Algorithmic discovery of dynamic models from infectious disease data

Jonathan Horrocks[1] **and Chris T. Bauch**[1,*]

[1]Department of Applied Mathematics, University of Waterloo, Waterloo, Canada N2L 3G1
[*]cbauch@uwaterloo.ca

ABSTRACT

Theoretical models are typically developed through a deductive process where a researcher formulates a system of dynamic equations from hypothesized mechanisms, for instance. However, recent advances in algorithmic methods are designed to discover sparse dynamic models inductively–directly from data–using minimal prior system knowledge. Most previous research has tested these methods by rediscovering models from synthetic data generated by the already known model. Here we apply Sparse Identification of Nonlinear Dynamics (SINDy) to discover mechanistic equations for disease dynamics from case notification data for measles, chickenpox, and rubella. The discovered models provide a good qualitative fit to the observed dynamics for all three diseases. Moreover, when SINDy uses a library of second-order functions, the discovered models tend to include mass action incidence and a seasonally forced transmission rate–a common feature of existing epidemiological models for childhood infectious diseases. We also show that the measles model discovered by SINDy is capable of out-of-sample prediction of a dynamical regime shift in measles case notification data. These results demonstrate the potential for algorithmic model discovery to enrich scientific understanding by providing a complementary approach to developing theoretical models.

Introduction

Dynamic models that capture the governing mechanisms of systems lie at the heart of many scientific theories of natural systems, ranging from planetary motion to epidemiology [1]. Since Isaac Newton's *Principia mathematica*, research have been devoted to creating models that accurately describe and predict the behaviour of these systems [2]. These models are typically arrived at through a deductive process by hypothesizing mechanisms, formulating dynamic mathematical models that represent those mechanisms, and testing them against data. This tried-and-true approach remains essentially unchanged today.

In the late twentieth century, methods of reconstructing phase spaces or differential equation models from time series data were proposed [3, 4]. More recently, advances in algorithmic sophistication, computational power, and increased data availability have renewed the development of 'model discovery' methods for dynamical system that determine a system of governing dynamical equations from a given dataset [5]. A seminal paper in model discovery uses symbolic regression to recover nonlinear differential equations [6]. This approach automated 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, since in principle it allows prediction under a broad range of possible conditions and not just a replication of the given dataset. In other words, the technique is intended to automatically uncover the nonlinearities that govern system dynamics.

However, early attempts at dynamical systems model discovery were subject to overfitting, as well as being computationally expensive and lacking the ability to scale well to systems with higher dimensionality. Deriving dynamic models from data faces several challenges stemming from large dimensionality. The simplest method to obtain a model that explains the data well minimizes the residual squared error between the predicted response and the data (OLS). This tends to create very complicated models with high descriptive value. However, the models tend to be overfitted to any noise present in the data, compromising their predictive ability [7]. Highly complicated models also exceed human analytic ability, thus detracting from the interpretability of the model.

Most alternative methods to OLS either use subset selection, which attempts to identify some subset of the predictors that adequately describes the system while disregarding the rest [8, 9], or shrinkage (regularization), which fits the model using all of the available predictors but forces the coefficients of select predictors towards zero, thereby performing a kind of variable selection. For instance, SINDy (Sparse Identification of Nonlinear Dynamics) is a recent breakthrough that automates model discovery through sparsity-promoting regression techniques [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. SINDy begins with a large library of nonlinear terms. The algorithm fits the model to the data with the current library, removes any nonlinear terms from the library that have small fitted coefficients, and repeats the process. This progressively shrinks the size of the library until a

relatively small system of differential or difference equations with good explanatory power for the given dataset is obtained. This inductive approach contrasts with the deductive approach of first formulating a model, inferring its parameters from data, and then testing its predictive power [20].

Epidemiological systems have long been studied with dynamic models, on account both of their public health relevance and the complex patterns exhibited by epidemics (Figure 1) [21, 22, 23, 24, 25, 26, 27, 28]. Contemporary epidemic modelling approaches can be traced to the work of Kermack and McKendrick in 1927 [29]. Their compartmental model partitions a population into a series of mutually exclusive compartments–such as susceptible, infected or recovered–and models how individuals move between the compartments [30, 31, 23]. Nonlinearity typically arises from infection transmission. For instance, the commonly used mass-action mixing principle posits that the number of new infections is the product of the number of susceptible and infectious individuals [30, 23, 31]. This principle was originally used to describe the rate of a well-mixed chemical reaction by relating it to the concentration of reactants [32], hence compartmental models conceptualize new infections as resulting from random mixing between susceptible and infectious individuals. These models have been expanded in many ways to account for observed epidemic patterns. For instance, seasonal variation in the transmission rate can give rise to complex dynamics such as bifurcations, oscillations and deterministic chaos [21, 22, 23, 24, 25, 26, 27, 28]. The spatial and temporal dynamics of many childhood infectious disease such as measles, pertussis, rubella and chickenpox are well-described by these dynamic models.

Epidemic models are one of several model systems that have been used to validate SINDy through an approach known as model rediscovery [11, 18, 33]. In this approach, synthetic data are generated by adding noise to the output from a pre-specified model, and the algorithm is applied to the synthetic data to study the conditions under which it can discover the original model [10]. Through this process, SINDy has shown it can not only successfully rediscover the original epidemic models, but it many cases it can also generate hierarchies of models of varying complexity, including new models that can fit the synthetic data but have different model equations from the original model [11, 18, 33]. SINDy has also been used for discovery of new models from empirical data in experimental mechanics and optics [34, 35] as well as to discovery of new models and model reductions from synthetic fluid mechanics data [36, 37]. Thus far, however, SINDy has not yet been tested in model discovery from empirical data on infectious disease dynamics. This represents a very different challenge for model discovery algorithms: the complicated and noisy nature of real-world epidemiological data represents a challenge for generating sparse, low-dimensional models that can provide mechanistic insights or predict real-world dynamics.

Here, using minimal prior epidemiological knowledge, we apply SINDy to time series data from measles, rubella and chickenpox to discover the dynamic models that govern their epidemic patterns, and we compare the resulting models to a standard compartmental model for these infections. We choose epidemiological systems because complex biological systems present a nontrivial challenge for SINDy, and these three infectious diseases display a wide range of dynamical behaviour. At the same time, a significant amount of research in compartmental epidemic models suggests that we have a good idea of what kind of discovered models would satisfy the criteria of interpretability and predictive power we expect from a dynamic model. This exercise allows us to determine whether inductive model discovery techniques have potential to identify new dynamic epidemiological models. Perhaps more importantly, this exercise could also tell us whether model discovery can enrich and nuance our understanding of existing study systems by uncovering a role for nonlinear mechanisms that were previously not considered. In the next section, we first apply SINDy to model rediscovery of a compartmental model with a seasonally varying transmission rate. Then we apply SINDy to model discovery from infectious disease case notification data for measles, rubella and chickenpox, and compare it to a standard compartmental epidemic model.

Results

Model Rediscovery from Simulated Data

For model rediscovery we use a discrete-time Susceptible-Infectious-Recovered (SIR) model accounting for demographic processes (birth and death) and seasonal variation in the transmission rate:

$$S_{t+1} = S_t + \nu - \beta(t) S_t I_t - \mu S_t, \quad (1)$$

$$I_{t+1} = I_t + \beta(t) S_t I_t - \gamma I_t - \mu I_t, \quad (2)$$

$$R_{t+1} = R_t + \gamma I_t - \mu R_t, \quad (3)$$

where S_t (I_t, R_t) is the number of susceptible (infectious, recovered) persons, t is the timestep, ν (μ) is the per capita birth (death) rate per timestep, γ is the per capita recovery rate per timestep, and $\beta(t)$ is the seasonally-varying transmission rate with form given by

$$\beta(t) = \beta_0(1 + \beta_1 \cos(2\pi t/T - \phi)), \quad (4)$$

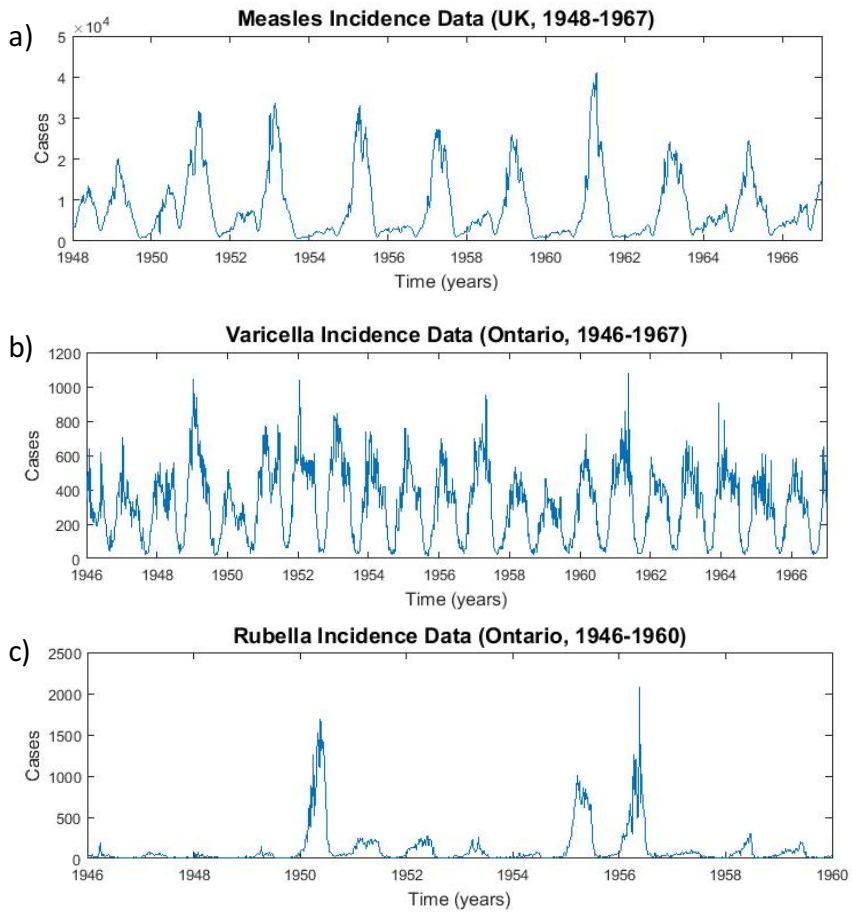


Figure 1. Case notifications (number of cases reported each week) for (a) measles in the United Kingdom, (b) chickenpox in Ontario and (c) rubella in Ontario. Data are from the International Infectious Disease Data Archive (http://iidda.mcmaster.ca/).

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. We interpreted one timestep to correspond to one week. The mass action incidence term ($\beta_t S_t I_t$) appears in both the S_{t+1} and I_{t+1} equations. Note that we may treat the state variable R as redundant since it does not appear in the other equations, and thus we may exclude it from simulations.

We used a discrete-time instead of a continuous-time (differential equation) model because SINDy necessitates evaluating the derivative of the input data, which for a continuous time model applied to empirical data can yield noisy and unpredictable values, making it difficult for the sparse regression algorithm to obtain a global minimum. In contrast, applying SINDy to a discrete-time modelling framework lends itself naturally to the weekly temporal resolution of the empirical data we used. Hence we use a discrete-time compartmental model for both model rediscovery as well as for model discovery from empirical data. However, we note that model rediscovery with SINDy also works with a continuous-time compartmental model.

We solved the discrete-time SIR model numerically and added Gaussian noise to the output state variables to generate a simulated dataset. SINDy was then applied to the simulated dataset using a function library consisting of all polynomials involving S_t and I_t up to and including second order, for both constant and seasonally-varying coefficients (see Methods). This process was repeated for a range of noise magnitudes.

For sufficiently small noise, SINDy was able to recover the original model with a very high accuracy by correctly recovering the coefficients of the SIR model (Figure 2). However, at higher (but still relatively small) noise magnitudes, SINDy begins to overfit the noise in the simulated data and, while it still correctly identifies the seasonally-varying mass action mixing term as having the largest coefficient, it also incorrectly identifies other nonlinear terms as having a role in the model, such as I_t^2 (Figure 2). The relatively small values for noise at which these spurious terms are introduced illustrates the challenges that noisy data present for SINDy. (However, it is also possible that other methods for regularization not explored here might enable SINDy to fit the data well even for the higher noise case.) Results for other noise levels appear in Supplementary Figures 1-4.

Model Discovery from Empirical Data

We studied three datasets consisting of case notifications for measles in England and Wales (1948-1967); chickenpox in Ontario, Canada (1946-1967); and rubella in Ontario, Canada (1946-60) (Figure 1). Each provides a different example of an attractor class: measles is biennial, chickenpox is predominantly annual, and rubella is multiennial with a weaker annual signatures as well [27, 25]. Hence, these data represent an interesting test for whether a SINDy model discovered from a library of annually-forced transmission rates can generate a model that can predict not only annual but also biennial or multiennial dynamics. Our baseline analysis used a second-order polynomial library. However, we also tried a third-order polynomial library, which in principle allows capturing more features of the data but also risks more overfitting of noise. The data in Figure 1 was first smoothed to reduce the risk of overfitting (see Methods).

The datasets describe the case incidence (number of new cases per week), but our SINDy state variables concern the prevalence of epidemiological states: the number of susceptible individuals (S_t) and infectious individuals (I_t, the infection prevalence) at any given time t. Hence we converted case incidence to case prevalence to obtain the required input time series of the number of infectious individuals (see Methods).

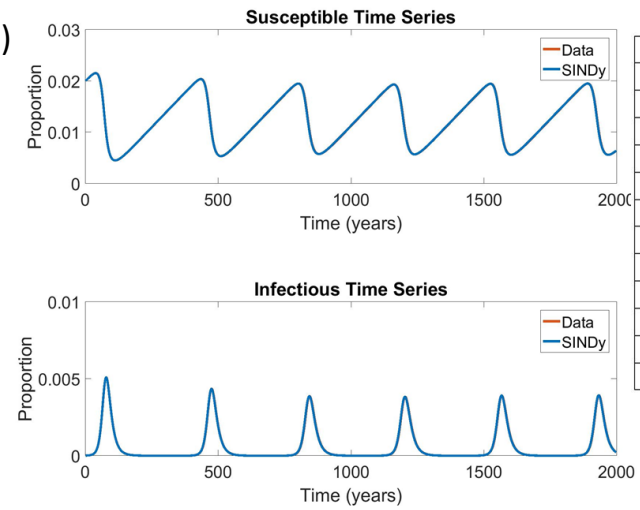
We do not have data on the temporal dynamics of susceptible individuals, so we reconstructed a time series of the number of susceptible individuals using a standard method (see Methods) [38]. The method requires the initial number of susceptible persons, S_0, and this initial condition can strongly influence disease dynamics. The number of susceptible individuals in a given year is not known for any of our historical datasets. Similarly, as mentioned in the Introduction, coefficients below a certain sparsity threshold value (λ, the "sparsity knob") are removed from the library at each iteration of SINDy. But, there is no *a priori* knowledge to guide the selection of the value of λ. Hence, we applied SINDy to each point of the $S_0 - \lambda$ parameter grid. The Akaike Information Criterion (AIC) score of the model identified at each point on the parameter grid was computed to give us a way of measuring model parsimony across the grid [39]. To quantify sparsity of the coefficient matrix we introduced a *sparsity index* which is the ratio of the number of library functions with zero coefficients to the total number of functions in the library:

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

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

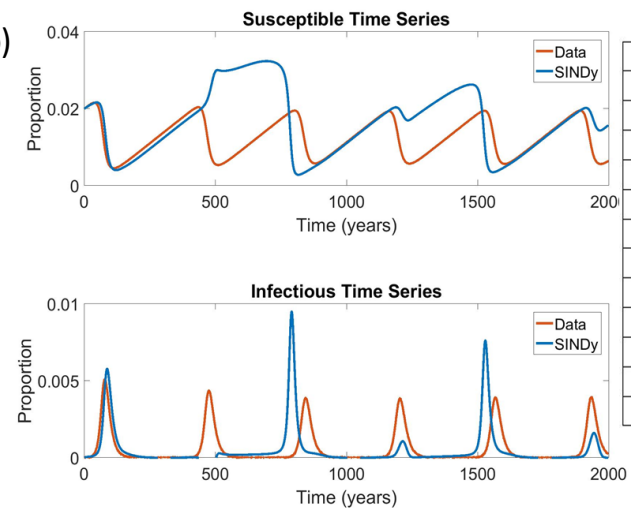
We found that SINDy discovered models that reflect the observed dynamics of infectious individuals for both measles and chickenpox (Figures 3 and 4). This includes both the biennial attractor of measles and the annual attractor of chickenpox. The models are not very sparse ($r = 0.25$) although the remaining coefficients differ greatly in the values SINDy assigns to them. For measles, SINDy assigned the largest coefficients to the bilinear incidence terms SI and βSI, corresponding to constant and seasonally varying mass-action incidence terms respectively, as in the discrete-time SIR model (Figure 3). SINDy also assigned large values for the SI and βSI coefficients for chickenpox, although the βII term was assigned the largest value overall (Figure

a)



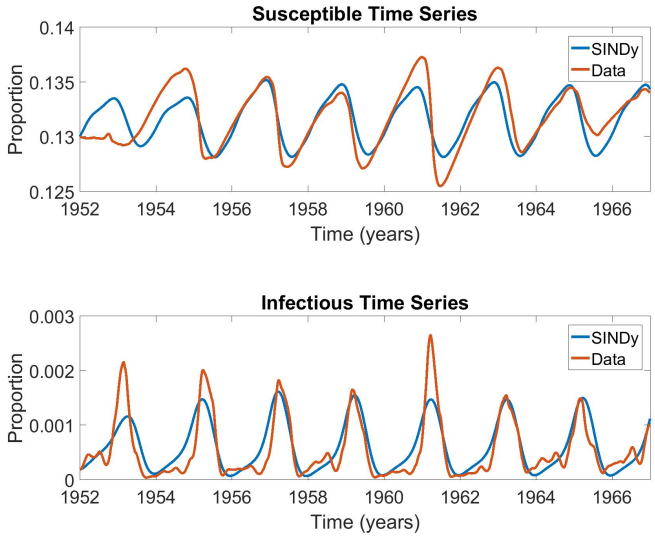
Term	S Eq.	I Eq.
1	5.4784×10^{-5}	0
S	0.9999	0
I	-1.2615×10^{-5}	0.9
S^2	-8.6338×10^{-5}	0
SI	-7.9968	7.9985
I^2	-0.0033	0.0013
$\beta(t)$	0	0
$\beta(t)S$	0	0
$\beta(t)I$	1.7848×10^{-5}	-7.4721×10^{-6}
$\beta(t)S^2$	1.1559×10^{-5}	0
$\beta(t)SI$	-2.0036	2.0016
$\beta(t)I^2$	0.0028	-0.0011

b)



Term	S Eq.	I Eq.
1	4.6829×10^{-5}	6.9399×10^{-5}
S	1.0014	-0.0103
I	-0.0034	0.9447
S^2	-0.05888	0.3845
SI	-6.3191	-14.4837
I^2	-3.2957	39.9876
$\beta(t)$	0	-3.542×10^{-5}
$\beta(t)S$	-2.5394×10^{-5}	0.0069
$\beta(t)I$	0.0069	-0.0688
$\beta(t)S^2$	0.0092	-0.3059
$\beta(t)SI$	-3.9521	26.4513
$\beta(t)I^2$	3.0368	-38.9909

Figure 2. Comparison of the simulated SIR model with additive noise of $\varepsilon = 1\times10^{-7}$ (top) and $\varepsilon = 2\times10^{-5}$ (bottom) with the corresponding coefficients of terms in the discovered model. $\beta_0 = 8$/wk, $\beta_1 = 0.25$, $\gamma = 0.1$/wk, $\mu = \nu = 5.4795 \times10^{-5}$/wk. The results in the table display the SINDy-discovered coefficients of the corresponding terms in (Eqs. 1 - 3). In panel (a) the Data and SINDy curves are very closely overlapping, hence the appearance of only a single blue curve.



Term	S Eq.	I Eq.
1	0.025	0.002
S	0.606	-0.037
I	1.084	-1.554
S^2	1.541	0.139
SI	-10.295	20.618
I^2	0	0
$\beta(t)$	-0.009	-0.013
$\beta(t)S$	0.146	0.2
$\beta(t)I$	-3.122	0
$\beta(t)S^2$	-0.591	-0.0779
$\beta(t)SI$	26.409	0
$\beta(t)I^2$	0	0

Figure 3. Comparison between measles incidence data and coefficients of the best SINDy-discovered model using a function library of polynomials up to 2nd order, and showing the model with the lowest AIC scores across the $S_0 - \lambda$ parameter grid. 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 excluded six terms, giving $r = 0.25$. $S_0 = 0.11286$ and $\lambda = 0.00517$. The parameter grids appear in SI Appendix, Figure 5. The results in the table display the SINDy-discovered coefficients of the corresponding terms in (Eqs. 1 - 3).

4). The model that SINDy discovered for rubella was more sparse ($r = 0.7$), and SINDy also assigned the largest coefficients to the mass-action mixing terms SI and βSI, although the model predicts an annual attractor instead of the observed multiennial attractor (Figure 5). This occurs despite the fact that models with bilinear incidence are capable of generating multiennial attractors a rubella [25, 27].

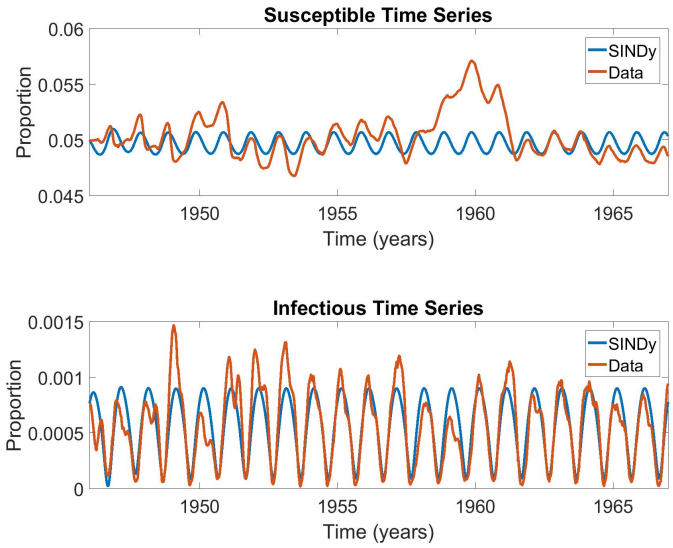
The agreement between model and data for measles and chickenpox is low compared to many well-controlled examples from physics [10], but it is comparable to the agreement achieved in many studies in epidemiological modelling [40]. This is because biological and social systems are complex and have multiple nonlinear feedbacks, in addition to numerous environmental heterogeneities. We discuss the case of rubella further in a following subsection.

The SINDy-discovered models confirm that seasonal forcing plays an important role in epidemic dynamics. The seasonally forced transmission rate (B*S*I) is always the first or second largest term in the discovered models. This provides a separate line of evidence in support of deductively derived models [27, 25] that support the role of seasonal variation in the transmission rate. Also, despite the fact that seasonal forcing occurs at a period of one year, SINDy can recover a seasonally forced model that exhibits a biennial attractor, hence these results provide another line of evidence that seasonal forcing can generate attractors of multiple different periods [25], some of which are epidemiological relevant. The emergence of a biennial attractor from seasonal (annual) forcing happens due to the interplay between forcing dynamics, the gradual build-up of new susceptible individuals through births, and its rapid depletion during epidemic periods.

Power spectral density

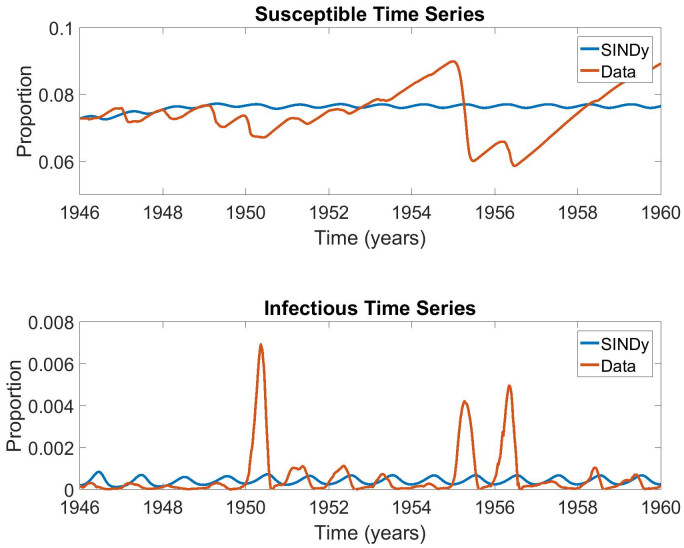
The results for rubella (Figure 5) highlight a limitation of using goodness-of-fit to time series as the criterion. As noted, a bilinear incidence term can generate multiennial attractors under seasonal forcing, but SINDy selected a model that generates an annual attractor under a second-order library. In terms of qualitative descriptions of the data, a model that generates multiennial attractors of the right frequency but with epidemic peaks at different years than what is observed in the dataset is perhaps more desirable, and says more about underlying epidemiological mechanisms, than a model that simply yields an annual attractor.

Hence, we developed a modification of our approach such that the goodness-of-fit between the power spectral density (PSD) of the model and the data is used as the criterion for assessing models, instead of fit between the prevalence time series (see Methods). Using this approach for rubella with a second-order polynomial library yields a model that reproduces the large multiennial outbreaks observe in the empirical dataset, as well as the notable tendency for seasonal rubella incidence to ramp up slowly in the first part of the season, but decline very quickly in the second part (Figure 6). The discovered model also includes a strong contribution from the mass-action incidence term $\beta(t)SI$.



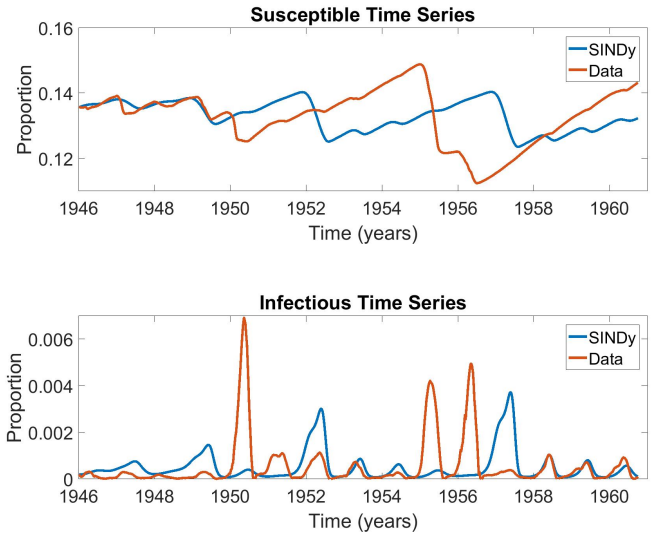
Term	S Eq.	I Eq.
1	0	-0.002
S	1.011	0.081
I	1.043	0.953
S^2	-0.037	-0.776
SI	-26.601	0
I^2	0	0
$\beta(t)$	0	0.001
$\beta(t)S$	-0.024	0.001
$\beta(t)I$	0.215	-1.163
$\beta(t)S^2$	0.267	0
$\beta(t)SI$	-3.749	19.24
$\beta(t)I^2$	0	114.765

Figure 4. Comparison between chickenpox incidence data and coefficients of the best SINDy-discovered model using a function library of polynomials up to 2nd order, and showing the model with the lowest AIC scores across the $S_0 - \lambda$ parameter grid. 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 excluded six terms, giving $r = 0.25$. The parameter grids appear in SI Appendix, Figure 7. The results in the table display the SINDy-discovered coefficients of the corresponding terms in (Eqs. 1 - 3).



Term	S Eq.	I Eq.
1	0.002	0
S	0.95	-0.006
I	0.915	-1.118
S^2	0.351	0
SI	-15.966	27.44
I^2	0	0
$\beta(t)$	0	0
$\beta(t)S$	0	0
$\beta(t)I$	0	2.177
$\beta(t)S^2$	0	0
$\beta(t)SI$	0	-27.746
$\beta(t)I^2$	0	0

Figure 5. Comparison between rubella incidence data and coefficients of the best SINDy-discovered model using a function library of polynomials up to 2nd order, and showing the model with the lowest AIC scores across the $S_0 - \lambda$ parameter grid. The algorithm was unable to discover a model that exhibited the multi-annual 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 excluded 14 terms, giving $r = 0.7$. The parameter grids appear in SI Appendix, Figure 10. The results in the table display the SINDy-discovered coefficients of the corresponding terms in (Eqs. 1 - 3).



Term	S Eq.	I Eq.
1	0.0022	0.0033
S	0.9718	-0.0452
I	-0.3461	-2.2542
S^2	0.0968	0.1547
SI	0	23.7858
I^2	55.9012	0
$\beta(t)$	-0.0015	0
$\beta(t)S$	0.0204	0
$\beta(t)I$	-0.6015	-3.5303
$\beta(t)S^2$	-0.0690	0
$\beta(t)SI$	5.5582	25.6070
$\beta(t)I^2$	0	0

Figure 6. Comparison between rubella incidence data and coefficients of the best SINDy-discovered model using a function library of polynomials up to 2nd order with power spectral density as the criterion for AIC selection. The results in the table display the SINDy-discovered coefficients of the corresponding terms in (Eqs. 1 - 3).

The estimated power spectral densities for the empirical data and the best SINDy models appears in Supplementary Figures 5-7, for all three diseases. These figures show that dominant peak of the SINDy model PSD matches that of the empirical data, for all three diseases: 1 year for chickenpox, 2 years for measles, and 5 years for rubella. However, the SINDy model for chickenpox fails to capture some of the lower frequencies present in the empirical data.

Effect of changing sparsity knob λ

Our approach applies SINDy to a grid of possible values for the initial number of susceptible individuals S_0 and the sparsity knob λ which determines the threshold for removing terms from the library in each iteration of SINDy. The sparsity knob is particularly important because if it values is set too low, the discovered 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 does not resemble the data in a meaningful way.

Hence, in order to balance sparsity with goodness-of-fit, we focussed on the models that yielded the lowest AIC score across the grid for the foregoing results (Figures 3-5). This approach shows there is an optimal region in the $S_0 - \lambda$ plane that ensures the SINDy algorithm can generate regularized, accurate models from empirical disease data (Supplementary Figures 8-15). For instance, in the case of measles, the best AIC corresponded to a model using an initial susceptible value of $S_0 = 0.11286$ and a sparsity knob of $\lambda = 0.00517$ (Figure 3). In contrast, a much lower sparsity setting ($\lambda = 0.0001$) discovers a model that is overfitted to apparently random features of the data and that includes all the terms of the library, whereas a much higher sparsity setting ($\lambda = 0.1$) discovers a model that exhibits annual attractor instead of the characteristics biennial attractor of measles (Supplementary Figure 9).

Third-order polynomial library

When a third-order library is used instead, the discovered models for measles and chickenpox capture the observed epidemiological dynamics as well as the second-order library does, and the models have similar sparsity indices (Supplementary Figures 16-18). SINDy continues to assign significant weight to the bilinear incidence terms SI and βSI, but SINDy assigned even stronger weight to the S^2I and SI^2 terms (and their corresponding seasonal terms). This could indicate overfitting, or it may indicate a more general nonlinear incidence mechanism in the underlying system. It has been shown 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 used [41, 42]. In the case of rubella, SINDy generates a model that captures the multiennial attractor observed in rubella dynamics (Supplementary Figure 15) although the model is not very sparse ($r = 0.15$) and is strongly dependent on trilinear incidence terms S^2I and I^3. Parameter planes for S_0 and λ and examples of predictions for very low and high sparsity thresholds appear in Supplementary Figures 19-27.

Comparison with Compartmental Epidemic Models

We compared the models discovered by SINDy to the classical discrete-time SIR compartmental model with seasonal forcing and mass-action mixing. We fitted Equations (1)-(4) to the case prevalence and reconstructed susceptible time series for all three infections by sweeping across parameter grids and minimizing the least-squared error between model and data time series (see Methods). We thereby inferred the parameter values ν, μ, β_0, β_1 and γ and compared these parameter values to the coefficients of functions determined from SINDy for the models in Figures 3-5.

These comparisons show that SINDy tends to select the coefficients corresponding to mass-action mixing, often with seasonal forcing as well, and that these coefficients have a similar magnitude and sign as the inferred parameters of the compartmental SIR model (Figure 7). The similarities are strongest for measles and chickenpox. This demonstrates that SINDy is effective in capturing the theoretical principles of mass action incidence–a core mechanism of most epidemiological models–as well as seasonal forcing, which is required to explain endemic patterns of many childhood infectious diseases in the pre-vaccine era. The SINDy models also depend on the linear terms for each of the S and I equations. Usually these have a similar magnitude and sign as in the SIR model. These terms correspond to vital dynamics (births and deaths). However, several other features are noticeably different from the inferred SIR model. For instance, SINDy often infers a different magnitude and/or sign than the SIR model in ways that have no obvious and immediate interpretation.

SINDy model can predict qualitative shifts in empirical measles dynamics out-of-sample

To test whether models discovered by SINDy can be predict real-world dynamics, we studied the ability of the second-order SINDy measles model to make out-of-sample prediction of regime shifts in measles dynamics. We used a classic example of nonlinear measles dynamics in the United Kingdom [25, 27]. During 1948-1967, the recruitment rate of new susceptible individuals in England & Wales was roughly constant and measles incidence exhibited a clear biennial pattern (Figure 1). However, the recruitment rate of new susceptible individuals dropped significantly in 1967, due to the introduction of mass vaccination and a decline in birth rates, causing dynamics to shift suddenly to a more irregular pattern dominating the time period 1968 to 1988 [25, 27]. Previous research has shown that simple compartmental epidemic models can predict this shift, as well as the patterns observed before and after the shift [25, 27]. These models characterize the dynamics from 1948-1967 as a biennial attractor [25] that is relatively stable to perturbations from noise (on account of the fact that the period of the biennium's Floquet multiplier–which predicts response to noise–is two years and thus exactly matches the two-year period of the biennial attractor [27]. The dynamics from 1968-1988 are characterized as an annual attractor that is more easily perturbed by noise (on account of the Floquet multiplier of the annual attractor having a non-annual period greater than one year [27]). These differences have implications for the power spectra of the time series of infection incidence. In the power spectra of both the classical models and the empirical data, we observe a shift from a dominant peak at two years and very little power elsewhere in the spectrum (except for a supporting annual peak) during the biennial era, to a dominant peak at one year and a second peak at a period of approximately 2.5 years (corresponding to the period of the attractor's Floquet multiplier), during the era of irregular dynamics [27].

The second-order SINDy model for measles was discovered from the incidence patterns observed during the biennial era (Figure 3). We hypothesized that if the SINDy model is discovering real-world mechanisms and not over-fitting the data, it should predict the same transition observed in the empirical data and predicted by previous models [27, 25]. To test this hypothesis, we reduced the susceptible recruitment rate in the SINDy model (the coefficient of the S term) to match the drop in the susceptible recruitment rate observed in the United Kingdom from 1948-1967 to 1968-1988 and we compared the time series and power spectra of infection incidence and susceptible individuals before and after the drop. We also added white noise to the simulations to test the response of the attractors to noise. We found that the SINDy model predicts the same transition. In the biennial era, the SINDy model predicts a stable biennial cycle that is relatively robust to noise: the time series of infection incidence shows a clear biennial pattern and the power spectrum shows a strong peak at a period of two years, a supporting peak at one year, and little else (Figure 8 and 3). In simulations where the coefficient of the S term is reduced to capture a drop in the recruitment rate, the SINDy model predicts a noisy annual cycle: the time series of infection incidence shows a dominant annual signature in the presence of significant irregularities and perhaps an envelope of a non-seasonal periodicity, and the corresponding power spectrum shows a strong annual peak with a secondary peak at a higher period, caused by the response of the annual attractor to noise. The time series also resemble the empirical post-vaccine dynamics [27, 25]. Hence, the SINDy model is predicting the empirically observed regime shift in measles dynamics out-of-sample. (Predicted susceptible dynamics are also shown in Figures 8, 9 but were not analyzed in Refs. [25, 27]. Ten additional stochastic realizations with different starting random number seeds and their power spectra appear in the SI appendix: Figures 28-37. These show similar patterns to Figures 8 and 9.)

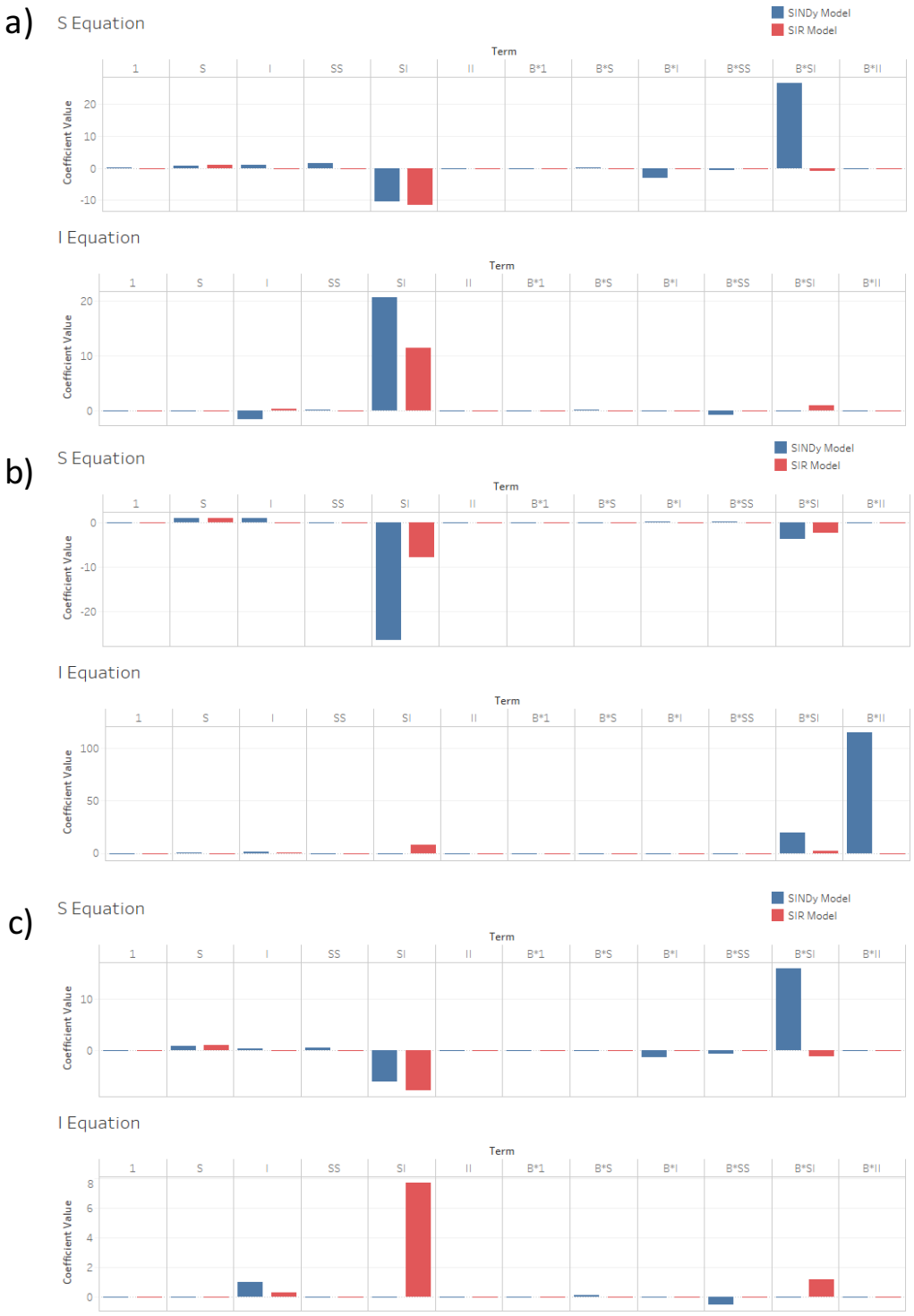


Figure 7. Comparison of coefficients between SINDy-discovered model (using a function library of 1st and 2nd order polynomials) and fitted SIR model for (a) measles, (b) chickenpox and (c) rubella.

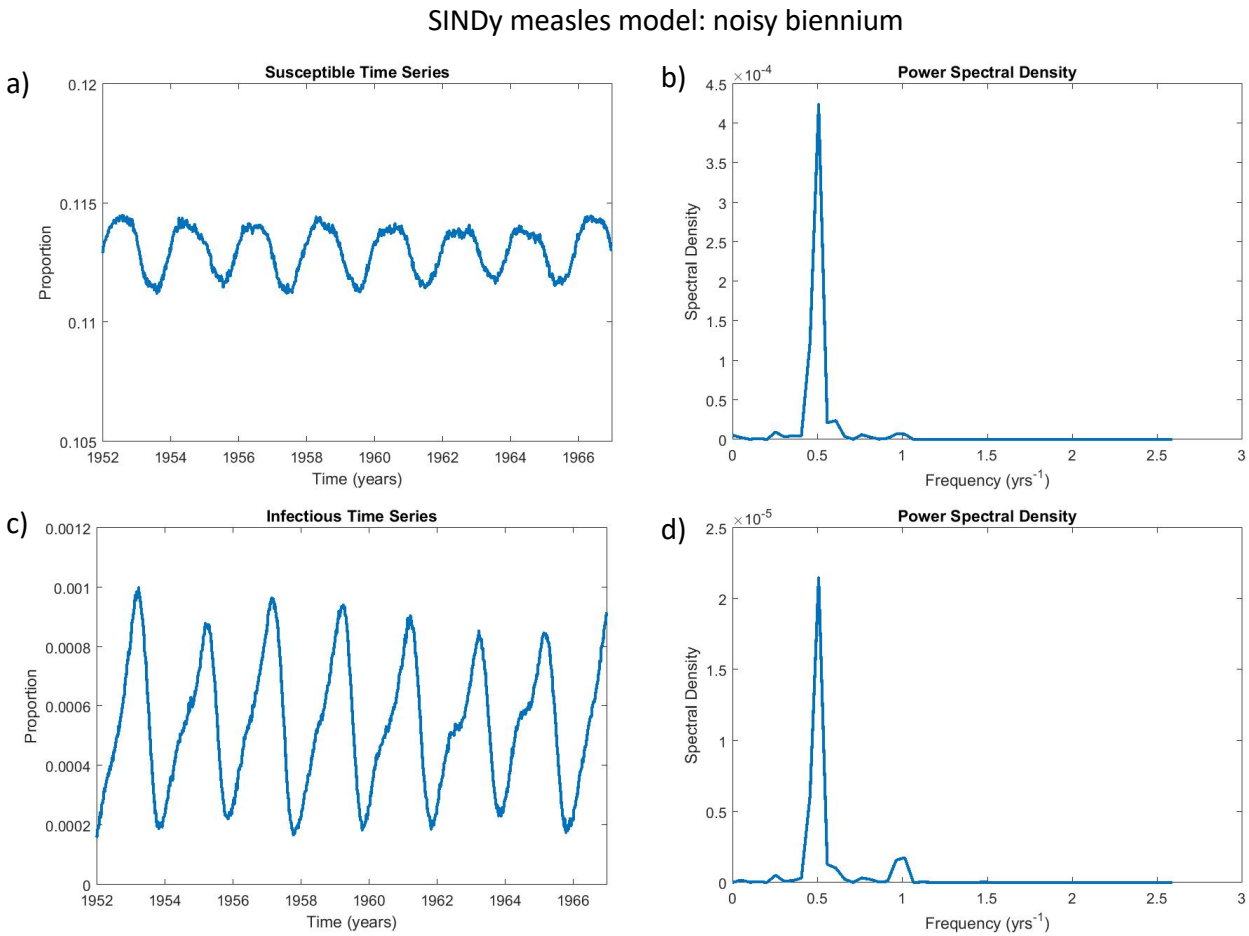


Figure 8. The SINDy measles biennium is robust to the addition of noise. This figure depicts the simulated timeseries of the SINDy measles model under additive noise: subpanels show the proportion of susceptible (a) and infected (c) individuals over time and the corresponding power spectral density plots for the susceptible (b) and infectious (d) time series. The power spectral density plots show strong power at a frequency of $0.5/\text{year}$ and a lesser peak at $1/\text{year}$, corresponding to a prominent biennial cycle. White noise with a coefficient of 1.5×10^{-3} was added to the right-hand side of the SINDy-discovered system of differential equations to generate these plots. See Methods for details about computation of the power spectral density.

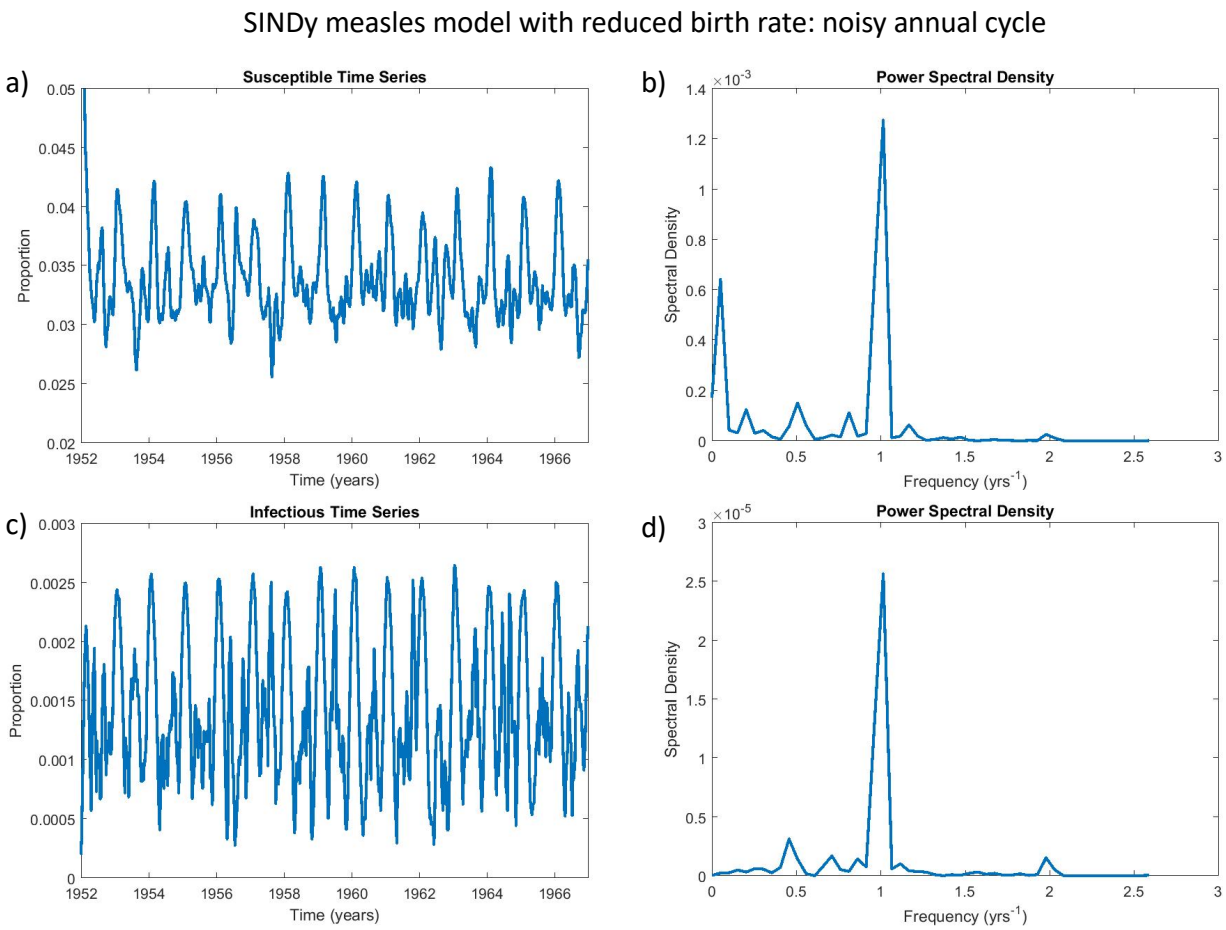


Figure 9. The SINDy measles model predicts a noisy annual attractor when the recruitment rate is reduced. This figure depicts the simulated timeseries of the SINDy measles model with a reduced birth rate, under additive noise: subpanels show the proportion of susceptible (a) and infected (c) individuals over time and the corresponding power spectral density plots for the susceptible (b) and infectious (d) time series. The power spectral density plots show strong power at a frequency of $0.5/\text{year}$ and a lesser peak at $1/\text{year}$, corresponding to a prominent biennial cycle. White noise with a coefficient of 1.5×10^{-3} was added to the right-hand side of the SINDy-discovered system of differential equations to generate these plots. See Methods for details about computation of the power spectral density. To simulate a reduced recruitment rate of newly-born susceptible individuals from 2.60/year to 1.36/year (expressed as the total fertility rate of susceptible offspring) between 1948-1967 and 1968-1988 in the United Kingdom due to falling birth rates and mass vaccination [43] the coefficient of S was changed from 0.606 (Figure 3) to 0.317.

Discussion

Model discovery provides a way of generating models inductively from data, using minimal prior knowledge about the system. This differs from the deductive approach that currently dominates model development in most fields, including theoretical epidemiology. Here we demonstrated that Sparse Identification of Nonlinear Dynamics (SINDy) can discover dynamical system models from empirical data on childhood infectious diseases from the pre-vaccine era. These inductively derived models (1) reproduce the observed dynamics of measles, chickenpox and rubella, (2) recover prominent features of deductively derived models, such as the fundamental mechanisms of mass-action mixing and seasonal forcing that underpin decades of research in theoretical epidemiology, (3) confirm the important role of seasonal variation in the transmission rate for dynamics of childhood infectious diseases, and (4) can predictregime shifts in dynamical patterns for measles, out-of-sample. Hence, our results show that model discovery methods could lend insight to both model creation as well as understanding of epidemiological mechanisms. Because the approach to developing the dynamic models is fundamentally different, it raises the possibility that real-world mechanisms could be discovered that would otherwise be difficult or impossible to find through conventional deductive modelling.

Limitations to the approach stem from data quality and availability, regression algorithms and criterion for selecting parameters like the sparsity threshold. Case notifications are typically under-reported even when they are available, and our approach requires reconstruction of the susceptible time series, which necessitates making assumptions [44, 38]. SINDy may be sensitive to the method used to reconstruct the susceptible time series, and these methods might also bias SINDy toward selecting certain terms. Unfortunately, susceptible reconstruction is necessary due to the lack of high-quality longitudinal serological data, but a careful sensitivity analysis in future work could help us better understand the impact of differing methods of susceptible reconstruction.

Applying SINDy to vaccine era data will require further thought because the vaccination in the 20th century changed the dynamics of childhood infectious disease dramatically [44, 45, 27, 25] and introduced the additional dimension of human behaviour through vaccine decision-making [46, 47, 48]. Fortunately, data are increasingly abundant in an era of digital data, open sharing, and online social media [49, 50], although this does not necessarily translate into higher quality data. And, the risk of overfitting to noise is always present and will only increase as more data becomes available [51]. In this case, we tested whether the model was over-fitted by analysing its out-of-sample prediction of empirical measles dynamics. However, over-fitting could also be tested varying the window of the moving average filter or subsampling the data before filtering.

The sensitivity of SINDy to noise (despite improvements over previous algorithms in this respect) is another limitation, and was illustrated in our model rediscovery subsection. Re-discovered models may describe the datasets very well in the presence of noise, but the spurious inclusion of terms that do not appear in the original model suggests caution when interpreting models discovered from noisy empirical data. We hypothesize that noise is also the cause of unexpected terms in our models discovered from empirical data, especially for the results using a third-order library. Part of the solution to this problem will be improved regression approaches that are less likely to overfit noise. However, the other part of the solution is inevitably the thoughtful application of subject expertise by humans when interpreting discovered models.

Another limitation is the choice of functional basis. This might be the largest limitation, since human bias is introduced when formulating the function library that SINDy starts with [10]. Our choice of seasonally varying transmission rates and polynomial functions may be limiting the discovery of a more parsimonious and interpretable model based on other terms, at least in principle. The state variables themselves are also pre-defined and may not be the best choice to recover the dynamics of the data.

There remains much opportunity 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 [52, 53] which incorporate functions that extend beyond a simple polynomial basis constructed from state variables. Specifically, a general nonlinear incidence (a transmission function of the form $S^p I^q$, where $p, q > 0$) could be explored. 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 optimal [25]. While an alternative based on forcing from the school year calendar was tested in our research, further analysis in this area is necessary.

In conclusion, we have shown that SINDy can be applied to epidemiological data to yield sparse models that describe observed epidemic patterns and correspond well with canonical deductive models from the past century. Although inductive model discovery methods come with their own set of challenges and limitations, their radically different approach to model generation suggests they can form a powerful complement to traditional modelling approaches, thereby improving our scientific understanding for many natural systems.

Material and Methods

Sparse Identification of Nonlinear Dynamics (SINDy)

This work builds on the sparse regression methods outlined in Ref. [10]. Given the recent advances in both compressed sensing [54, 55, 56] and sparse regression [57, 7] 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

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}(t), t), \tag{5}$$

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. [10] is total variation regularization [58, 59] 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

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

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

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

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} 1 & \boldsymbol{X} & \boldsymbol{X}^{P_2} & \boldsymbol{X}^{P_3} & \ldots & \sin(\boldsymbol{X}) & \cos(\boldsymbol{X}) & \sin(\boldsymbol{2X}) & \cos(\boldsymbol{2X}) & \ldots \end{bmatrix}, \tag{6}$$

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

$$\dot{\boldsymbol{X}} = \Theta(\boldsymbol{X})\Xi, \tag{7}$$

where $\Xi = (\xi_1, \xi_2, \ldots, \xi_p)$ 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) [57, 7], 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. [7].
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. [7]. We can then reconstruct the kth row of the dynamical system by taking

$$\dot{\boldsymbol{x}}_k = \Theta(\boldsymbol{x}_k^T)\xi_k, \tag{8}$$

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

$$\dot{\boldsymbol{x}} = \Xi^T \Theta(\boldsymbol{x}^T)^T. \tag{9}$$

The code for this algorithm, along with several examples that demonstrate its application, can be found at Ref. [60]. The modified repository used for all computation done can be found at Ref. [61].

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 [44, 62, 63]. 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 . When using up to 2nd order polynomials, the function library used was

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

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

$$\Theta(\boldsymbol{X}) = [\boldsymbol{1} \quad \boldsymbol{S} \quad \boldsymbol{I} \quad \boldsymbol{S^2} \quad \boldsymbol{I^2} \quad \boldsymbol{SI} \quad \boldsymbol{S^3} \quad \boldsymbol{S^2I} \quad \boldsymbol{SI^2} \quad \boldsymbol{I^3} \quad \boldsymbol{\beta} \quad \boldsymbol{\beta S} \quad \boldsymbol{\beta I} \quad \boldsymbol{\beta S^2} \quad \boldsymbol{\beta I^2} \quad \boldsymbol{\beta SI} \quad \boldsymbol{\beta S^3} \quad \boldsymbol{\beta S^2I} \quad \boldsymbol{\beta SI^2} \quad \boldsymbol{\beta I^3}]^{'}$$

where β is the seasonally-varying transmission rate given in Eq. 22. When the model coefficients are given in figures describing SINDy-discovered models, this parameter is represented by B.

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 [64, 65, 66]. 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) [39] and is derived from use of maximum likelihood. The AIC value for a given candidate model i is defined by

$$AIC_i = 2k - 2\ln(L(\mathbf{x}, \hat{\mu}), \tag{10}$$

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.

Data Sources and 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 1952-1967 (from Ref. [67]), chickenpox in Ontario (Canada) from 1946-1967, and rubella in Ontario from 1946-1960 (both from [27]). These diseases and time periods were chosen as they exhibit contrasting dynamic behaviour, most notably in the period of the epidemic cycle. For each of these diseases, the time frame chosen is before the vaccines for the respective diseases became commonly available. The raw data were also smoothed using a Savitzky-Golay filter [68] of order 3 with a window length of 19 in order to reduce the risk of SINDy overfitting the data, resulting in the smoother time series shown in Figs. 2-6.

Once this data was imported and both the time and case vectors were labelled, both the birth and population data (taken from [67, 69, 70, 71, 72]) were imported and linearly interpolated to be given per week, the same scale as the disease data. Time series of birth rates for Ontario and the United Kingdom appear in Supplementary Figure 38.

Discrete Time Model

In the limit as $\Delta t \to 0$ the discrete model (Eqs. 1 - 3) converges to the continuous-time compartmental SIR model [31, 27]. Applying SINDy to discover a continuous-time model involves determining the derivative vector $\dot{\boldsymbol{x}}(\boldsymbol{t}) = \langle \dot{S}(t), \dot{I}(t), \dot{R}(t) \rangle$. 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. Hence, we used the discrete-time SIR model for our analysis.

Solving Eq. 4 for β and iterating over empirical data for measles, chickenpox and rubella give the time series found in Supplementary Figure 25. Given that each of these diseases is most common among school-aged children [44, 73, 45] 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. [73, 74], though more analysis by Ref. [44] indicate the peak in transmission rate occurs several weeks earlier, alleging this effect may also be attributed to weather fluctuations.

Susceptible Reconstruction

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

$$S_{t+1} = S_t - \alpha C_{t,t+1} + B_{t,t+1}, \tag{11}$$

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) [44]. 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 [63] as well as changing temporally [38]. It is also difficult to estimate explicitly, due to the lack of serological data available.

An extension of this method is derived in Ref. [38]. They assume the discrete relation

$$S_{t+1} = S_t - \alpha_t C_{t,t+1} + B_{t-d,t-d+1} + u_t, \tag{12}$$

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.

$$S_t = \bar{S} + Z_t. \tag{13}$$

By substituting Eq. [13] into Eq. [12] we see that Z_t also satisfies the relation

$$Z_{t+1} = Z_t - \alpha_t C_{t,t+1} + B_{t-d,t-d+1} + u. \tag{14}$$

Iterating this expression results in the relation

$$Z_t = Z_0 - \sum_{i=1}^{t} \alpha_i C_{i,i+1} + \sum_{i=1}^{t} B_{i-d,i-d+1} + \sum_{i=1}^{t} u_i \tag{15}$$

Ref. [38] uses 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. [15] to

$$Z_t = Z_0 - \bar{\alpha} X_t + Y_t - R_t + U_t. \tag{16}$$

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

$$Y_t = \bar{\alpha} X_t + (Z_t - Z_0). \tag{17}$$

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 $\bar{\alpha}$.

We call this the 'global regression method'. Applying this reconstruction method yields time series of of the proportion of susceptible individuals for the data in Figure 1 (Supplementary Figure 39). From these figures it can be seen that each reconstruction (especially for the chickenpox and rubella case notification data) suffers from local shifts in the mean, caused by the assumption that the reporting rate is temporally invariant. Ref. [38] correct for this by assuming that the dominant fluctuations in Eq. 16 are caused by variation in the reporting rate α_t rather than in external noise (u_t). Eq. 16 can then be expressed as

$$Y_{t+1} = R_t - U_t Z_0 - (\alpha_{t+1} - \bar{\alpha})X_t + \alpha_{t+1}X_{t+1} + Z_{t+1} - u_{t+1}. \tag{18}$$

Local linear regression techniques can then be applied to estimate both the reporting rate and the susceptible class. This method is sensitive to the bandwidth parameter, and must be tuned beforehand to minimize large-scale fluctuations from the global mean. We use this 'locally linear regression method' [38] for all of the results reported in our paper. The susceptible reconstruction time series from the incidence data in Figure 1 appears in Supplementary Figure 40, and the resulting transmission rate reconstruction time series also appears in Supplementary Figure 41.

Incidence to Prevalence Conversion

The *prevalence* of the disease is defined by the number (or proportion) of infectious individuals at any given time. Compartmental epidemic models usually predict infection prevalence. However, data are usually in the form of newly occurring cases, referred to as *incidence*. Hence both the proportion of susceptible individuals and the prevalence of infection must be recovered from the infection incidence data before the SINDy algorithm can be applied.

Given temporal case incidence 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

$$\langle P_t \rangle = \frac{pD_i}{L} \tag{19}$$

From the relation

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

we then obtain

$$P_t = \frac{C_t p D_i}{\langle C_t \rangle L} \tag{20}$$

which is used to construct the prevalence (infectious) class given incidence data. We assumed $D_i = 2$ weeks for all i [75, 27, 25], $L = 65$ years [76] and $p = 0.95$ [75].

Weighted Thresholding

SINDy is based on 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. [10] is the iterated thresholding method, outlined in Section . 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. [10] and subsequent papers this parameter is taken as constant, though Ref. [11] 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 (the proportion susceptible, for instance, is $O(0.1)$ while the proportion infected is $O(0.0001)$. 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

$$\lambda_w^{(k)} = \frac{\lambda_c}{|\Theta^{(k)}(\boldsymbol{X})|}, \tag{21}$$

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 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. 21.

Choice of 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 [10]. 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. The dynamics of the prevalence of both measles and chickenpox are strongly dictated the seasonal forcing function [27, 25]. Hence, a new parameter β is constructed such that

$$\beta = \beta_0(1+\beta_1\cos(2\pi t/T-\phi)), \tag{22}$$

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.

Model dynamics are sensitive to ϕ. Hence, we included ϕ in a three-dimensional parameter sweep involving ϕ, S_0, and λ as follows. For each point of the $S_0-\lambda$ parameter grid ($\lambda \in [0.0001, 0.1]$, $S_0 \in [0.05, 0.13]$, see Supplementary Figures 5-12), we also conducted a parameter sweep for ϕ ranging from 0 to 52 weeks in increments of $\Delta\phi = 0.5$wk. The SINDy model was generated for each value of ϕ in this parameter range, and the SINDy model with lowest AIC values was selected to represent that point in the $S_0-\lambda$ parameter grid. (We opted for this approach to facilitate visualization of the three-parameter sweep in the two-dimensional $S_0-\lambda$ plane and to focus on the $S_0-\lambda$ relationship. Altering the values of ϕ away from the optimal values usually worsens the model fit in predictable ways. For instance, a value of ϕ that causes transmission to peak in the summer months also causes an epidemic peak in the summer, which is rarely observed in the empirical datasets.)

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 .

Power Spectral Density

Estimates of the power spectral density of the prevalence time series can be useful for model selection when qualitative features such as attractor class are seen as more important than goodness-of-fit. However, a model selection method that values sparsity is still desirable. 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. We followed standard procedures for computing power spectra of time series [77, 78, 27]. First, the infection time series data from both the discovered SINDy model and the relevant data were smoothed using a moving average window with a span of 23 timesteps. Second, the smoothed data were linearly trend-corrected. Third, 20 % of the time series was tapered using a split cosine bell. The periodogram of the resulting time series was then computed. We used the Matlab functions smooth, detrend, and periodogram [79]. The AIC score was then computed using the residual sum of squares between the empirical and SINDy model power spectral density estimates, taking the number of free parameters from the discovered model. The methodology for computed the power spectral density described in the out-of-sample prediction subsection was the same.

SIR Model Fitting

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. [25, 75]. 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 $\mathscr{R}_0$ and recovery rate γ, the mean transmission rate is completely determined by $\beta_0 = \mathscr{R}_0 \cdot \gamma$. The parameters that were varied, with corresponding ranges and step sizes, were $R_0 \in (6, 16)$, $\gamma \in (0.55, 1.25)$, $\beta_1 \in (0.05, 0.35)$, $\mu \in (3\times10^{-4}, 6\times10^{-4})$ and $\in (0, 51.5)$. 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 $(S_0, I_0) = (0.1, 5\times10^{-5})$. The resulting models with the minimal sum of squares error when compared with the data were selected and plotted in Supplementary Figures 42-44.

Data Availability

The code used to generate the results is publicly available at Ref. [61]. The infectious disease data for measles, rubella and chickenpox can be obtained from the International Infectious Disease Data Archive (http://iidda.mcmaster.ca/). Demographic

data are available from published sources [67, 69, 70, 71, 72].

References

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

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

[3] N. H. Packard et al. "Geometry from a Time Series". In: *Phys. Rev. Lett.* 45 (9 Sept. 1980), pp. 712–716.

[4] James P Crutchfield and Bruce S McNamara. "Equation of motion from a data series". In: *Complex systems* 1.417-452 (1987), p. 121.

[5] Bryan C Daniels and Ilya Nemenman. "Automated adaptive inference of phenomenological dynamical models". In: *Nature communications* 6 (2015), p. 8133.

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

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

[8] 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.

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

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

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

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

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

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

[15] 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.

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

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

[18] Niall M Mangan et al. "Model selection for hybrid dynamical systems via sparse regression". In: *arXiv preprint arXiv:1808.03251* (2018).

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

[20] Daihai He, Edward L Ionides, and Aaron A King. "Plug-and-play inference for disease dynamics: measles in large and small populations as a case study". In: *Journal of the Royal Society Interface* (2009).

[21] Lars Folke Olsen and William Morris Schaffer. "Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics". In: *Science* 249.4968 (1990), pp. 499–504.

[22] S. P. Ellner, B. A. Bailey, and G. V. Bobashev. "Noise and Nonlinearity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling". In: *The American Naturalist* 151.5 (1998), pp. 425–440.

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

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

[25] David J. D. Earn et al. “A Simple Model for Complex Dynamical Transitions in Epidemics”. In: *Science* 287.667 (2000).

[26] Pejman Rohani, David JD Earn, and Bryan T Grenfell. “Opposite patterns of synchrony in sympatric disease metapopulations”. In: *Science* 286.5441 (1999), pp. 968–971.

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

[28] Matthew J Ferrari et al. “The dynamics of measles in sub-Saharan Africa”. In: *Nature* 451.7179 (2008), p. 679.

[29] W. O. Kermack and A. G. McKendrick. “A contribution to the mathematical theory of epidemics”. In: *Proc. R. Soc. Lond. A* 115.772 (1927).

[30] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[31] Herbert W Hethcote. “The mathematics of infectious diseases”. In: *SIAM review* 42.4 (2000), pp. 599–653.

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

[33] Niall M Mangan et al. “Model selection for hybrid dynamical systems via sparse regression”. In: *Proceedings of the Royal Society A* 475.2223 (2019), p. 20180534.

[34] Zhilu Lai and Satish Nagarajaiah. “Sparse structural system identification method for nonlinear dynamic systems with hysteresis/inelastic behavior”. In: *Mech. Sys. & Sig. Proc.* 117 (2019), pp. 813–842.

[35] Mariia Sorokina, Stylianos Sygletos, and Sergei Turitsyn. “Sparse identification for nonlinear optical communication systems: SINO method”. In: *Optics express* 24.26 (2016), pp. 30433–30443.

[36] Magnus Dam et al. “Sparse identification of a predator-prey system from simulation data of a convection model”. In: *Physics of Plasmas* 24.2 (2017), p. 022310.

[37] J.-C. Loiseau and S. L. Brunton. “Constrained Sparse Galerkin Regression”. In: *Journal of Fluid Mechanics* 838 (2018), pp. 42–67.

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

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

[40] Neil M Ferguson, Christl A Donnelly, and Roy M Anderson. “Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain”. In: *Nature* 413.6855 (2001), p. 542.

[41] Wei-min Liu, Herbert W. Hethcote, and Simon A. Levin. “Dynamical behavior of epidemiological models with nonlinear incidence rates”. In: *Journal of mathematical biology* 25.4 (1987), pp. 359–380.

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

[43] WHO. URL: http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tswucoveragemcv1.html.

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

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

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

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

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

[49] Marcel Salathe et al. "Digital epidemiology". In: *PLoS computational biology* 8.7 (2012), e1002616.

[50] A Demetri Pananos et al. "Critical dynamics in population vaccinating behavior". In: *Proceedings of the National Academy of Sciences* (2017), p. 201704093.

[51] David Lazer et al. "The parable of Google Flu: traps in big data analysis". In: *Science* 343.6176 (2014), pp. 1203–1205.

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

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

[54] David L Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

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

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

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

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

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

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

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

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

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

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

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

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

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

[68] Ronald W Schafer et al. "What is a Savitzky-Golay filter". In: *IEEE Signal processing magazine* 28.4 (2011), pp. 111–117.

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

[70] *200 years of the Census in Wales*. URL: https://web.archive.org/web/20090319202324/http://www.statistics.gov.uk/census2001/bicentenary/pdfs/wales.pdf.

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

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

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

[74] H. E. Soper. "The Interpretation of Periodicity in Disease Prevalence". In: *Journal of the Royal Statistical Society* 92.1 (1929), pp. 34–73.

[75] Roy M. Anderson and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[76] George W Leeson. "Increasing longevity and the new demography of death". In: *International Journal of Population Research* 2014 (2014).

[77] Peter J Brockwell, Richard A Davis, and Stephen E Fienberg. *Time Series: Theory and Methods: Theory and Methods*. Springer Science & Business Media, 1991.

[78] Maurice Bertram Priestley. *Spectral analysis and time series*. Vol. 1. Academic press London, 1981.

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

Acknowledgments

The authors are grateful to David J.D. Earn for providing data for this project from the International Infectious Disease Data Archive (IIDDA), and to Sri Namachivaya, Giang Tran, and two anonymous reviewers for helpful comments.

Author contributions statement

CTB conceived the study. CTB and JH developed the methodology. JH conducted the analysis and drafted the manuscript. CTB revised the manuscript.

Additional information

The authors declare no competing interests.