	-6,621	-6,614
0.005179475	-7,191	-7,118	-7,021	-6,992	-6,919	-6,852	-6,793	-6,746	-6,713	-6,690	-6,669	-6,647	-6,637	-6,629	-6,622
0.01	-7,163	-7,100	-7,031	-6,932	-6,922	-6,851	-6,793	-6,750	-6,716	-6,690	-6,669	-6,651	-6,637	-6,629	-6,622
0.019306977	-7,197	-7,107	-7,031	-6,941	-6,846	-6,793	-6,804	-6,757	-6,722	-6,692	-6,669	-6,654	-6,643	-6,633	-6,628
0.037275937	-7,162	-7,107	-7,031	-6,941	-6,854	-6,796	-6,762	-6,727	-6,701	-6,679	-6,664	-6,652	-6,639	-6,630	-6,627
0.071968567	-7,165	-7,086	-7,039	-6,942	-6,858	-6,799	-6,762	-6,727	-6,701	-6,679	-6,664	-6,652	-6,639	-6,630	-6,624
0.138949549	-7,179	-7,090	-6,910	-6,840	-6,854	-6,793	-6,743	-6,723	-6,691	-6,679	-6,664	-6,652	-6,639	-6,630	-6,624
0.268269580	-7,179	-7,088	-6,967	-6,807	-6,639	-6,684	-6,608	-6,517	-6,440	-6,588	-6,610	-6,628	-6,630	-6,630	-6,630
0.517947468	-6,406	-6,439	-6,436	-6,807	-6,639	-6,604	-6,543	-6,449	-6,374	-6,313	-6,265	-6,230	-6,203	-6,194	-6,500
1	-6,383	-6,406	-6,620	-6,697	-6,703	-6,480	-6,362	-6,449	-6,374	-6,313	-6,265	-6,230	-6,205	-6,194	-6,183

Figure A.7: The AIC values for SINDy models across a range of both initial susceptible and threshold values, utilizing the varicella dataset and a 3rd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model. Cells without value refer to a model which, when simulated, resulted in a diverging time series.

Varicella (Third Order Library), Low Threshold

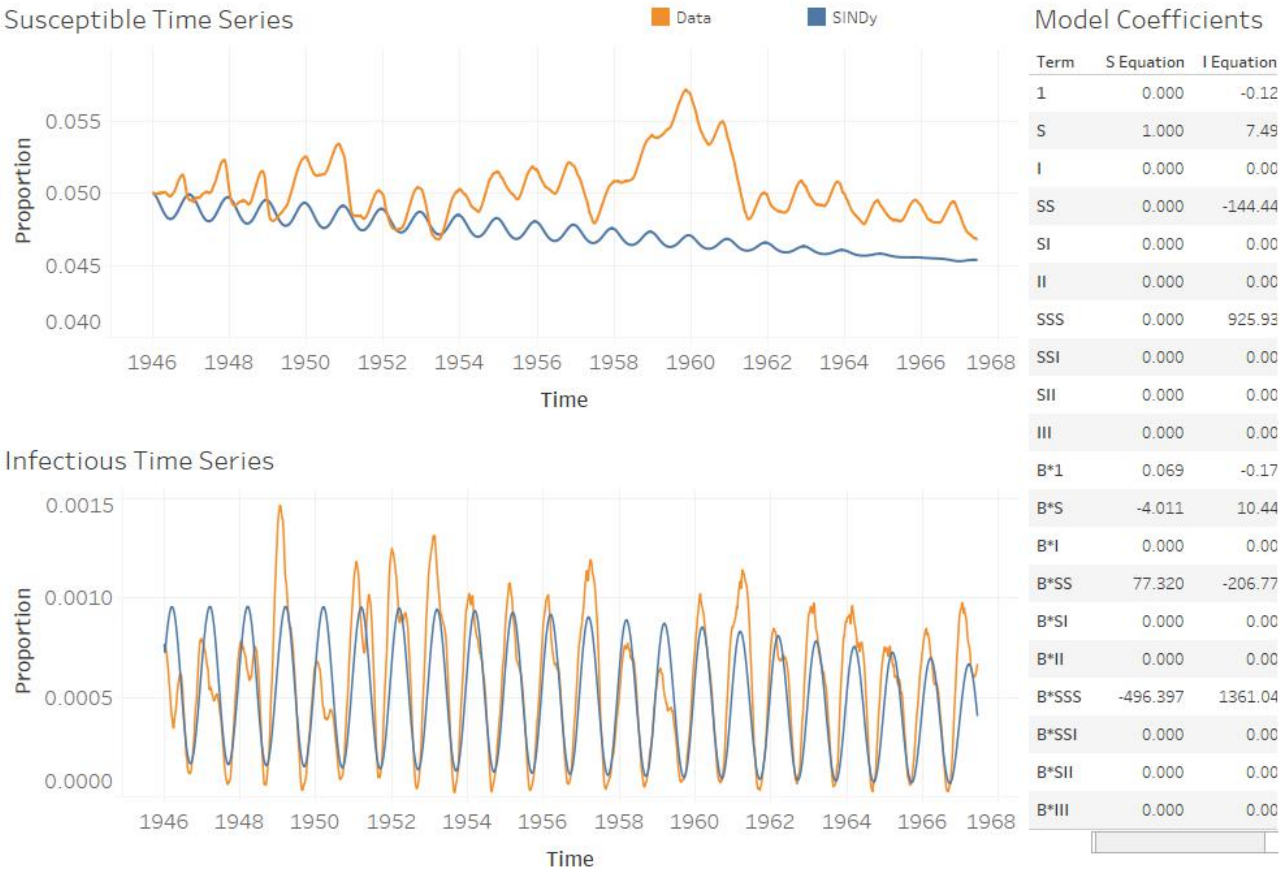


Model Coefficients

Term	S Equation	I Equation
1	-0.147	0.022
S	9.295	-1.247
I	23.803	-8.923
SS	-155.720	23.386
SI	-968.498	377.996
II	2712.909	345.389
SSS	972.259	-146.094
SSI	9664.872	-3616.596
SII	-48146.974	-6361.606
III	-160474.497	-16344.801
B*1	0.005	0.036
B*S	-0.041	-2.204
B*I	0.538	1.940
B*SS	-4.108	44.490
B*SI	12.825	-82.836
B*II	-2629.091	1860.829
B*SSS	58.074	-297.660
B*SSI	-398.769	861.189
B*SII	43755.271	-30140.671
B*III	234042.774	-243853.009

Figure A.8: Resulting time series and coefficients from SINDy-discovered models of the varicella dataset (with a 3rd order polynomial library) using a relatively low threshold ($\lambda = 0.0001$). The resulting model exhibits a good fit and accurate recovery of attractor class, but all possible terms are active, which is an indicator of an overfit model.

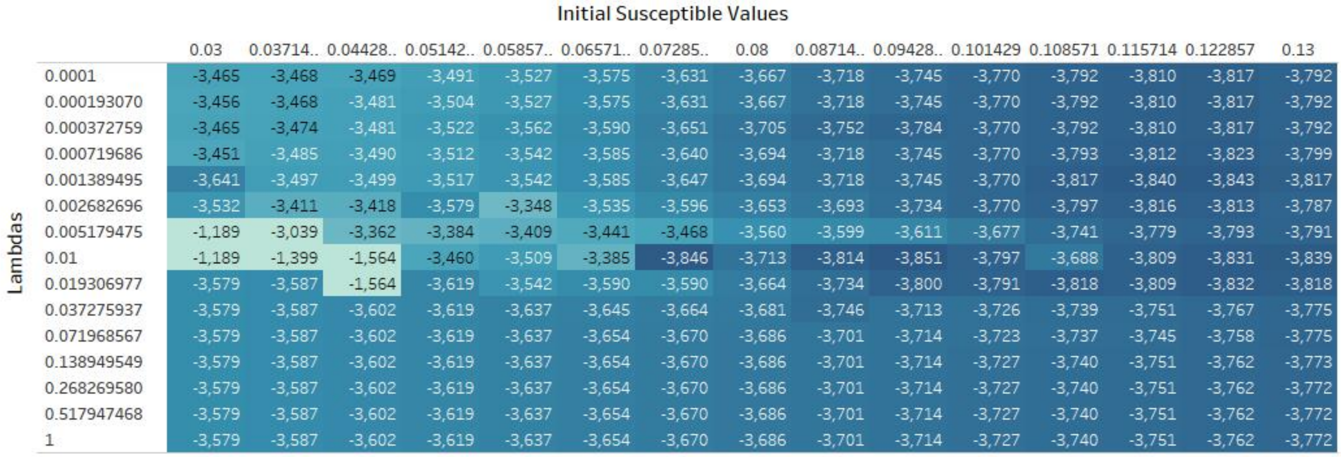
Varicella (Third Order Library), High Threshold



Term	S Equation	I Equation
1	0.000	-0.12
S	1.000	7.49
I	0.000	0.00
SS	0.000	-144.44
SI	0.000	0.00
II	0.000	0.00
SSS	0.000	925.93
SSI	0.000	0.00
SII	0.000	0.00
III	0.000	0.00
B*1	0.069	-0.17
B*S	-4.011	10.44
B*I	0.000	0.00
B*SS	77.320	-206.77
B*SI	0.000	0.00
B*II	0.000	0.00
B*SSS	-496.397	1361.04
B*SSI	0.000	0.00
B*SII	0.000	0.00
B*III	0.000	0.00

Figure A.9: Resulting time series and coefficients from SINDy-discovered models of the varicella dataset (with a 3rd order polynomial library) using a relatively high threshold ($\lambda = 1$). The resulting model exhibits sparsity and recovers the attractor class, but the asymptotic behaviour appears to diverge from the data.

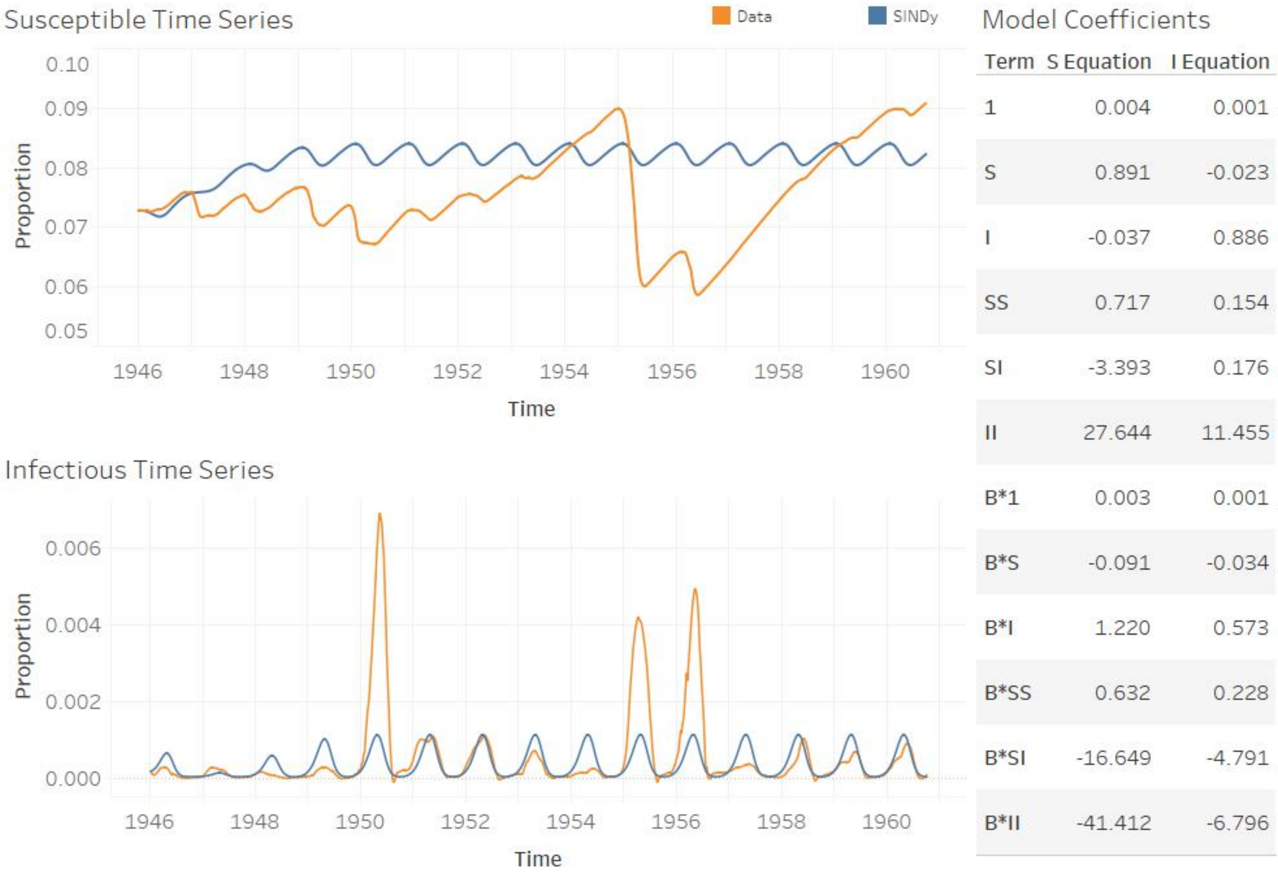
Rubella (Second Order Library), AIC Values



	Initial Susceptible Values														
Lambdas	0.03	0.03714..	0.04428..	0.05142..	0.05857..	0.06571..	0.07285..	0.08	0.08714..	0.09428..	0.101429	0.108571	0.115714	0.122857	0.13
0.0001	-3,465	-3,468	-3,469	-3,491	-3,527	-3,575	-3,631	-3,667	-3,718	-3,745	-3,770	-3,792	-3,810	-3,817	-3,792
0.000193070	-3,456	-3,468	-3,481	-3,504	-3,527	-3,575	-3,631	-3,667	-3,718	-3,745	-3,770	-3,792	-3,810	-3,817	-3,792
0.000372759	-3,465	-3,474	-3,481	-3,522	-3,562	-3,590	-3,651	-3,705	-3,752	-3,784	-3,770	-3,792	-3,810	-3,817	-3,792
0.000719686	-3,451	-3,485	-3,490	-3,512	-3,542	-3,585	-3,640	-3,694	-3,718	-3,745	-3,770	-3,793	-3,812	-3,823	-3,799
0.001389495	-3,641	-3,497	-3,499	-3,517	-3,542	-3,585	-3,647	-3,694	-3,718	-3,745	-3,770	-3,817	-3,840	-3,843	-3,817
0.002682696	-3,532	-3,411	-3,418	-3,579	-3,348	-3,535	-3,596	-3,653	-3,693	-3,734	-3,770	-3,797	-3,816	-3,813	-3,787
0.005179475	-1,189	-3,039	-3,362	-3,384	-3,409	-3,441	-3,468	-3,560	-3,599	-3,611	-3,677	-3,741	-3,779	-3,793	-3,791
0.01	-1,189	-1,399	-1,564	-3,460	-3,509	-3,385	-3,846	-3,713	-3,814	-3,851	-3,797	-3,688	-3,809	-3,831	-3,839
0.019306977	-3,579	-3,587	-1,564	-3,619	-3,542	-3,590	-3,590	-3,664	-3,734	-3,800	-3,791	-3,818	-3,809	-3,832	-3,818
0.037275937	-3,579	-3,587	-3,602	-3,619	-3,637	-3,645	-3,664	-3,681	-3,746	-3,713	-3,726	-3,739	-3,751	-3,767	-3,775
0.071968567	-3,579	-3,587	-3,602	-3,619	-3,637	-3,654	-3,670	-3,686	-3,701	-3,714	-3,723	-3,737	-3,745	-3,758	-3,775
0.138949549	-3,579	-3,587	-3,602	-3,619	-3,637	-3,654	-3,670	-3,686	-3,701	-3,714	-3,727	-3,740	-3,751	-3,762	-3,773
0.268269580	-3,579	-3,587	-3,602	-3,619	-3,637	-3,654	-3,670	-3,686	-3,701	-3,714	-3,727	-3,740	-3,751	-3,762	-3,772
0.517947468	-3,579	-3,587	-3,602	-3,619	-3,637	-3,654	-3,670	-3,686	-3,701	-3,714	-3,727	-3,740	-3,751	-3,762	-3,772
1	-3,579	-3,587	-3,602	-3,619	-3,637	-3,654	-3,670	-3,686	-3,701	-3,714	-3,727	-3,740	-3,751	-3,762	-3,772

Figure A.10: The AIC values for SINDy models across a range of both initial susceptible and threshold values, utilizing the rubella dataset and a 2nd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model. Cells without value refer to a model which, when simulated, resulted in a diverging time series.

Rubella (Second Order Library), Low Threshold

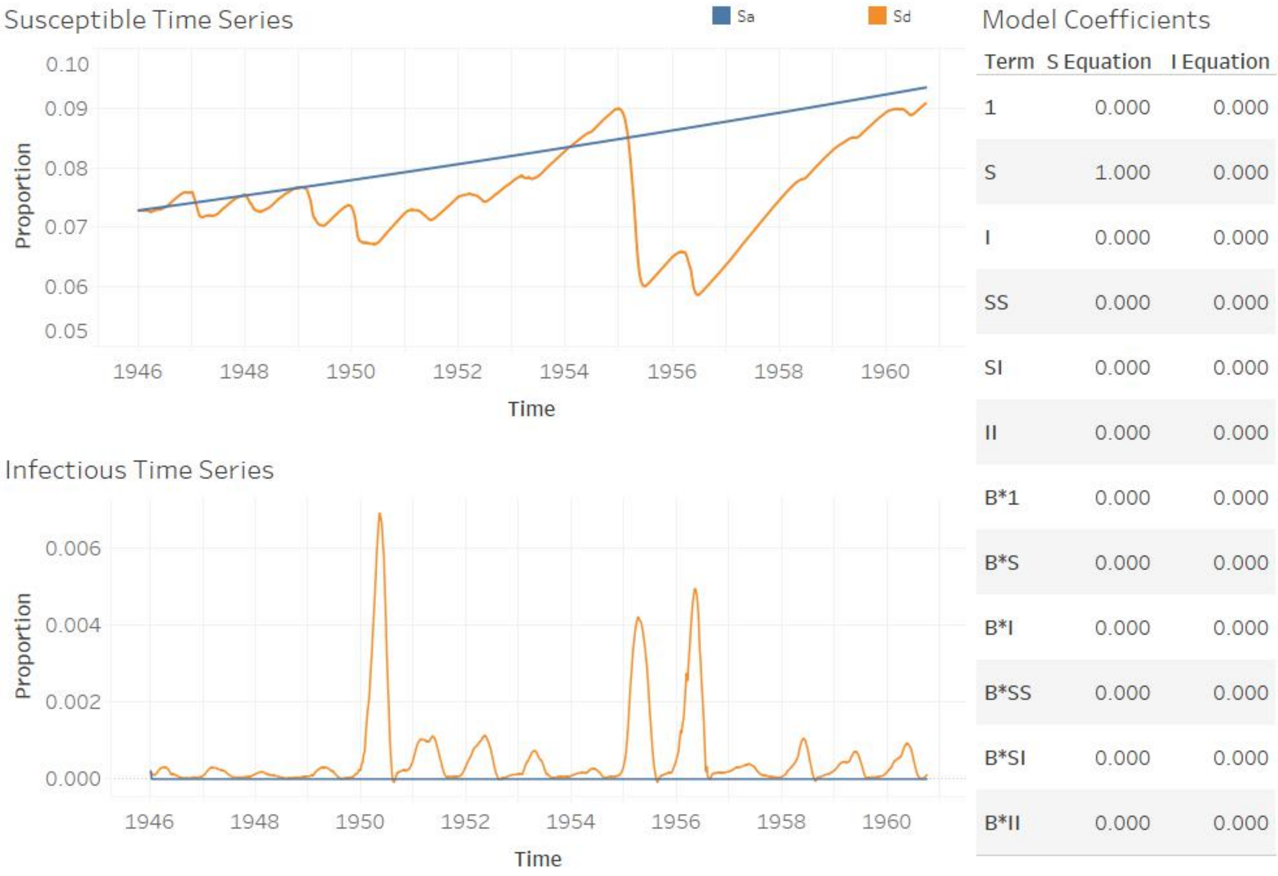


Model Coefficients

Term	S Equation	I Equation
1	0.004	0.001
S	0.891	-0.023
I	-0.037	0.886
SS	0.717	0.154
SI	-3.393	0.176
II	27.644	11.455
B*1	0.003	0.001
B*S	-0.091	-0.034
B*I	1.220	0.573
B*SS	0.632	0.228
B*SI	-16.649	-4.791
B*II	-41.412	-6.796

Figure A.11: Resulting time series and coefficients from SINDy-discovered models of the rubella dataset (with a 2nd order polynomial library) using a relatively low threshold ($\lambda = 0.0001$). The resulting model exhibits neither sparsity nor a well-fitting time series, indicating there is no benefit decreasing the sparsity threshold.

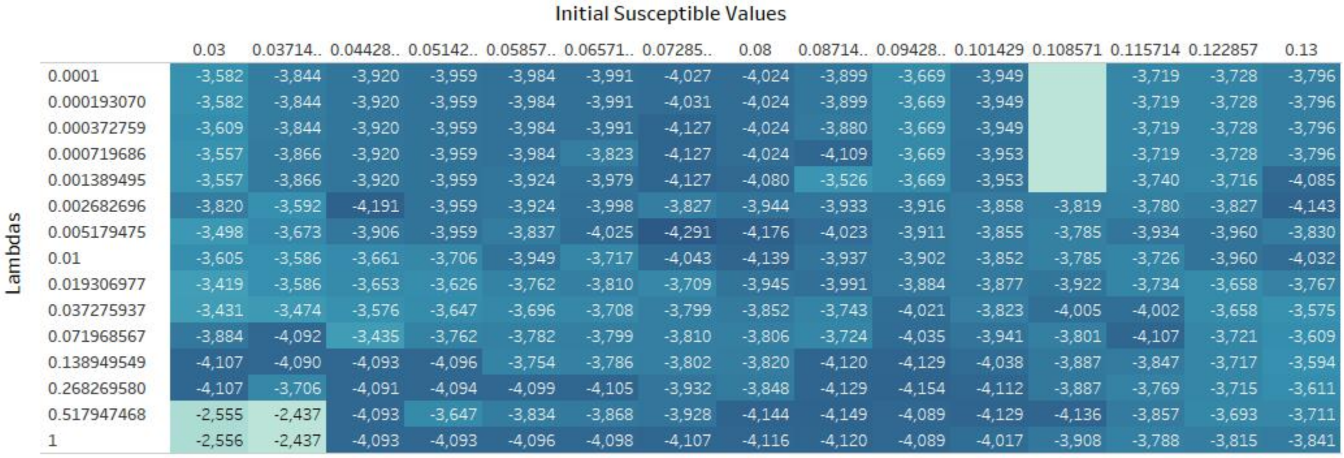
Rubella (Second Order Library), High Threshold



Term	S Equation	I Equation
1	0.000	0.000
S	1.000	0.000
I	0.000	0.000
SS	0.000	0.000
SI	0.000	0.000
II	0.000	0.000
B*1	0.000	0.000
B*S	0.000	0.000
B*I	0.000	0.000
B*SS	0.000	0.000
B*SI	0.000	0.000
B*II	0.000	0.000

Figure A.12: Resulting time series and coefficients from SINDy-discovered models of the rubella dataset (with a 2nd order polynomial library) using a relatively high threshold ($\lambda = 1$). Once the threshold is increased past a critical value, the model is reduced to a linear, which is certainly sparse but does not match the dynamics of the system.

Rubella (Third Order Library), AIC Values

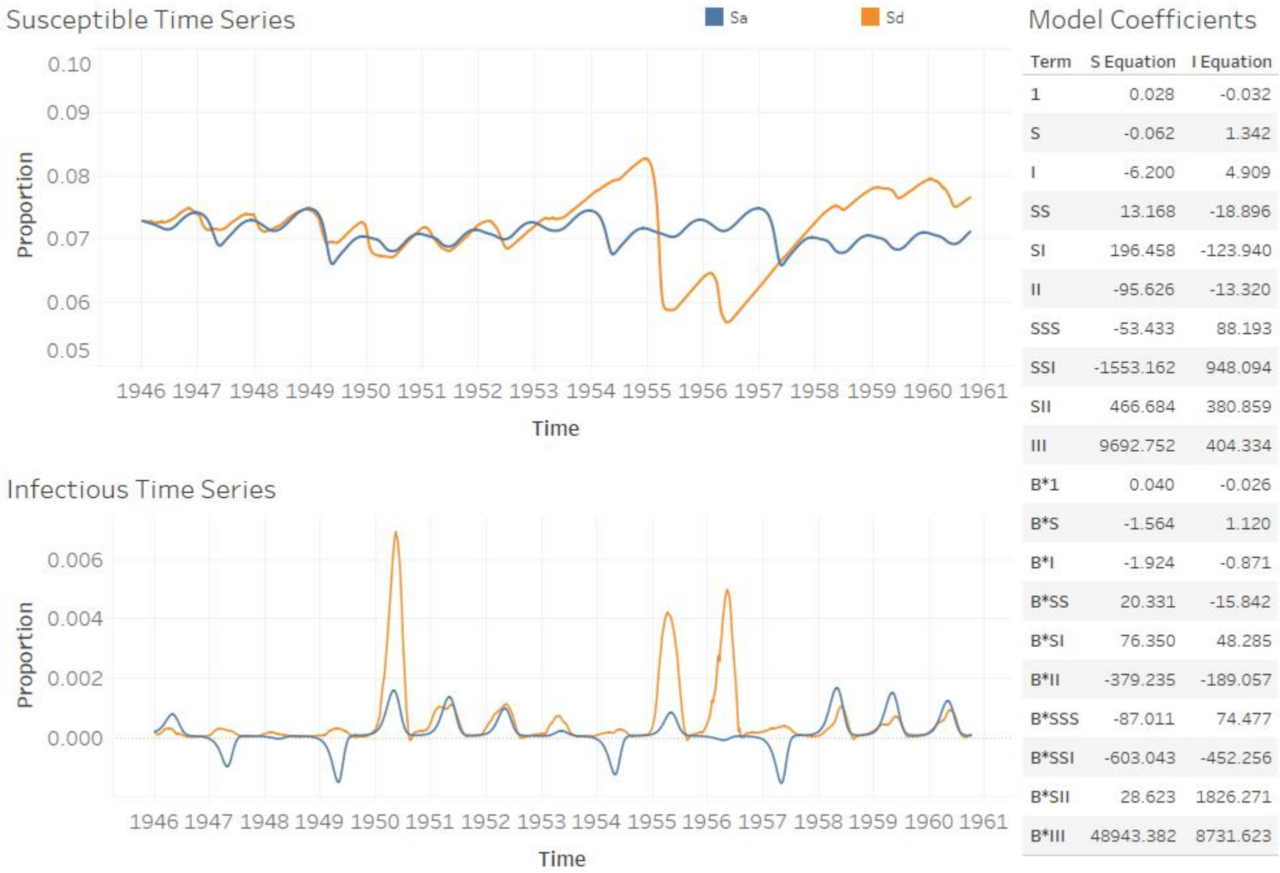


Initial Susceptible Values

Lambdas	0.03	0.03714..	0.04428..	0.05142..	0.05857..	0.06571..	0.07285..	0.08	0.08714..	0.09428..	0.101429	0.108571	0.115714	0.122857	0.13
0.0001	-3,582	-3,844	-3,920	-3,959	-3,984	-3,991	-4,027	-4,024	-3,899	-3,669	-3,949		-3,719	-3,728	-3,796
0.000193070	-3,582	-3,844	-3,920	-3,959	-3,984	-3,991	-4,031	-4,024	-3,899	-3,669	-3,949		-3,719	-3,728	-3,796
0.000372759	-3,609	-3,844	-3,920	-3,959	-3,984	-3,991	-4,127	-4,024	-3,880	-3,669	-3,949		-3,719	-3,728	-3,796
0.000719686	-3,557	-3,866	-3,920	-3,959	-3,984	-3,823	-4,127	-4,024	-4,109	-3,669	-3,953		-3,719	-3,728	-3,796
0.001389495	-3,557	-3,866	-3,920	-3,959	-3,924	-3,979	-4,127	-4,080	-3,526	-3,669	-3,953		-3,740	-3,716	-4,085
0.002682696	-3,820	-3,592	-4,191	-3,959	-3,924	-3,998	-3,827	-3,944	-3,933	-3,916	-3,858	-3,819	-3,780	-3,827	-4,143
0.005179475	-3,498	-3,673	-3,906	-3,959	-3,837	-4,025	-4,291	-4,176	-4,023	-3,911	-3,855	-3,785	-3,934	-3,960	-3,830
0.01	-3,605	-3,586	-3,661	-3,706	-3,949	-3,717	-4,043	-4,139	-3,937	-3,902	-3,852	-3,785	-3,726	-3,960	-4,032
0.019306977	-3,419	-3,586	-3,653	-3,626	-3,762	-3,810	-3,709	-3,945	-3,991	-3,884	-3,877	-3,922	-3,734	-3,658	-3,767
0.037275937	-3,431	-3,474	-3,576	-3,647	-3,696	-3,708	-3,799	-3,852	-3,743	-4,021	-3,823	-4,005	-4,002	-3,658	-3,575
0.071968567	-3,884	-4,092	-3,435	-3,762	-3,782	-3,799	-3,810	-3,806	-3,724	-4,035	-3,941	-3,801	-4,107	-3,721	-3,609
0.138949549	-4,107	-4,090	-4,093	-4,096	-3,754	-3,786	-3,802	-3,820	-4,120	-4,129	-4,038	-3,887	-3,847	-3,717	-3,594
0.268269580	-4,107	-3,706	-4,091	-4,094	-4,099	-4,105	-3,932	-3,848	-4,129	-4,154	-4,112	-3,887	-3,769	-3,715	-3,611
0.517947468	-2,555	-2,437	-4,093	-3,647	-3,834	-3,868	-3,928	-4,144	-4,149	-4,089	-4,129	-4,136	-3,857	-3,693	-3,711
1	-2,556	-2,437	-4,093	-4,093	-4,096	-4,098	-4,107	-4,116	-4,120	-4,089	-4,017	-3,908	-3,788	-3,815	-3,841

Figure A.13: The AIC values for SINDy models across a range of both initial susceptible and threshold values, utilizing the rubella dataset and a 3rd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model. Cells without value refer to a model which, when simulated, resulted in a diverging time series.

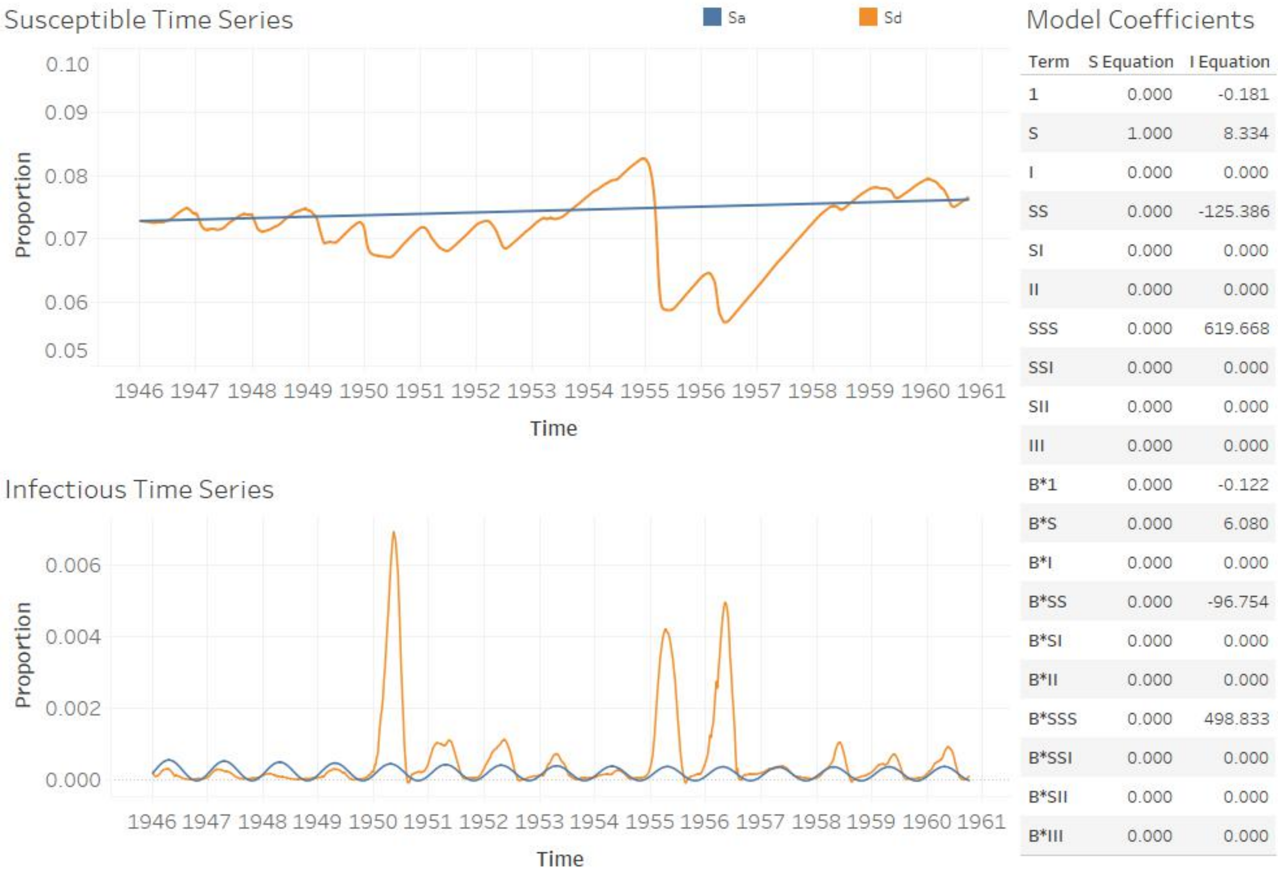
Rubella (Third Order Library), Low Threshold



Term	S Equation	I Equation
1	0.028	-0.032
S	-0.062	1.342
I	-6.200	4.909
SS	13.168	-18.896
SI	196.458	-123.940
II	-95.626	-13.320
SSS	-53.433	88.193
SSI	-1553.162	948.094
SII	466.684	380.859
III	9692.752	404.334
B*1	0.040	-0.026
B*S	-1.564	1.120
B*I	-1.924	-0.871
B*SS	20.331	-15.842
B*SI	76.350	48.285
B*II	-379.235	-189.057
B*SSS	-87.011	74.477
B*SSI	-603.043	-452.256
B*SII	28.623	1826.271
B*III	48943.382	8731.623

Figure A.14: Resulting time series and coefficients from SINDy-discovered models of the rubella dataset (with a 3rd order polynomial library) using a relatively low threshold ($\lambda = 0.0001$). The resulting model attempts to capture the peaks of the underlying multiennial attractor, but at the cost of a high number of active nonlinearities. Note that the negative peaks in the infectious time series are not biologically relevant, but can be adjusted using a constraint when simulating.

Rubella (Third Order Library), High Threshold



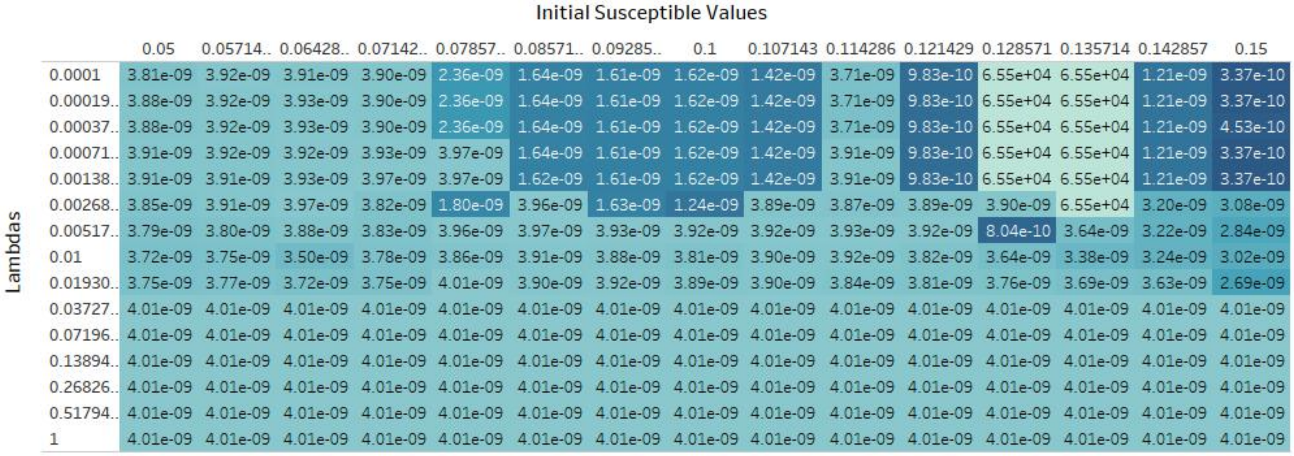
Model Coefficients

Term	S Equation	I Equation
1	0.000	-0.181
S	1.000	8.334
I	0.000	0.000
SS	0.000	-125.386
SI	0.000	0.000
II	0.000	0.000
SSS	0.000	619.668
SSI	0.000	0.000
SII	0.000	0.000
III	0.000	0.000
B*1	0.000	-0.122
B*S	0.000	6.080
B*I	0.000	0.000
B*SS	0.000	-96.754
B*SI	0.000	0.000
B*II	0.000	0.000
B*SSS	0.000	498.833
B*SSI	0.000	0.000
B*SII	0.000	0.000
B*III	0.000	0.000

Figure A.15: Resulting time series and coefficients from SINDy-discovered models of the rubella dataset (with a 3rd order polynomial library) using a relatively high threshold ($\lambda = 1$). The resulting model exhibits sparsity but is linear in the susceptible time series and has an annual oscillation in the infectious time series, neither of which match the dynamics of the system.

A.2 Models Discovered Using PSD for Model Selection

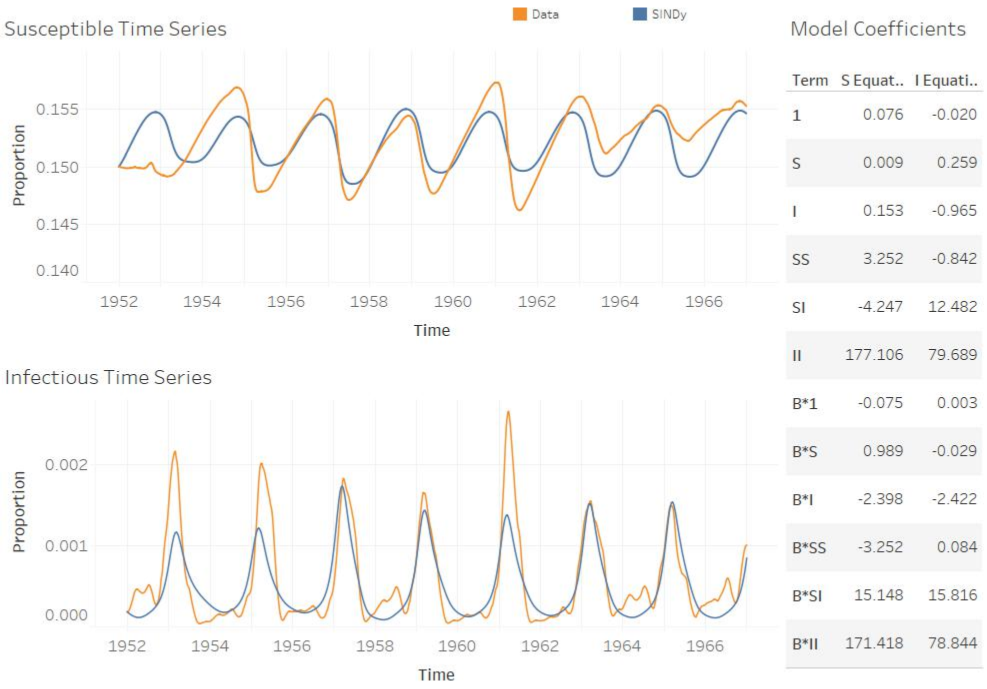
Measles - AIC Values, Best Model, and Selected Model



Initial Susceptible Values

Lambdas	0.05	0.05714..	0.06428..	0.07142..	0.07857..	0.08571..	0.09285..	0.1	0.107143	0.114286	0.121429	0.128571	0.135714	0.142857	0.15
0.0001	3.81e-09	3.92e-09	3.91e-09	3.90e-09	2.36e-09	1.64e-09	1.61e-09	1.62e-09	1.42e-09	3.71e-09	9.83e-10	6.55e+04	6.55e+04	1.21e-09	3.37e-10
0.00019..	3.88e-09	3.92e-09	3.93e-09	3.90e-09	2.36e-09	1.64e-09	1.61e-09	1.62e-09	1.42e-09	3.71e-09	9.83e-10	6.55e+04	6.55e+04	1.21e-09	3.37e-10
0.00037..	3.88e-09	3.92e-09	3.93e-09	3.90e-09	2.36e-09	1.64e-09	1.61e-09	1.62e-09	1.42e-09	3.71e-09	9.83e-10	6.55e+04	6.55e+04	1.21e-09	4.53e-10
0.00071..	3.91e-09	3.92e-09	3.92e-09	3.93e-09	3.97e-09	1.64e-09	1.61e-09	1.62e-09	1.42e-09	3.91e-09	9.83e-10	6.55e+04	6.55e+04	1.21e-09	3.37e-10
0.00138..	3.91e-09	3.91e-09	3.93e-09	3.97e-09	3.97e-09	1.62e-09	1.61e-09	1.62e-09	1.42e-09	3.91e-09	9.83e-10	6.55e+04	6.55e+04	1.21e-09	3.37e-10
0.00268..	3.85e-09	3.91e-09	3.97e-09	3.82e-09	1.80e-09	3.96e-09	1.63e-09	1.24e-09	3.89e-09	3.87e-09	3.89e-09	3.90e-09	6.55e+04	3.20e-09	3.08e-09
0.00517..	3.79e-09	3.80e-09	3.88e-09	3.83e-09	3.96e-09	3.97e-09	3.93e-09	3.92e-09	3.92e-09	3.93e-09	3.92e-09	8.04e-10	3.64e-09	3.22e-09	2.84e-09
0.01	3.72e-09	3.75e-09	3.50e-09	3.78e-09	3.86e-09	3.91e-09	3.88e-09	3.81e-09	3.90e-09	3.92e-09	3.82e-09	3.64e-09	3.38e-09	3.24e-09	3.02e-09
0.01930..	3.75e-09	3.77e-09	3.72e-09	3.75e-09	4.01e-09	3.90e-09	3.92e-09	3.89e-09	3.90e-09	3.84e-09	3.81e-09	3.76e-09	3.69e-09	3.63e-09	2.69e-09
0.03727..	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09
0.07196..	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09
0.13894..	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09
0.26826..	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09
0.51794..	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09
1	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09	4.01e-09

Figure A.16: The AIC values for SINDy models generated from power spectral density estimates of the infectious time series across a range of both initial susceptible and threshold values, using the measles dataset and a 2nd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model.



Model Coefficients

Term	S Equat..	I Equati..
1	0.076	-0.020
S	0.009	0.259
I	0.153	-0.965
SS	3.252	-0.842
SI	-4.247	12.482
II	177.106	79.689
B*1	-0.075	0.003
B*S	0.989	-0.029
B*I	-2.398	-2.422
B*SS	-3.252	0.084
B*SI	15.148	15.816
B*II	171.418	78.844

Figure A.17: Comparison between measles data and the best SINDy-discovered model using a function library of polynomials up to 2nd order and spectral density estimates for model selection.

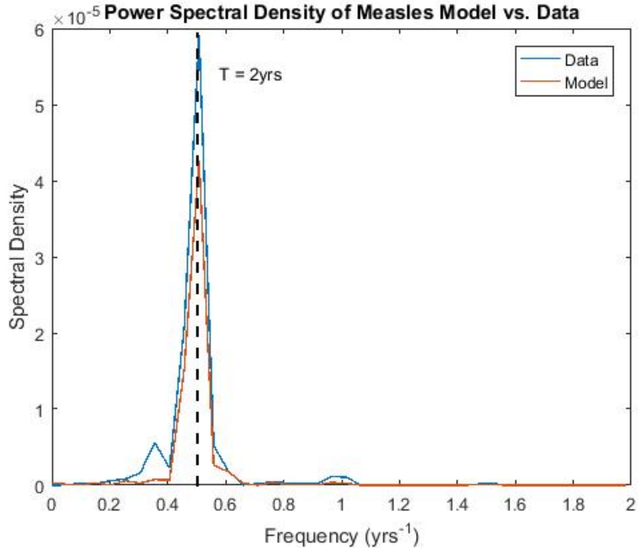
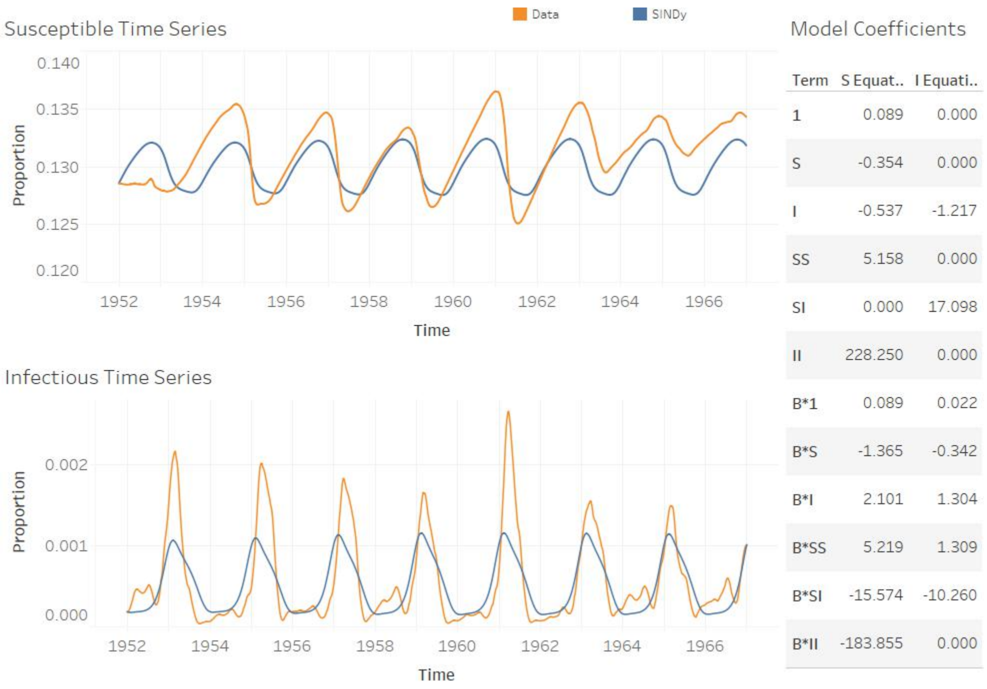


Figure A.18: Comparison of power spectral density estimates of measles data and the most parsimonious SINDy-discovered model. The peak corresponding to the most notable attractor present within the data (with period of 2 years) are noted with the dashed line.



Model Coefficients

Term	S Equat..	I Equati..
1	0.089	0.000
S	-0.354	0.000
I	-0.537	-1.217
SS	5.158	0.000
SI	0.000	17.098
II	228.250	0.000
B*1	0.089	0.022
B*S	-1.365	-0.342
B*I	2.101	1.304
B*SS	5.219	1.309
B*SI	-15.574	-10.260
B*II	-183.855	0.000

Figure A.19: Comparison between measles data and a selected parsimonious SINDy-discovered model using a function library of polynomials up to 2nd order and spectral density estimates for model selection.

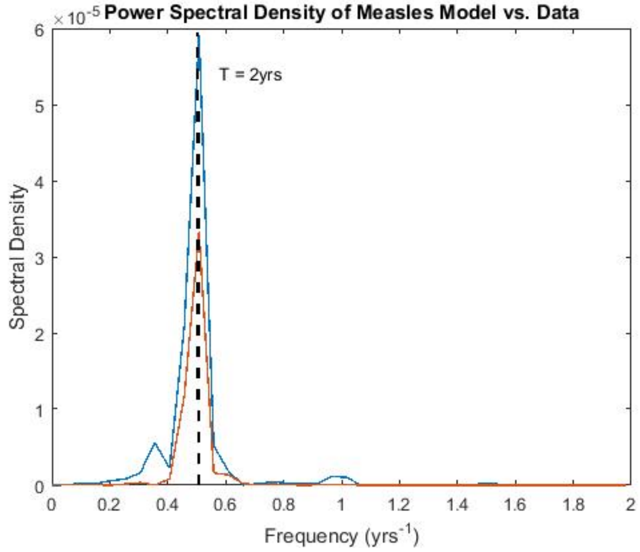
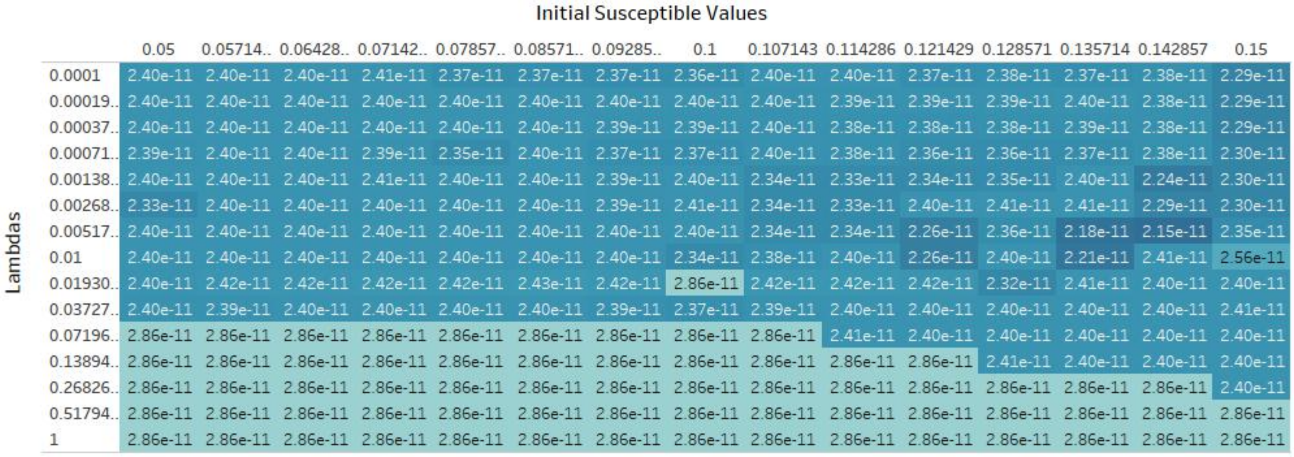


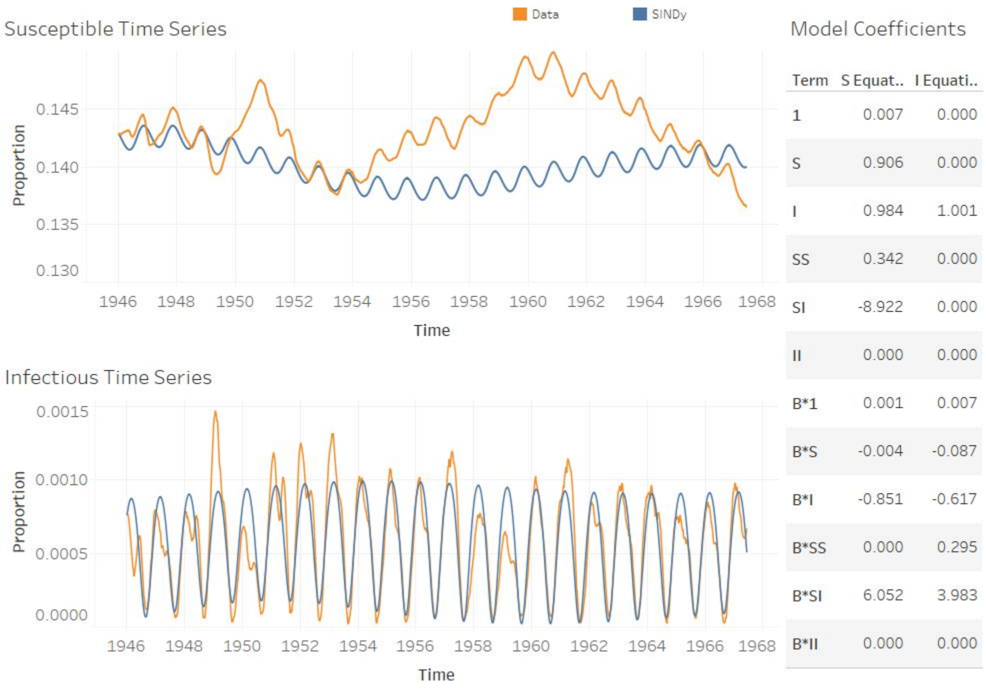
Figure A.20: Comparison of power spectral density estimates of measles data and a selected SINDy-discovered model. The peak corresponding to the most notable attractor present within the data (with period of 2 years) are noted with the dashed line.

Varicella - AIC Values, Best Model, and Selected Model



	Initial Susceptible Values														
Lambdas	0.05	0.05714..	0.06428..	0.07142..	0.07857..	0.08571..	0.09285..	0.1	0.107143	0.114286	0.121429	0.128571	0.135714	0.142857	0.15
0.0001	2.40e-11	2.40e-11	2.40e-11	2.41e-11	2.37e-11	2.37e-11	2.37e-11	2.36e-11	2.40e-11	2.40e-11	2.37e-11	2.38e-11	2.37e-11	2.38e-11	2.29e-11
0.00019..	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.39e-11	2.39e-11	2.39e-11	2.40e-11	2.38e-11	2.29e-11
0.00037..	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.39e-11	2.39e-11	2.40e-11	2.38e-11	2.38e-11	2.38e-11	2.39e-11	2.38e-11	2.29e-11
0.00071..	2.39e-11	2.40e-11	2.40e-11	2.39e-11	2.35e-11	2.40e-11	2.37e-11	2.37e-11	2.40e-11	2.38e-11	2.36e-11	2.36e-11	2.37e-11	2.38e-11	2.30e-11
0.00138..	2.40e-11	2.40e-11	2.40e-11	2.41e-11	2.40e-11	2.40e-11	2.39e-11	2.40e-11	2.34e-11	2.33e-11	2.34e-11	2.35e-11	2.40e-11	2.24e-11	2.30e-11
0.00268..	2.33e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.39e-11	2.41e-11	2.34e-11	2.33e-11	2.40e-11	2.41e-11	2.41e-11	2.29e-11	2.30e-11
0.00517..	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.34e-11	2.34e-11	2.26e-11	2.36e-11	2.18e-11	2.15e-11	2.35e-11
0.01	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.34e-11	2.38e-11	2.40e-11	2.26e-11	2.40e-11	2.21e-11	2.41e-11	2.56e-11
0.01930..	2.40e-11	2.42e-11	2.42e-11	2.42e-11	2.42e-11	2.43e-11	2.42e-11	2.86e-11	2.42e-11	2.42e-11	2.42e-11	2.32e-11	2.41e-11	2.40e-11	2.40e-11
0.03727..	2.40e-11	2.39e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.39e-11	2.37e-11	2.39e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.41e-11
0.07196..	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.41e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11	2.40e-11
0.13894..	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.41e-11	2.40e-11	2.40e-11	2.40e-11
0.26826..	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.40e-11
0.51794..	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11
1	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11	2.86e-11

Figure A.21: The AIC values for SINDy models generated from power spectral density estimates of the infectious time series across a range of both initial susceptible and threshold values, using the varicella dataset and a 2nd order polynomial library. Darker colour refers to a lower AIC value, which indicates a higher quality model.



Term	S Equat..	I Equati..
1	0.007	0.000
S	0.906	0.000
I	0.984	1.001
SS	0.342	0.000
SI	-8.922	0.000
II	0.000	0.000
B*1	0.001	0.007
B*S	-0.004	-0.087
B*I	-0.851	-0.617
B*SS	0.000	0.295
B*SI	6.052	3.983
B*II	0.000	0.000

Figure A.22: Comparison between varicella data and the best SINDy-discovered model using a function library of polynomials up to 2nd order and spectral density estimates for model selection.

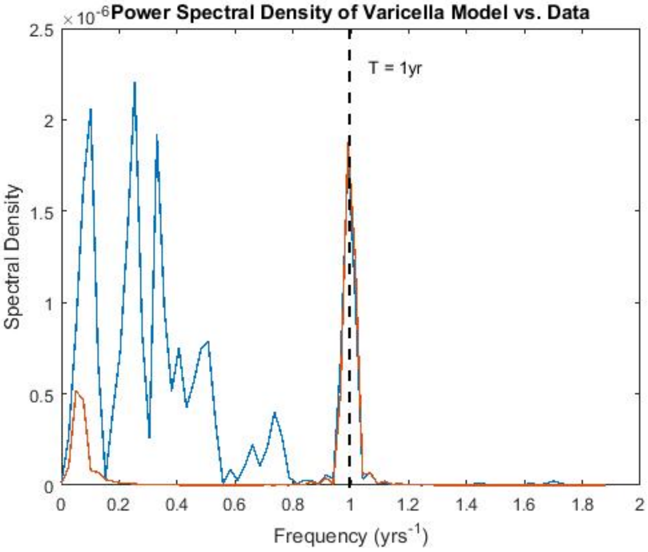


Figure A.23: Comparison of power spectral density estimates of varicella data and the most parsimonious SINDy-discovered model. The peak corresponding to the most notable attractor present within the data (with period of 2 years) are noted with the dashed line.