PLOS Computational Biology

RESEARCH ARTICLE

Learning structured population models from data with WSINDy

Rainey Lyons*, Vanja Dukic, David M. Bortz

Department of Applied Mathematics, University of Colorado, Boulder, Colorado, United States of America

* rainey.lyons@colorado.edu

Abstract

Characteristics of individuals in a population, such as age and size, play a key role in determining how populations change over time. In contexts of population dynamics, identifying effective model features, such as fecundity and mortality rates, is generally a complex and computationally intensive process, especially when the dynamics are heterogeneous across the population. In this work, we propose a Weak form Scientific Machine Learning-based method for selecting appropriate model ingredients from a library of scientifically feasible functions used to model structured populations. This paper presents extensions of the Weak form Sparse Identification of Nonlinear Dynamics (WSINDy) method to select the best-fitting ingredients from noisy time-series histogram data. This extension includes learning heterogeneous dynamics and also learning the boundary processes (such as birth) of the model directly from the data. We additionally incorporate a cross-validation method which helps fine tune the recovered boundary process hyperparameters to the data.

Several test cases are considered, demonstrating the method's performance for several standard models from population modeling, including age and size-structured models. Through these examples, we examine both the advantages and limitations of the method, with a particular focus on the distinguishability of terms in the library.

Author summary

Physiological characteristics of individuals, such as age and size, play a key role in determining how populations change over time. Developing effective mathematical models to describe the population dynamics requires determining how vital rates, such as mortality and fertility rates, depend on the individuals' state. In this work, we propose a method for selecting these state-dependent rates from a library of scientifically plausible options, using time-series population data. Our approach builds on and adapts an existing Weak form Scientific Machine Learning technique, originally developed for discovering underlying dynamical systems from data. We test the method using both artificial data, generated from known models,

OPEN ACCESS

Citation: Lyons R, Dukic V, Bortz DM (2025) Learning structured population models from data with WSINDy. PLoS Comput Biol 21(12): e1013742. https://doi.org/10.1371/journal.pcbi.1013742

Editor: Jifan Shi, Fudan University, CHINA

Received: July 1, 2025

Accepted: November 13, 2025

Published: December 8, 2025

Data availability statement: All code used in this manuscript will be available on the repository github.com/MathBioCU/WSINDyStructuredPopulations.

Funding: The research reported in this publication was supported in part by the NIGMS Division of Biophysics, Biomedical

Technology and Computational Biosciences (grant R35GM149335 to DMB); NSF Division of Molecular and Cellular Biosciences MODULUS (grant 2054085 to DMB); NSF Division Of Environmental Biology (EEID grant DEB-2109774 to VD), and NIFA Biological Sciences (grant 2019-67014-29919 to VD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and real population data from a previous study. Through these case studies, we evaluate the method's ability to recover the correct model ingredients and predict the future population distribution, even when the data are heavily corrupted by noise. We also examine the limitations of the approach, particularly in cases where different candidate terms produce similar effects and are therefore difficult to distinguish.

1 Introduction

The study of structured population dynamics is an active field in mathematical biology where the resulting models and theory provide a rigorous framework for describing how individual-level traits, such as age, size, or physiological state, influence the dynamics of a population. Among these models, hyperbolic partial differential equations (PDEs) have proven particularly effective in contexts where one can model individuals' state progression as a deterministic transport process. This class of models arises naturally in applications ranging from cell growth and organism development [1,2] to disease progression and epidemiology [3,4].

Two of the most well-studied models in this setting are the age-structured model, also known as the McKendrick–von Foerster model [2,4], and the size-structured model, or Sinko–Streifer model [5]. These equations exemplify the typical hyperbolic structure of these models: transport and reaction processes, which make use of *growth*, *death*, and *birth* functions to encode both internal and external regulatory mechanisms. For instance, an age-structured model describes the evolution of the population number density, n, using the system

$$\begin{cases} \partial_t n + \alpha \partial_a n = -d[n](a)n, & (t,a) \in (0,T) \times (0,\infty) \quad (1a) \\ n(t,0) = \displaystyle\int_0^\infty \beta[n](a)n(t,a)\,\mathrm{d}a, & t \in (0,T), \quad (1b) \\ n(0,a) = n_0(a), & a \in \mathbb{R}^+, \quad (1c) \end{cases}$$

where $a \in \mathbb{R}^+ := [0,\infty)$ represents the age of individuals, and a size-structured version is given by

$$\begin{cases} \partial_t n + \partial_x(g[n](x)n) = -d[n](x)n, & (t,x) \in (0,T) \times (x_1,x_2), \quad (2a) \\ g[n](x_1)\,n(t,x_1) = \displaystyle\int_{x_1}^{x_2} \beta[n](x)n(t,x)\,\mathrm{d}x, & t \in (0,T), \quad (2b) \\ g[n](x_2)\,n(t,x_2) = 0, & t \in (0,T), \quad (2c) \\ n(0,x) = n_0(x), & x \in [x_1,x_2], \quad (2d) \end{cases}$$

where $x \in [x_1,x_2] \subset \mathbb{R}^+$ represents the size of individuals. Here, we use the notation $f[n]$ to denote a non-pointwise dependency on the population density which is usually nonlocal in nature. Details on the specific structure are provided in the next section.

In both the age structure and size structured setting, the accuracy and interpretability of the model depend heavily on the functional forms and parameters of the growth, death, and birth terms, denoted in the above equations by g, d, and β, respectively. Generally, these forms can be inferred by studies at the individual level, but in many instances such studies are expensive or infeasible, and one may not be able to obtain an effective dependence of the model ingredients on the structural variable [6,7]. Additionally, these vital components may depend directly on environmental variables such as the total population [8] or resource abundance [9], which tend to be processes that cannot be directly measured and whose functional form is unknown and is usually taken as an *ansatz*.

Classical approaches to model calibration rely on specifying parametric forms for these functions and estimating parameters by minimizing the discrepancy between simulated and observed population dynamics (see, for instance, [10] and the references therein). While effective, this approach is computationally demanding, as it requires repeated forward simulations of the PDE system. Furthermore, parameter estimation problems are often ill-posed, especially in the presence of noise, limited data, or structural uncertainty in the model. Recently, Weak form Scientific Machine Learning (WSciML) methods such as the Weak form Sparse Identification of Nonlinear Dynamics (WSINDy) – a WSciML extension of the well known Sparse Identification of Nonlinear Dynamics [11,12] for equation discovery. [13,14], and the the Weak-form Estimation of Nonlinear Dynamics (WENDy) [15,16] have emerged as promising alternatives for learning governing equations and estimating parameters directly from data. These methods bypass the need for repeated forward simulations by instead minimizing an equation error residual over all possible combinations of candidate terms in the library. In particular, the weak-form algorithms, WSINDy and WENDy, have been shown to be robust to noise, while retaining high accuracy and computational efficiency (for a general overview of weak-form methods see e.g., [17,18]).

In this work, we apply the WSINDy framework to the discovery of hyperbolic structured population equations. Given noisy population data, our goal is to identify effective model components from a library of biologically plausible functions. From the list of aforementioned methods, WSINDy is the most natural choice of method for the considered problem, as structured population models are commonly studied in a weak sense (as smoothness is not always guaranteed [19–21]). We demonstrate the performance of this approach on both synthetic and real data, showing that WSINDy recovers relevant and effective dynamics with significantly reduced computational effort compared to traditional parameter estimation methods.

Finally, we highlight several novel aspects of our work. Previous implementations of the WSINDy algorithm have focused primarily on discovering homogeneous, pointwise nonlinearities of the form $f(n)$. Here, we explore the method's ability to identify both heterogeneous model ingredients and boundary processes directly from the data. To our knowledge, both of these extensions are yet unexplored in the context of weak form model discovery. Additionally, we consider nonlocal nonlinearities of the form $f(s, N(t))$, which are ubiquitous in structured population models and capture well-known density-dependent effects such as those found in logistic-type nonlinearities, Ricker suppression terms, and Beverton-Holt-type population effects [8,22,23]. These features introduce new challenges that we address in the sections that follow.

The manuscript is organized as follows: in Sect 2, we present the modified WSINDy method for a general structured population model and include details in how to incorporate the identification of the boundary process. In Sect 3, we discuss the assumptions made on the data and present an array of test problems which are focused on the two most popular structured population models (1) and (2). In Sect 3.2.1, we explore the performance of the method, including how the method performs in high noise cases and how well the method can distinguish between similar terms in the library. Finally, in Sect 4 we conclude with a discussion of the method and directions for future exploration.

2 Models and methods

In this section, we present the general framework of the manuscript. We begin by introducing a general structured population model which encompasses the well-studied models given by Eqs (1) and (2) as well as other commonly seen structured population models. We then present the method applied to this general model and describe a validation procedure used to improve the recovery of the boundary terms.

2.1 A general structured population model

Throughout the manuscript, we assume that the noise-free data follows underlying dynamics governed by a hyperbolic structured population equation. To distinguish between noisy and clean data, we make use of the superscript $\star$ to denote the noise-free population density and true model ingredients. In such models, the population is distributed over some structural variable (such as age, size, or a combination with other physiological traits), and its distribution is denoted by $n^\star(t,s)$, where $t \in [0,T]$ denotes time and $s \in \Omega \subseteq \mathbb{R}^d$ denotes an arbitrary structural variable, assumed to lie in a connected domain Ω with boundary $\partial\Omega = \partial\Omega^+ \cup \partial\Omega^-$. Another natural interpretation of this density $n^\star$ commonly found in ecology is that the number of individuals whose structure lies in a measurable subset $S \subset \Omega$ at time t is given by the quantity $\int_S n^\star(t,s)\,\mathrm{d}s$. (Note that the number density integrates to the total population size, $N(t) = \int_\Omega n(t,s)\,\mathrm{d}s$.) While the rest of this work focuses primarily on the most common case $d = 1$, we note that there is no inherent restriction of the method to one-dimensional models.

The following system gives a general form of the dynamics:

$$\begin{cases} \partial_t n^\star(t,s) + \nabla_s \cdot (g^\star[n^\star](s)\, n^\star(t,s)) = f^\star[n^\star](s,n^\star), & (t,s) \in (0,T)\times\Omega, \quad (3a)\\ g^\star[n^\star](s)\, n^\star(t,s)\cdot\vec{\eta}(s) = \displaystyle\int_\Omega \beta^\star[n^\star](\sigma) n^\star(t,\sigma)\,\mathrm{d}\sigma, & (t,s) \in [0,T]\times\partial\Omega^+, \quad (3b)\\ g^\star[n^\star](s)\, n^\star(t,s)\cdot\vec{\eta}(s) = 0, & (t,s) \in [0,T]\times\partial\Omega^-, \quad (3c)\\ n^\star(0,s) = n_0^\star(s), & s \in \Omega. \quad (3d) \end{cases}$$

In the equations above, $\vec{\eta}(s)$ denotes the inward-pointing unit normal vector at the boundary point s. The model terms $g^\star$, $f^\star$, and $\beta^\star$ represent biological processes which modeled via transport, source, and boundary terms. Common examples of such processes are growth/aging, mortality or division, and reproduction, respectively. These terms will be assumed to be smooth or at least globally Lipschitz in all arguments to be consistent with the well-posedness theory for Eq (3) (see, e.g., [21]). The notation $f[n]$ denotes a non-pointwise dependence on the population density, typically through a weighted average of the population. For instance, $f[n](\cdot) = f(\cdot, \int_\Omega \gamma(s) n(t,s)\,\mathrm{d}s)$ for some known weight function γ. While many types of kernels are biologically relevant, well-studied, and have many interesting mathematical properties [19,24,25]; in this manuscript, we focus on the most common biologically relevant nonlinearity (where $\gamma \equiv 1$), meaning that the model ingredients depend on the total population size $N(t) = \int_\Omega n(t,s)\,\mathrm{d}s$.

As one of the crucial components of the WSINDy method is the eponymous *weak form*, integration over both the temporal and structural variables will play a vital role. To streamline notation, we define the L^2 inner product over time and space by

$$\langle f, g\rangle := \int_0^T \int_\Omega f(t,s)\cdot g(t,s)\,\mathrm{d}s\,\mathrm{d}t.$$

We will also abuse this notation slightly to represent discrete approximations of this inner product in the same way.

The weak form of Eq (3) is then given by:

$$-\langle \partial_t \phi, n^\star \rangle - \langle \nabla_s \phi, g^\star[n^\star] n^\star \rangle = \langle \phi, f^\star[n^\star](\cdot, n^\star) \rangle, \tag{4}$$

where $\phi(t, s)$ is a smooth real-valued function compactly supported in $(0, T) \times \Omega$, i.e, $\phi \in C_c^p((0, T) \times \Omega)$ for some $p \geq 2$. The test function ϕ plays a role similar to a Gaussian smoother, while also maintaining the quantitative relationship given by Eq (4), allowing us to simultaneously smooth the data and utilize equation error methods. The smoothness and compact support of the test function allows us to exploit the rapid convergence of the trapezoidal rule [14, Lemma 2] allowing for highly accurate computations of the integrals appearing in the weak form. However, one down side of these test functions is that the boundary conditions (3b)-(3c) and the initial condition (3d) are absent from Eq (4). This presents a challenge when attempting to identify the boundary process (commonly representing birth), a critical component in describing the population dynamics. To address this, we provide a method for identifying the boundary process in the following section.

2.2 Weak Sparse Identification of Nonlinear Dynamics (WSINDy)

In this section, we extend the WSINDy algorithm to accommodate structured populations. This extension is novel as it enables the identification of heterogeneous dynamics and boundary processes directly from the given data. In the subsequent section, we introduce a cross-validation procedure designed to leverage the accuracy of the learned boundary processes for hyperparameter tuning.

Let $\mathbf{n} := \{n_j\}_{j=1}^J$ denote (after a suitable indexing) a set of possibly noisy observations of the number density over a disjoint partition $\{\Lambda_j\}_{j=1}^J \subset \Omega$, i.e., $n_j := \frac{1}{|\Lambda_j|} \int_{\Lambda_j} n(t_j, s)\, \mathrm{d}s$. As stated before, we assume the noise-free continuous density $n^\star$ evolves according to a true model of the form (3). The WSINDy algorithm utilizes the weak form (4) to construct a sparse regression problem, selecting the proper model ingredients from a set of given trial functions, known as the "library". More precisely, we assume the true model ingredients $g^\star$, $f^\star$, and $\beta^\star$ can be represented as a (sparse) linear combination of a given set of trial functions. The set of these functions is denoted by $\{g_m(s, N)\}_{m=1}^{M^g}$, $\{f_m(s, n, N)\}_{m=1}^{M^f}$, and $\{\beta_m(s, n, N)\}_{m=1}^{M^\beta}$, respectively. Then, for a given set of test functions $\{\phi_k\}_{k=1}^K \subset C_c^p((0, T) \times \Omega)$, we construct the linear system

$$\mathbf{b}^{\text{pde}} = G^{\text{pde}} \mathbf{w} = (G^g | G^f) \begin{pmatrix} \mathbf{w}^g \\ \mathbf{w}^f \end{pmatrix}, \tag{5}$$

where the vector $\mathbf{b}^{\text{pde}} \in \mathbb{R}^{K \times 1}$ and matrix $G^{\text{pde}} \in \mathbb{R}^{K \times (M^g + M^f)}$ are given by

$$\mathbf{b}_k^{\text{pde}} := -\langle \partial_t \phi_k, n \rangle, \quad G_{k,m}^g := \langle \nabla_s \phi_k, g_m(\cdot, N) n \rangle, \text{ and } G_{k,m}^f := \langle \phi_k, f_m(\cdot, n, N) \rangle.$$

As for the choice of test functions, we opt to make use of piecewise-polynomial test functions of the form $\phi(t, s) = \varphi(t) \prod_{i=1}^{\mathrm{d}} \psi_i(s_i)$, where

$$\varphi(t) = \begin{cases} C_t (t - t_1)^p (t_2 - t)^q, & t \in (t_1, t_2) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi_i(s_i) = \begin{cases} C_i (s_i - a_i)^p (b_i - s_i)^q, & s_i \in (a_i, b_i) \\ 0 & \text{otherwise} \end{cases}$$

The constants C_t and the C_i are chosen such that $\|\varphi\|_\infty = \|\psi_i\|_\infty = 1$. The supports of the test functions are determined from the data in such a way that the intervals $[a_i, b_i]$ or $[t_1, t_2]$ account for a given percentage of the full domain (denoted by r_s and r_t, respectively). For the numerical results to follow, we will set $p = q = 14$ and $r_s = r_t = 0.5$. These choices were determined through numerical experiments, and properly choosing these parameters from the data is currently a frontier of active research. We point out that while there exist methods for choosing the support and smoothness of test functions from the given data (such as, e.g., [13,26]), these methods are not tuned for heterogeneous libraries such as those considered here and we find that these methods result in radii which are too small to fully distinguish the terms in the library. However, from the numerical experiments (see for example S2 Fig) we see that there is a relatively large set of pairs (r_t, r_x) which can recover the proper dynamics in the presence of noise. Additionally, while this class of test functions is commonly used, we make no claims that it is indeed the best choice.

As discussed in the previous section, since this choice of test function is compactly supported in Ω, the boundary condition (3b) is absent from the weak form (4). One natural way to account for the boundary is to couple Eq (3) with the dynamics of the total population $N(t) := \int_\Omega n(t,s)\,\mathrm{d}s$ given by the ordinary differential equation acquired by integrating Eq (3):

$$\frac{\mathrm{d}}{\mathrm{d}t}N = \int_\Omega \beta^\star[n](s)n(t,s)\,\mathrm{d}s + \int_\Omega f^\star[n](s,n(t,s))\,\mathrm{d}s, \tag{6}$$

with the initial condition $N(0) = \int_\Omega n_0(s)\,\mathrm{d}s$. This equation then has the corresponding weak form

$$-\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\varphi(t)\,N(t)\,\mathrm{d}t = \langle\varphi, \beta^\star[n]n\rangle + \langle\varphi, f^\star[n](\cdot,n)\rangle, \tag{7}$$

for a smooth test function φ with compact support in $(0,T)$.

This allows us to extend or "stack" the linear system (5) by concatenating it with the linear system

$$\mathbf{b}^{\mathrm{ode}} = \Xi\mathbf{v} := \left(\Xi^f|\Xi^\beta\right)\begin{pmatrix}\mathbf{w}^f\\ \mathbf{w}^\beta\end{pmatrix}, \tag{8}$$

with $\mathbf{b}_k^{\mathrm{ode}} \in \mathbb{R}^{K\times 1}$ and $\Xi \in \mathbb{R}^{K\times(M^f+M^b)}$ given by

$$\mathbf{b}_k^{\mathrm{ode}} := -\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\varphi_k(t)N(t)\,\mathrm{d}t, \quad \Xi^f_{k,m} := \langle\varphi_k, f_m(\cdot,n,N)\rangle, \quad \text{and} \quad \Xi^\beta_{k,m} := \langle\varphi_k, \beta_m(\cdot,N)n\rangle.$$

We then concatenate the two systems, resulting, finally, in the linear system

$$\begin{pmatrix}\mathbf{b}^{\mathrm{pde}}\\ \mathbf{b}^{\mathrm{ode}}\end{pmatrix} =: \mathbf{b} = G\mathbf{w} := \left(\begin{array}{c|c|c} G^g & G^f & \mathbf{0}\\ \hline \mathbf{0} & \Xi^f & \Xi^\beta\end{array}\right)\left(\begin{array}{c}\mathbf{w}^g\\ \hline \mathbf{w}^f\\ \hline \mathbf{w}^\beta\end{array}\right). \tag{9}$$

The problem then becomes finding sparse $\mathbf{w} \in \mathbb{R}^{(M^g+M^f+M^\beta)\times 1}$ which minimizes the loss function

$$\mathcal{L}(\mathbf{x};\mathbf{b},G,\lambda) := \|\mathbf{b} - G\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_0, \tag{10}$$

where λ is a given sparsity parameter. The function $\mathcal{L}$ is minimized using the modified sequential-thresholding least-squares method (MSTLS) provided in [13], which has been successful in a variety of applications of sparse regression

techniques. To demonstrate the benefits of the MSTLS algorithm, we provide a small illustrative example of the performance of the method compared to ordinary least squares (OLS) in S4 Fig. This method has been studied for differential equation models of various types and contexts, including ordinary differential equations [14], partial differential equations [13,27], and hybrid models with multiple time scales [28].

Remark 2.1. *When studying population dynamics from an age-structured perspective, it is quite common to know a priori the age-time relationship α in Eq (1). In that case, one does not need to identify the transport term and can focus all efforts on the identification of the death and birth functions. One can easily modify the method presented above by simply adding the transport term into the vector $\boldsymbol{b}^{pde}$ in Eq (5). That is, Eq (5) would be given by*

$$\mathbf{b}_k^{\text{pde}} := -\langle \partial_t \phi_k, n\rangle - \langle \nabla_s \phi_k, \alpha n\rangle \text{ and } G_{k,m}^{\text{pde}} := \langle \phi_k, f_m(\cdot, n, N)\rangle.$$

2.2.1 Boundary bagging. While sparse regression on system (9) often yields accurate identification of PDE terms, in practice, the system tends to be dominated by its PDE component. This is due to the significantly greater number of test functions used in the PDE part compared to the ODE part, resulting in many more rows in system (5) than in system (8). As a consequence, the PDE residual becomes disproportionately weighted during optimization, causing the method to "push" errors into the boundary conditions. This typically leads to poor, and generally non-sparse, term selection performance in the boundary equations. Traditionally, this issue can be solved by a reweighting of the different components in the objective function. However for our method, we found extremely sharp behavior in the scaling coefficient where slightly different coefficients had drastic behaviors and caused the method to completely forgo the identification of either the PDE component or the ODE component. This made it difficult to systematically find the critical reweighting coefficient which resulted in a balanced identification of the full system.

To address this, we introduce a cross-validation procedure for the ODE term selection. This method is inspired by the library bagging technique used in Ensemble versions of SINDy and WSINDy [29] (see comments in the discussion in Sect 4). The difference here is in the method of selecting which terms to discard from the boundary process component of the library. Specifically, we fix the learned source weights $\mathbf{w}^f$ and apply sparse regression to the modified system

$$\mathbf{b}^{\text{ode}} - \Xi^f \mathbf{w}^f = \Xi^\beta \widehat{\mathbf{w}}^\beta, \tag{11}$$

solving for $\widehat{\mathbf{w}}^\beta$. We then compare the supports of the original and cross-validated boundary weights, $\text{supp}(\widehat{\mathbf{w}}^\beta)$ and $\text{supp}(\mathbf{w}^\beta) := \{i \in \{1, 2, \dots, M^\beta\} : \mathbf{w}_i^\beta \neq 0\}$. Often, the ODE-focused component returns a sparser subset of learned boundary terms, which tend to be more accurate representations of the true dynamics when tested against synthetic data. The terms that are not common between the methods are then removed from the boundary component of the library before refitting. We summarize the method in Algorithm 1 and provide an example of the performance gain from this method in S3 Fig.

3 Results

In the sections to follow, we will test the algorithm on several biologically motivated examples, which in particular focus on two well-studied models: age-structured population models (1) and size-structured population models (2). The uncorrupted measurements of each problem will be constructed using a standard flux-limiter finite volume schemes (details provided in S1 Appendix) [30,31]. We begin with a discussion on the noise assumptions and the performance metrics commonly used to assess these methods.

While standard explorations of the noise sensitivity of similar methods have traditionally involved using additive Gaussian noise, such an approach could result in the artificial measurements of population data being negative, especially for

Algorithm 1 WSINDyStructuredPop.

```
function WSINDyStructuredPop(n_data, N_data, s, t, {φ_k}, {g_j, f_j, β_j}, ...)
    Construct G^g, G^f, Ξ^f, Ξ^β, b^pde, and b^ode from Eqs (5) and (8)
    w^g, w^f, w^β ← MSTLS([G^g, G^f, Ξ^f, Ξ^β], b)
    ŵ^β ← MSTLS(Ξ^β, b^ode − Ξ^f w^f)
    if supp(ŵ^β) ≠ supp(w^β) then
        if supp(ŵ^β) ⋂ supp(w^β) ≠ ∅ then
            idx ← supp(ŵ^β) ⋂ supp(w^β)
        else
            idx ← supp(ŵ^β) ⋃ supp(w^β)
        w^g, w^f, w^β ← MSTLS([G^g, G^f, Ξ^f, Ξ^β(:, idx)], b)
    else
        w^β ← arg min_{v∈{w^β, ŵ^β}} R([w^g; w^f; v])
    return w^g, w^f, w^β
```

large noise levels. Therefore, to address this problem, we opt to use multiplicative log-normal noise to preserve the realistic nonnegative structure of the population data. More precisely, we assume the elements in the dataset $\mathbf{n}$ are of the form $n_j = \varepsilon_j n_j^\star$ such that $\varepsilon_j = \exp(z_j)$, where $\{z_j\}_{j=1}^J$ are i.i.d. Gaussian variables with zero-mean and constant variance σ^2, $z_j \sim \mathcal{N}(0, \sigma^2)$. To measure the effective noise level from this distribution, we define the expected noise-to-signal ratio by

$$\sigma_{NR} := \frac{\mathbb{E}[(n - n^\star)^2]}{||n^\star||^2_{RMS}} = \mathbb{E}[(\varepsilon - 1)^2] = e^{\sigma^2}(e^{\sigma^2} - 1) + (e^{\sigma^2/2} - 1)^2.$$

One benefit of the WSINDy algorithm is that it can not only select proper model ingredients, but also provides an estimate of the linear coefficient present in front of the chosen functions. This is a major benefit when population dynamics are concerned, as the coefficients in front of the total population variable, N, provide insight into critical parameters of the dynamics. For example, in the Verhulst population model, $\dot{N} = rN(1 - N/k)$, k is often interpreted as the carrying capacity of the environment. However, a large bias introduced by the log-normally distributed noise can cause the estimates of N to be skewed and consequently result in major errors in the estimates of these linear parameters, leading to large errors in the dynamics of the total population. This is a critical observation in our setting as, due to the asymmetrical structure of the log-normal distribution, there is a bias encountered when calculating the total population from the noisy number density data. To correct for this, we approximate the variance of the population data using a local polynomial fit [32] and dividing by the estimated expected value of the log-normal noise. It is worth noting that this is unnecessary when the model ingredients do not depend on the total population, as Eq (9) would then only be affected by the dynamics of the total population and not the particular value of N. Precise details of this method are provided in S2 Appendix.

Table 1 summarizes the performance metrics used throughout the manuscript. To measure the accuracy of the recovered coefficients, we make use of two performance metrics commonly used in equation learning methods, denoted by $\mathbf{E}_\infty$ and $\mathbf{E}_2$. Essentially, $\mathbf{E}_\infty$ represents the upper bound of the relative error on the recovered coefficients and $\mathbf{E}_2$ provides information on the magnitude of the error on all coefficients. Additionally, we will make use of the true positivity ratio denoted by **TPR**, where TP represents the number of true positives, FP the number of false positives, and FN the number of false negatives identified by the method.

Finally, to measure the predictive power of the learned model, even when incorrect terms are learned, we define the training time interval by $[0, T_{\text{test}}]$ and whenever $T_{\text{test}} < T$, we measure the prediction accuracy of the model by simulating the evolution of the system using the learned model ingredients to construct a predicted solution $\tilde{n}$. We then define the prediction error $\mathbf{E}_p$ to be the relative L^2 norm of the true solution $n^\star$ and the predicted solution $\tilde{n}$ over the testing interval

Table 1. Performance metrics used throughout the paper.

Label	Formula	Description
$\mathbf{E}_\infty$	$\mathbf{E}_\infty(\mathbf{w}) := \max_{\{j:\mathbf{w}_j^\star \neq 0\}} \frac{\lvert\mathbf{w}_j - \mathbf{w}_j^\star\rvert}{\lvert\mathbf{w}_j^\star\rvert}$	Relative accuracy of true term estimators
$\mathbf{E}_2$	$\mathbf{E}_2(\mathbf{w}) := \frac{\lVert\mathbf{w}-\mathbf{w}^\star\rVert_2}{\lVert\mathbf{w}^\star\rVert_2}$	Magnitude of identified coefficients
TPR	$\mathbf{TPR}(\mathbf{w}) := \frac{TP}{TP+FP+FN}$	True positivity ratio
$\mathbf{E}_p$	$\mathbf{E}_p(\mathbf{w}) := \frac{\lVert\hat{n}-n^\star\rVert_{L2((T_{\text{test}},T)\times\Omega)}}{\lVert n^\star\rVert_{L2((T_{\text{test}},T)\times\Omega)}}$	Prediction error of learned model
R	$\mathbf{R}(\mathbf{w}) := \frac{\lVert\mathbf{b}-G\mathbf{w}\rVert_2}{\lVert\mathbf{b}\rVert_2}$	Residual of the linear system with learned weights

https://doi.org/10.1371/journal.pcbi.1013742.t001

(T_{test}, T). As we will see, when the library consists of similar functions, the learned equation could be analytically incorrect (i.e., **TPR** < 1); however, the learned model still fits the data as the difference between the model ingredients is small when weighted against the data.

3.1 Construction of the library

The construction of the matrix G is the most computationally intensive and arguably the most important part of the algorithm detailed above. The question of which trial functions to include in the library is often best answered by expert insight into the population dynamics. For example, depending on the population in question, age-structured mortality functions have been modeled using exponential and polynomial functions [33]. Therefore, if possible, one should treat the libraries as an array of distinct hypothesis functions the WSINDy algorithm will then select between. It is worth pointing out that while WSINDy can estimate linear parameters, all nonlinear parameters must currently be given in the library. Extending weak-form algorithms such as WSINDy and WENDy to estimate these nonlinear parameters is still an active research area with some sparse recent advancements [16,34], and so we postpone this update to our method for future work.

In this preliminary study, we will construct our libraries as if we are well informed regarding the general shape of the true model ingredients, e.g., we will assume that the fecundity rate is approximately Gaussian. While the usual goal of model discovery methods is to make use of a large library with many choices of functions, it is common for ecologist to have some prior knowledge of the general shape of the heterogeneous components of the model ingredients by studying individuals, e.g., knowing the reproductive age ranges of individuals. Accordingly, we will make use of functions commonly found in the literature of structured populations including polynomial $f_{\text{poly}}(s;p) = s^p$, exponential $f_{\text{e}}(s;k) = e^{ks}$, sigmoidal $f_{\text{sig}}(s;s_c,k) = \frac{1}{1+\exp(-k(s-s_c))}$, and Gaussian $f_{\text{Gauss}}(s;\mu,\sigma) = \exp(-(s-\mu)^2/(2\sigma^2))$. Additionally, we assume the dependence on the total population is multiplicative, i.e. $f(s,N) = f_1(x)f_2(N)$, as is common in the literature [8,35], however, we remark this is not necessary in general.

3.2 Linear example problems

We begin by assessing the method's ability to identify linear population models, that is, where the dynamics are independent of the total population size. We take four representative linear structured population models that differ in their transport, source, and boundary terms. The specific forms of the true functions used in each example are summarized in Table 2. In these examples, the true transport (growth/aging) term $g^\star(s)$, source (death) term $f^\star(s,n)$, and boundary (birth) term $\beta^\star(s)$ are chosen to be simple, but commonly used, functions of the structuring variable s (e.g., age or size). The specific libraries used in each example are listed in Table 3.

3.2.1 The effect of noise. No Noise: In Table 4, we present the coefficient errors and residual from the learned coefficients in the ideal case $\sigma_{NR} = 0$. This provides a baseline test of the algorithm where all of the errors are numerical in nature and not due to added noise.

Table 2. The critical parameters for the linear example problems.

Example	$g^\star(\boldsymbol{s},\boldsymbol{N})$	$\boldsymbol{f}^\star(\boldsymbol{s},\boldsymbol{n},\boldsymbol{N})$	$\beta^\star(\boldsymbol{s},\boldsymbol{N})$	$(\mathbf{0},\boldsymbol{T})\times\Omega$
L.1	1	$-0.3n$	0.4	$(0,10)\times[0,15]$
L.2	1	$-0.1\exp(0.08s)n$	$f_{\text{Gauss}}(s;10,5)$	$(0,10)\times[0,15]$
L.3	$0.2(3–s)$	$-0.3sn$	$f_{\text{sig}}(s;1,2)$	$(0,10)\times[0,3]$
L.4	$s(1–0.25s)$	$-0.7\exp(0.2s)n$	$0.6s$	$(0,10)\times[1,4]$

https://doi.org/10.1371/journal.pcbi.1013742.t002

Table 3. Libraries for the linear example problems in Table 2.

Example	Transport Terms	Source Terms	Boundary Terms
L.1	$\{s^p\}_{p=0}^2$	$\{s^pn\}_{p=0}^3$,	$\{s^p\}_{p=0}^3$,
L.2	$\{s^p\}_{p=0}^2$	$\{\exp(0.08(4k+1)s)n\}_{k=0}^4$	$\{f_{\text{Gauss}}(s;5k,5)\}_{k=\{1,2,3\}}$
L.3	$\{s^p\}_{p=0}^2$	$\{s^pn\}_{p=0}^4$	$\{f_{\text{sig}}(s;k,2)\}_{k=1}^4$
L.4	$\{s^p\}_{p=0}^2$	$\{\exp((0.2+0.5k)s)n\}_{k=0}^3$	$\{s^p\}_{p=1}^4$

https://doi.org/10.1371/journal.pcbi.1013742.t003

Table 4. E_∞, E_2, residual, and prediction error for the test cases and libraries presented above with $T_{\text{test}} = 0.5T$ and $\sigma_{NR} = 0$. All examples for the no noise case have **TPR** = 1.

Label	E_∞	E_2	R	E_p
Ex L.1	9.6e−06	6.1e−06	4.1e−06	5.3e−04
Ex L.2	5.6e−04	3.9e−04	1.1e−05	3.4e−03
Ex L.3	2.9e−04	2.6e−04	5.2e−05	3.9e−03
Ex L.4	8.2e−04	5.2e−04	2.3e−04	5.7e−04

https://doi.org/10.1371/journal.pcbi.1013742.t004

Noise: In Figs 1 and 2, we present the measured performance metrics for the linear models presented in Table 2 using 3000 points in the structural variable and 500 points in time. It is worth noting that as noise is added to the data, the **TPR** of the learned coefficients expectantly drops off. However, the prediction error maintains reasonable levels proportional to the noise level, indicating that the method can consistently learn effective models using the library terms even if they are analytically incorrect. This is demonstrated for example L.2 in Fig 3 where the learned model has a **TPR** value of 0.8, but the model fits the data well and maintains a reasonable prediction level. We provide similar examples of this with the other models in S1 Fig. Additionally, the average run time of each realization and model recovery procedure was less than 10 seconds and can be easily run on any modern laptop.

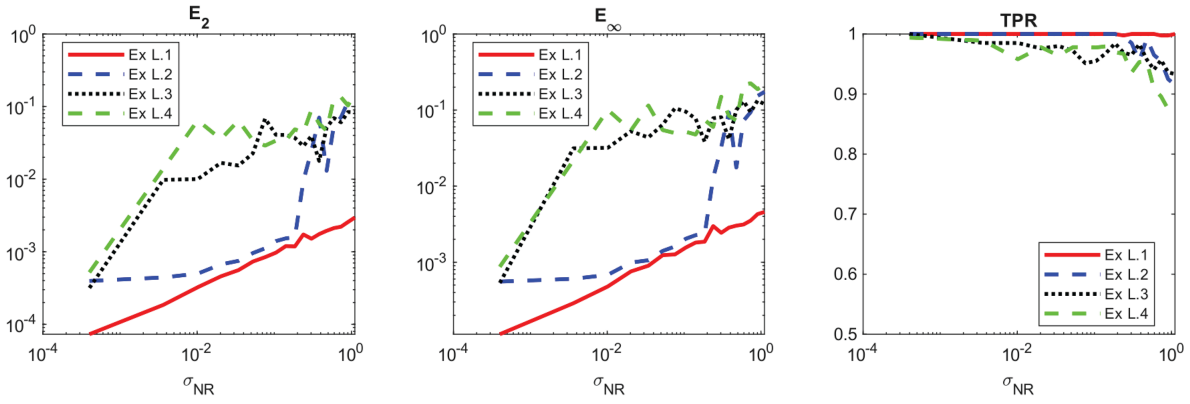

Fig 1. Average performance metrics for the linear models in Table 2 using the libraries in Table 3 plotted against various noise levels using 100 dataset realizations at each noise level.

https://doi.org/10.1371/journal.pcbi.1013742.g001

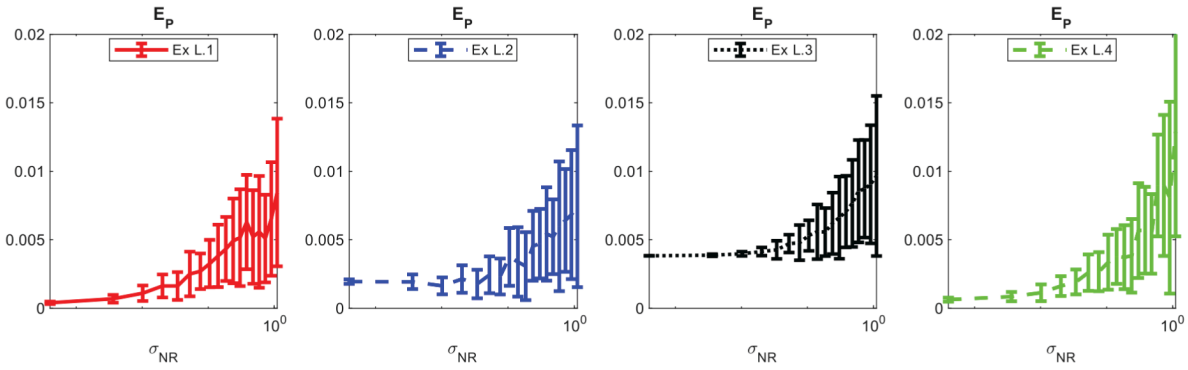

Fig 2. Average and inner quartile range of the prediction error, E_p, for the linear models in Table 2 using the libraries in Table 3 plotted against various noise levels using 100 dataset realizations at each noise level.

https://doi.org/10.1371/journal.pcbi.1013742.g002

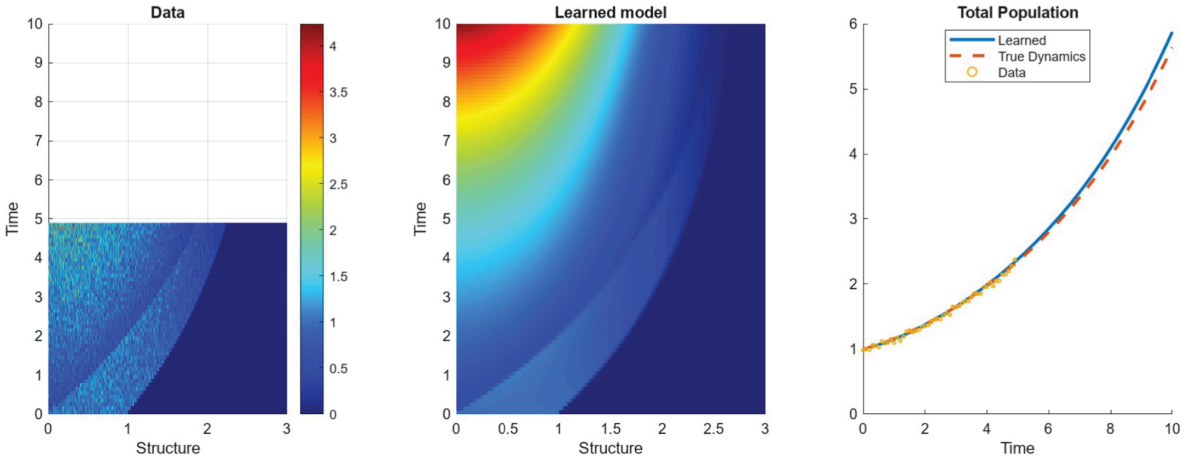

Fig 3. Typical results of using the WSINDy algorithm for example problem L.3 in Table 2 with the library presented in Table 3, $\sigma_{NR} = 0.66$, and 50 points in time. The prediction error and **TPR** of the learned model approximately 0.035 and 0.8, respectively.

https://doi.org/10.1371/journal.pcbi.1013742.g003

3.2.2 Distinguishability of library terms. In this section, we computationally investigate the distinguishability of the candidate functions within the library. How to mathematically characterize and evaluate the distinguishability of a given library for a given data set using sparse regression-based equation learning is currently an open question (see the discussion in Sect 4). Specifically, we examine how effectively the algorithm can distinguish between similar functions of the structural variable during the recovery process. This question is central to the reliability of sparse identification methods such as WSINDy, especially when multiple candidate terms produce similar features with respect to the data.

To explore this, we focus on Example L.2 and, to isolate the effects of term distinguishability, we simplify the transport component by restricting the library to include only the true transport term, $g^\star(s) = 1$. This ensures that the algorithm is not influenced by potential ambiguity in the transport dynamics, allowing us to study the source and boundary components in isolation.

We consider two test cases that assess the distinguishability of the source and boundary terms, respectively:

Case 1 (Source term): We fix the boundary term to include only the true birth rate, $f_{\text{Gauss}}(s; 10, 5)$. For the source term, we construct a series of libraries containing perturbed exponential functions of the form:

$$\{\exp\left((0.08 + \delta i)s\right) n\}_{i=-k}^{k},$$

where δ controls the spacing between candidate terms and k determines the total number of alternatives. This setup allows us to assess how sensitive the algorithm is to small variations in source dynamics and how well it can single out the correct term among closely spaced candidates. Additionally, the structure of the libraries guarantees that the true function $f^{\star}(s, n) = -0.1 \exp(0.08s)n$ is always present for every choice of δ and k.

Case 2 (Boundary term): We now fix the source term to include only the true function, $f^{\star}(s, n) = -0.1 \exp(0.08s)n$, and assess distinguishability in the boundary term. To this end, we consider libraries of Gaussian profiles of the form:

$$\{f_{\text{Gauss}}(s; 10 + \delta i, 5)\}_{i=-k}^{k},$$

where the means of these functions vary over different spacing.

We present the true positive ratio (**TPR**) results for both cases in Figs 4 and 5, assuming no noise ($\sigma_{NR} = 0$). The corresponding condition numbers of the libraries are also shown to indicate how difficult it is to distinguish terms based on their linear dependence. From these results, we observe a general trend: as the spacing δ decreases and candidate terms become increasingly similar, the algorithm's ability to correctly identify the true term deteriorates. In particular, the **TPR** drops sharply once the candidate functions are sufficiently close in structure, indicating that one should be aware and cautious of the similarity between the trial functions.

Such problems have been encountered before in model selection literature and are often addressed using some library regularization method, such as coherence-based pruning methods [36,37]. However, in our preliminary investigation, such methods only provide substantial improvement in the noise-free case and often remove the true terms from the weak-form library when substantial noise is present. Therefore, we opt not to include this regularization in the current method and leave the exploration of such techniques to future work.

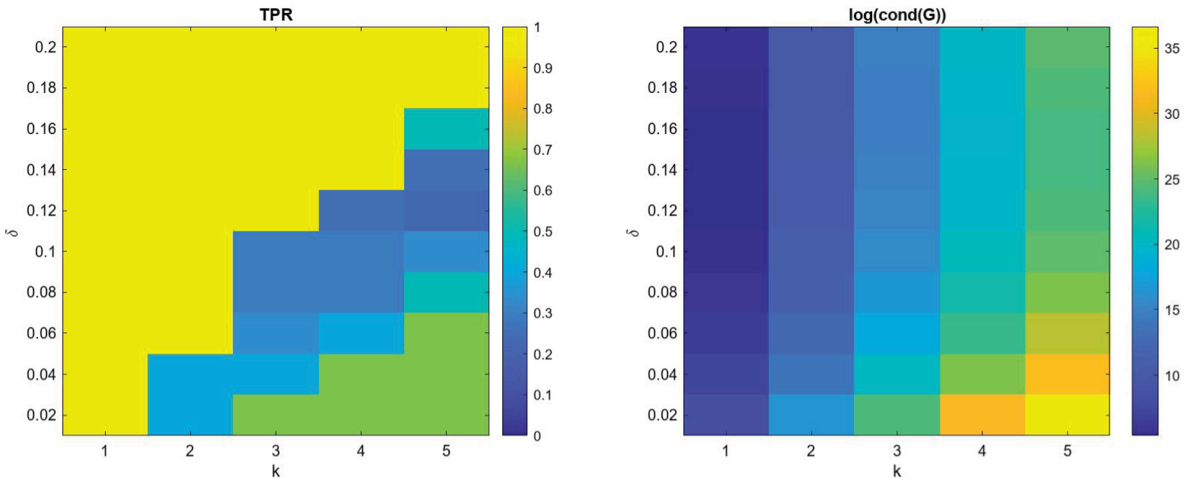

Fig 4. TPR and condition number over different choices of library parameters for case 1.

https://doi.org/10.1371/journal.pcbi.1013742.g004

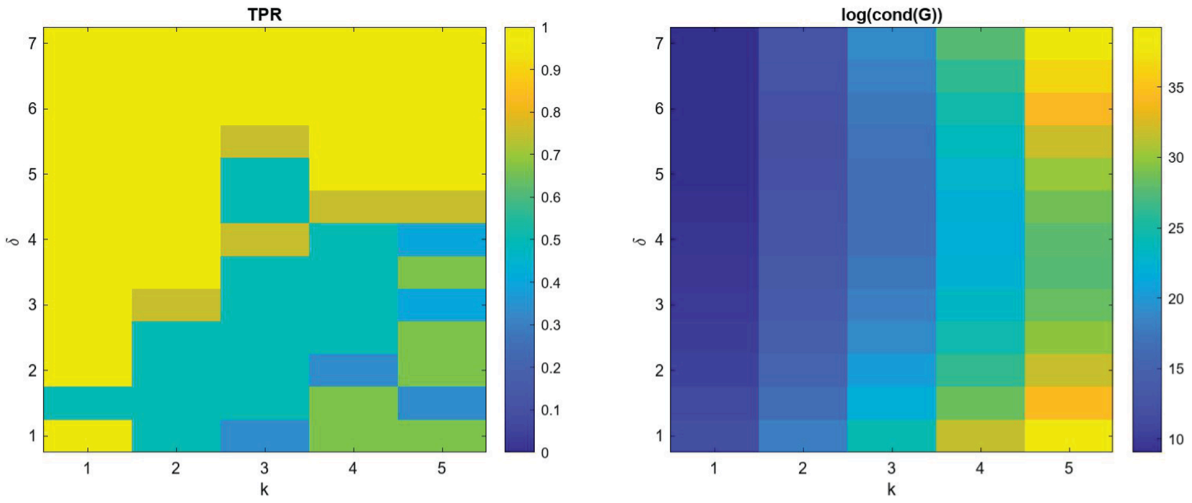

Fig 5. TPR and condition number over different choices of library parameters for case 2.

https://doi.org/10.1371/journal.pcbi.1013742.g005

3.2.3 Resolution of the histogram data. While it is ideal for the method to be provided a data set that is the result of coarse-graining the dynamics of many individuals, this is not always feasible. In many studies of structured populations, especially in studies of large numbers of individuals, it is common to aggregate similarly structured individuals into discrete "classes" or bins. Therefore, it is natural to consider effectiveness of the method when the number of these classes is small. To this end, we display in Fig 6 the performance metrics as a function of the number of structure classes for examples L.2 - L.4 (we opt not to display the results for example *L*.1 as the dynamics in this example are homogeneous and are therefore less affected by the coarsening of the dataset). Even in the absence of noise, we observe a significant drop in the true positive rate (**TPR**) around 20–35 structure classes, suggesting a critical resolution threshold below which WSINDy fails to reliably recover the correct model terms from the given libraries. Furthermore, the coefficient errors

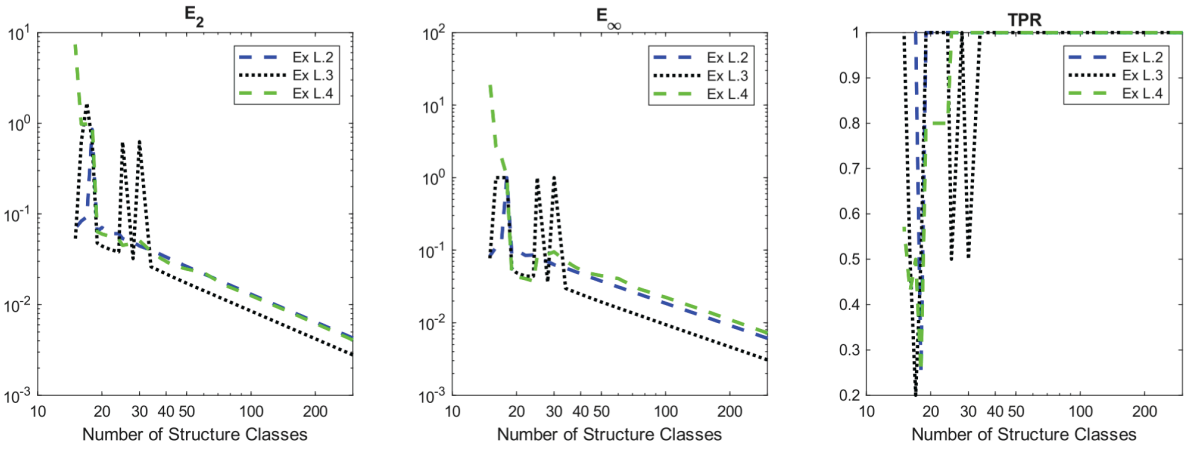

Fig 6. Performance metrics of the method tested against the number of structure classes in the noise-free case.

https://doi.org/10.1371/journal.pcbi.1013742.g006

increase gradually as the number of structure classes decreases, consistent with behavior typically observed in SINDy-type methods when resolution is reduced [29]. We point out that these thresholds are problem specific and the derivation of an analytical threshold for a general class of problems is an open problem.

3.3 Nonlinear example problems

In this section, we consider the more complex case of nonlinear models, where the governing biological processes depend on the total population, N. We will consider the following two examples listed in Table 5. We use libraries similar to those in Table 3 with the addition of the variable N and so we omit the libraries for brevity. We provide an example of using the WSINDy method for recovering Ex NL.2 in Fig 7.

While the method certainly can recover the true dynamics, the introduction of the dependence on the total population comes with several challenges for model discovery using WSINDy which we would like to discuss. First, nonlinear models often converge relatively quickly to an equilibrium, especially when the dynamics occur over long time scales. As a result, the data becomes nearly stationary, making it difficult to identify the underlying dynamical system. Second, convergence to equilibrium leads to many functions of the total population N appearing structurally similar, which complicates their identification within the candidate function library. This indistinguishability is further exacerbated by WSINDy's flexibility in adjusting the linear coefficients of selected terms. In particular, the algorithm may favor effective linear approximations at the cost of omitting the correct nonlinear structure.

To demonstrate these effects, we focus first on example problem NL.1 where the natural approach to recovering the true source term would be to use a library of monomials such as $\{N^k n\}_{k=0}^3$, however, even at very low levels of noise, the monomial library has distinguishability issues. This is demonstrated in Fig 8 where the method learns an effective, but incorrect, mortality term. The algorithm struggles to differentiate between candidate terms because the key differences among them occur near $t = 0$, where the compactly supported test functions are near zero. We point out, however, that at

Table 5. The critical parameters for the nonlinear example problems. The domains for both problems are given by $(0, 20)\times[0, 15]$ and $(0, 20)\times[0, 3]$, respectively.

Example	$g^\star(s, N)$	$f^\star(s, n, N)$	$\beta^\star(s, N)$
NL.1	1	$-0.6Nn$	$1.5\exp(-0.25N)$
NL.2	$0.2(3 - s)\exp(-0.25N)$	$-0.3s(1 + N)n$	$2f_{\text{sig}}(s; 1, 2)(1 - 0.1N)$

https://doi.org/10.1371/journal.pcbi.1013742.t005

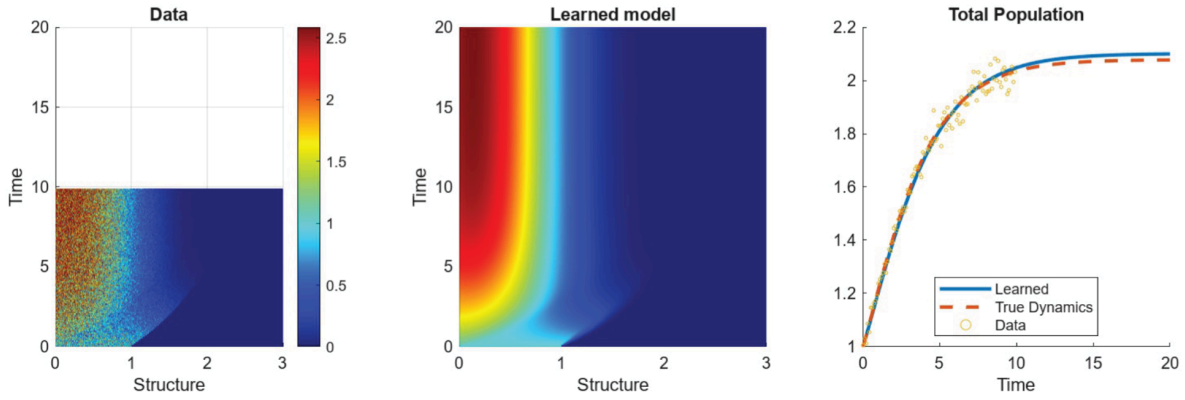

Fig 7. A typical run of Example NL.2 with $\sigma_{NR} = 0.66$ and 50 points in time. The prediction error and **TPR** of this learned model are 0.004 and 1, respectively.

https://doi.org/10.1371/journal.pcbi.1013742.g007

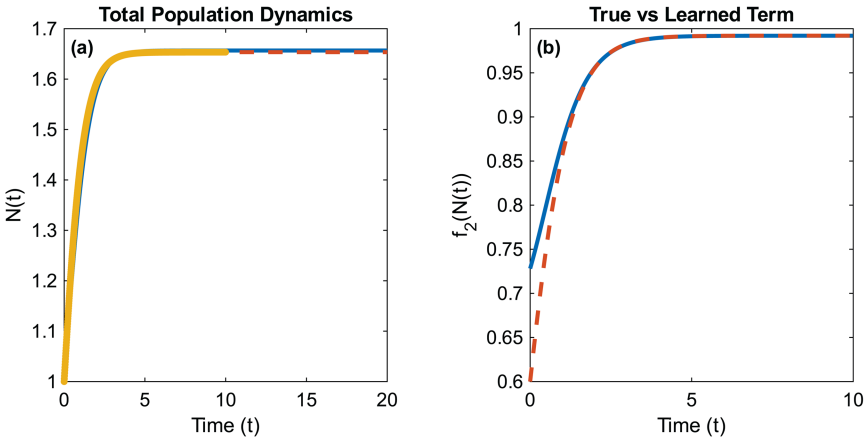

Fig 8. (a) The learned and true dynamics of the total population, *N*, plotted with the training data with noise ratio $\sigma_{NR} \approx 0.02$. (b) The learned nonlinear structure (of the form $a + bN^3$) plotted with the true structure over the training time interval.

https://doi.org/10.1371/journal.pcbi.1013742.g008

low noise levels, this issue can often be mitigated by reducing the radius of the test function support. However, doing so typically degrades performance at moderate to high noise levels.

Next, we consider example NL.2, which incorporates a commonly used logistic-type nonlinearity in the birth rate, $\beta^\star$. When the data is significantly corrupted by noise (e.g., noise level $\sigma_{NR} > 0.20$), the algorithm often replaces the nonlinear structure with an effective linear approximation. To illustrate this behavior, Fig 9 compares the learned and true birth rates, both multiplied by the true population density. In this case, the selected birth rate was given by $\beta(s) = 1.6638 f_{\text{sig}}(s; 1, 2) \approx 2(1 - 0.1\bar{N}) f_{\text{sig}}(s; 1, 2)$ where $\bar{N}$ is the average value of N over the training time. This structure, along with the corresponding plot, demonstrates that although the learned linear birth kernel is analytically incorrect, it is still effective (and sparser) in capturing the birth process from the data set.

While this issue can be resolved with more data, this is not always feasible in practice. So, it is natural to wonder when this linearization is acceptable form the modeling perspective. While the answer to this question will vary depending on the dataset, population, and research questions at hand, generally this substitution is most acceptable when

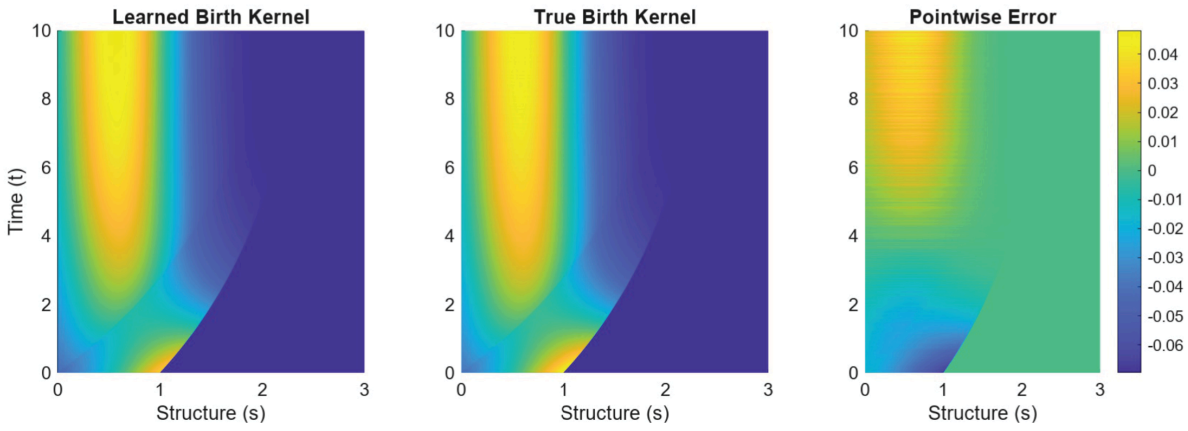

Fig 9. From left to right, we present the learned linear birth kernel $\beta(s)n^\star(t, s)$, the true nonlinear birth kernel $\beta^\star(s, N(t))n^\star(t, s)$, and their pointwise error over the training time.

https://doi.org/10.1371/journal.pcbi.1013742.g009

(i) the estimated effective terms yield predictive accuracy comparable to the true model (if known) within the relevant domain, and
(ii) the inferred dynamics remain qualitatively consistent with known mechanistic constraints (e.g., positivity, boundedness).

Regarding point (i), when the true model is known, one can bound the predictive accuracy of the model using Grönwall-type stability estimates of the model. Perhaps the most natural estimate for these equations is that provided in [38, Theorem 4.6], which states that

$$\|n - n^\star\|_{BL} \leq C_T \max\{\|\beta - \beta^\star\|_\infty, \|g - g^\star\|_\infty, \|f - f^\star\|_\infty\},$$

where $\|\cdot\|_{BL}$ is the dual bounded-Lipschitz norm discussed in many recent works for structured population models [20,21,38,39] (and for general transport equations [40]) and C_T is a constant which depends (exponentially) on the final time T. This can be used to provide a crude estimate of the prediction error when the learned terms are not analytically correct. However, in practice, the true model ingredients (if they exist) tend to be unknown and so this bound is difficult to estimate. Therefore, the development of a formal guiding criteria is left as future work.

3.4 Application to real-world age-structured data

In this section, we demonstrate the utility of WSINDy on a real-world, age-structured population data set that documents the demographic dynamics of a semi-captive population of Asian elephants (*Elephas maximus*) using a data set publicly available via the Dryad digital repository [41] and analyzed from a discrete time modeling perspective in [42]. The dataset records the age and demographic status of individual elephants, enabling us to construct an empirical approximation of the age-structured population density. To achieve this, we aggregate individuals into discrete age classes by binning at a resolution of one year, yielding a structured data array with 81 age classes and 31 temporal snapshots. Although the temporal resolution is relatively coarse for data-driven model discovery, WSINDy is still capable of identifying interpretable and biologically plausible model terms. The resulting equation was of the form

$$\begin{cases} \partial_t n(t,a) + \alpha \partial_a n(t,a) = -c \exp(0.06(a-85)) n(t,a), \\ n(t,0) = \displaystyle\int_0^\infty b f_{\text{Gauss}}(a; 33, 11) n(t,a)\, \mathrm{d}a, \end{cases}$$

with $(\alpha, b, c) \approx (0.9184, 0.3468, 0.0595)$. The library used in this experiment is constructed from a set of Gaussian bases for the birth rates and shifted exponential mortality rates which is common in population ecology [43]. What is meant by *shifted* here is that the exponential functions are of the form $\exp(\gamma(a-85))$ this is necessary as the data set has very few deaths in the population. Therefore, the coefficients in front of a usual exponential term would be extremely small and would then be thresholded out by the algorithm. For exponential functions, this shift acts as a rescaling trick in the sense that $\exp(\gamma(a-85)) = \exp(-\gamma 85)\exp(\gamma a)$ allowing such mortality rates to be recovered in the event of low actualized death. Figs 10 and 11 presents the results of applying WSINDy to the elephant data. Fig 10 compares the reconstructed population dynamics to the binned representation of the data, both at a density level and at the level of the total population. Fig 11 compares the inferred survival and fertility functions to those derived from the matrix population modeling as reported in [42]. These results show that the WSINDy-inferred model ingredients fall within reasonable ranges when compared to traditional modeling techniques.

4 Discussion

The framework of Scientific Machine Learning (SciML) aims to merge methods from scientific computing with those from machine learning to generate accurate and interpretable data-driven models. In this work, we have further developed

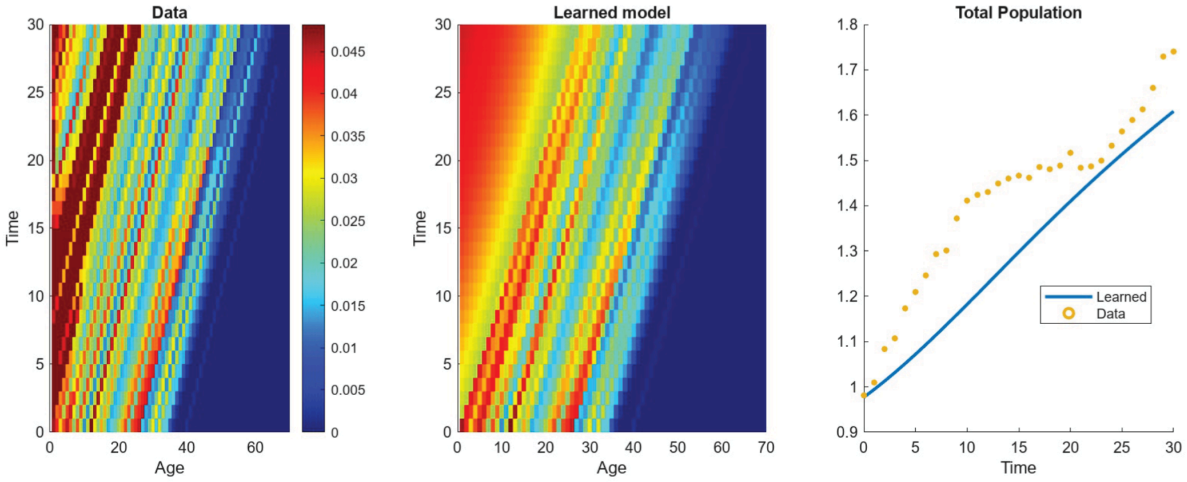

Fig 10. WSINDy-reconstructed age-structured population dynamics of the semi-captive Asian elephant population studied in [42] over a 31-year period.

https://doi.org/10.1371/journal.pcbi.1013742.g010

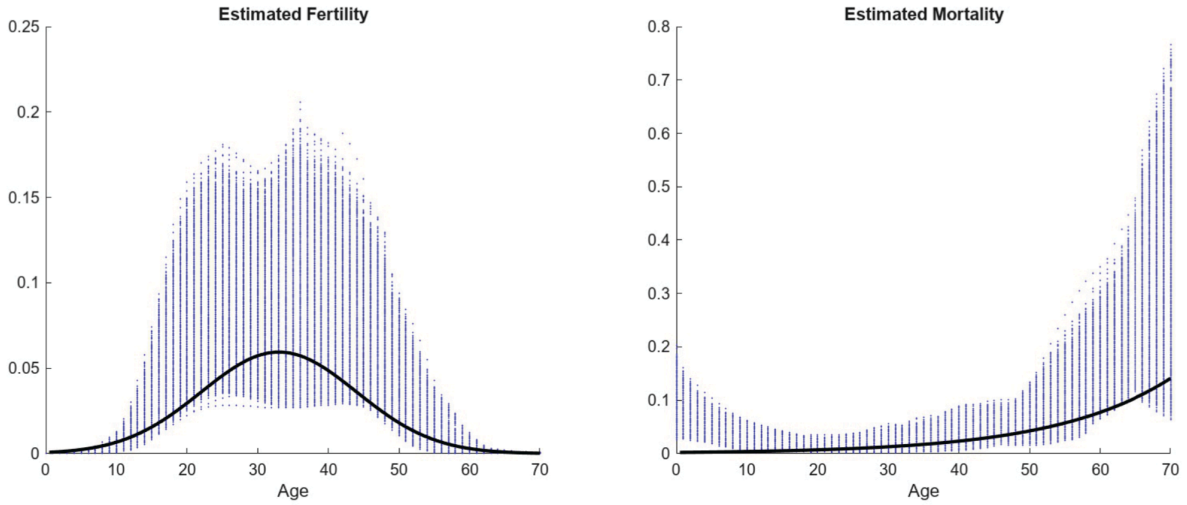

Fig 11. Estimated age-specific fertility and survival functions of the semi-captive Asian elephant population inferred by WSINDy (solid lines), shown alongside those derived from classical matrix projection modeling in [42] (blue dots).

https://doi.org/10.1371/journal.pcbi.1013742.g011

Weak form SciML (WSciML) to learn PDE-based structured population models. To the best of the authors' knowledge, this effort is the first to use SciML methods to learn this class of population models. To this end, we have presented the first iteration of a weak-form equation learning algorithm for structured population equations and have demonstrated the potential of the method on both synthetic and real datasets.

As with any approach, WSciML methods have advantages and disadvantages. In this section, we discuss both the advantages as well as the current limitations. The most notable advantage is the capability of the method to directly learn the governing equation, bypassing the time-consuming iteration between model form creation, numerical approximation, and validation after parameter fitting. It is of course, still necessary to carefully curate a library of mechanistically justified

model features from which to build the model. Given a suitable library, the WSciML process allows researchers to simultaneously test multiple interpretable models for, e.g., structuring relationships, instead of investigating one hypothesis at a time. As a result, WSciML methods can be orders of magnitude faster than the traditional model discovery and parameter inference methods [15].

Another important advantage of WSciML methods is their robustness to noise. As demonstrated in Fig 1 (as well as in other previous publications [13–15,17,44]), the weak form integral transform offers a SciML method which is highly robust to noise. Furthermore, while there is evidence that for certain classes of models, WSciML methods work very well with sparse data [44], structured population models present their own challenges. In particular, this (initial) version of using WSINDy to learn structured populations is moderately sensitive to the number of structure classes in the binned data as shown in Sect 3.2.3. It is also critical that the time series of the structured data include observations that are sufficiently rich in information content so as to allow statistically warranted inclusion or pruning of terms in the library. However, a precise quantification of the needed *richness* is beyond the scope of this work and will be the focus of future efforts. Regardless, with sufficient data, it's clear that WSINDy performs well at selecting effective models that not only fit the data but also possess solid predictive capabilities.

Several studies have also examined how data resolution affects the distinguishability of library terms. A common approach to addressing this issue involves using multiple datasets with varying initial conditions to gain a more comprehensive understanding of the system's dynamics [45,46]. In our setting, data resolution appears to be closely linked to the support of the population density: a larger support over the time series (relative to the size of Ω) tends to yield improved distinguishability across the library terms. A more rigorous study of this effect is left as future work.

As mentioned in Sect 2.2.1, for our structured population models, a naive sparse regression of the combined ODE/PDE model in (9) resulted in correctly learning the PDE at the expense of incorrectly learning the ODE. For this class of problems, it would thus be natural to consider the weak form version of the Ensemble-SINDy (E-SINDy) approach by Fasel et al., [29]. However, we found that E-WSINDy needed a (relatively) larger library to be effective, which resulted in ill-conditioned matrices (for our example problems in this paper). Accordingly, in our example problems, we made use of relatively small libraries, which necessitated a rather restrictive set of hyperparameters in the traditional E-WSINDy method. Because of this, we opted not to make use of the original E-WSINDy method and used our own extension, based on cross-validation (see Sect 2.2.1). In the case of a much larger or redundant library, however, E-WSINDy would remain an efficient method for model discovery.

Although the example models studied here are only in one-dimension, WSINDy has seen much success with PDEs in multiple dimensions [13,27]. Therefore, the generalization of the method to d dimensions is certainly possible. Most structured population models are of relatively low structure dimension (e.g., $d \leq 3$) with few analytical studies considering a large number of structured variables [47]. Since the computations closely follow those of the original WSINDy method (apart from the addition of a one-dimensional boundary process) the overall computational complexity can be made comparable to that reported in [13], namely $\mathcal{O}(N^{d+1}\log(N))$ for N data points in each of the d structure variables. This polynomial-time scaling with respect to the structure dimension is made possible by the use of separable test functions (as described in Sect 2.2) together with fast Fourier transform–based evaluation of the required integrals [13], both of which are essential for efficiently handling higher dimensional systems.

Lastly, we explored potential improvements to the method by focusing on two aspects. First, in this introductory study, we construct the library using a collection of structurally identical trial functions, each parameterized by different (but often similar) values. As demonstrated in Sect 3.2.2, including a large number of such near-redundant functions can significantly increase the condition number of the weak form matrix G. A high condition number amplifies numerical instability and can impair the algorithm's ability to reliably distinguish between candidate terms. This, in turn, may lead to poor term selection, where incorrect terms are favored due to small differences in fit quality being exaggerated by ill-conditioning. A possible avenue to remedy this issue is to iterate the selection step defined in Sect 2.2 with a library optimization step where the parameters in the selected trial functions are modified in such a way as to better fit the data [34]. While the

method presented here requires some minimal prior knowledge of the system (with respect to the choice of trial functions), such an improvement would allow for a more general library where any nonlinear parameters can be fine tuned to the data. Additionally, in the work above, the regression steps do not take into account any natural intuition from the model. Indeed, depending on the application at hand, the method could be tailored to the dynamics by including natural modeling constraints into the regression. Different applications of such models have a variety of natural restrictions on the model ingredients, such as a decaying survival probability in animal populations, symmetric division in cell populations, or mass conservation in structured coagulation-fragmentation equations. Adding these constraints to the regression can significantly improve the interpretability of the models and may result in more accurate identification of the dynamics at higher noise levels or lower data resolutions.

Second, in the case where the structure of the true model ingredients is completely unknown, one may seek to make use of nonparametric representations of the trial functions and aim to "build up" the true ingredients in a similar manner to forward-solver least-squares techniques [48,49]. One can construct such a library using nonparametric functions in the same way as presented in Sect 2.2 and use non-sparse regression techniques to find effective models. This method, however, allows the algorithm much freedom in selecting model ingredients and, as such, can result in highly variable and even unrealistic representations of the model ingredients and would possibly be better handled by parameter estimation methods such as WENDy. This is an active frontier of research.

Acknowledgments

This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder. The authors would like to thank D. Messenger (Los Alamos National Labs) and N. Heitzman-Breen (CU Boulder) for discussion regarding the implementation of the method.

Supporting information

S1 Appendix. Simulation details.
(PDF)

S2 Appendix Approximation of the total population.
(PDF)

S1 Fig. WSINDy results. Typical results of using the WSINDy algorithm for the linear models in Table 2 with libraries presented in Table 3, $\sigma_{NR} = 0.66$, and 50 points in time.
(EPS)

S2 Fig. Test Function Heatmap. An example heatmap of average TPR values over different values of r_t and r_x. This is for example L.3 and the results are averaged over 10 realizations of a noise level $\sigma_{NR} = 0.25$.
(EPS)

S3 Fig. Cross Validation. TPR of Ex L.2 over different noise levels with cross validation (blue) and no cross validation (red). The curves represent the average of 100 realizations of each noise level.
(EPS)

S4 Fig MSTLS Comparison. Comparison of the MSTLS algorithm and OLS for example L.3 for typical runs at noise levels $\sigma_{NR} = 0.0$, $\sigma_{NR} = 0.1$, and $\sigma_{NR} = 0.5$.
(EPS)

Author contributions

Conceptualization: Rainey Lyons, David M. Bortz.

Funding acquisition: David M. Bortz.

Methodology: Rainey Lyons, Vanja Dukic, David M. Bortz.

Software: Rainey Lyons.

Supervision: David M. Bortz.

Validation: Rainey Lyons.

Visualization: Rainey Lyons.

Writing – original draft: Rainey Lyons, Vanja Dukic, David M. Bortz.

Writing – review & editing: Rainey Lyons, Vanja Dukic, David M. Bortz.

References

1. Metz JAJ, Diekmann O, Levin S. The dynamics of physiologically structured populations. Berlin, Heidelberg: Springer; 1986.
2. Von Foerster H. Some remarks on changing populations. In: Stohlman F, editor. The kinetics of cellular proliferation. Grune & Stratton; 1959. p. 382–407.
3. Keyfitz BL, Keyfitz N. The McKendrick partial differential equation and its uses in epidemiology and population study. Mathematical and Computer Modelling. 1997;26(6):1–9. https://doi.org/10.1016/s0895-7177(97)00165-9
4. M'Kendrick AG. Applications of mathematics to medical problems. Proceedings of the Edinburgh Mathematical Society. 1925;44:98–130. https://doi.org/10.1017/s0013091500034428
5. Sinko JW, Streifer W. A new model for age□size structure of a population. Ecology. 1967;48(6):910–8. https://doi.org/10.2307/1934533
6. Milner FA, Rabbiolo G. Rapidly converging numerical algorithms for models of population dynamics. J Math Biol. 1992;30(7):733–53. https://doi.org/10.1007/BF00173266 PMID: 1522394
7. Pilant M, Rundell W. Determining a coefficient in a first-order hyperbolic equation. SIAM J Appl Math. 1991;51(2):494–506. https://doi.org/10.1137/0151025
8. Gurtin ME, Maccamy RC. Non-linear age-dependent population dynamics. Arch Rational Mech Anal. 1974;54(3):281–300. https://doi.org/10.1007/bf00250793
9. Cushing JM. Some competition models for size-structured populations. Rocky Mountain J Math. 1990;20(4). https://doi.org/10.1216/rmjm/1181073049
10. Wood SN. Inverse problems and structured-population dynamics. In: Tuljapurkar S, Caswell H, editors. Structured-population models in marine, terrestrial, and freshwater systems. Boston, MA: Springer US; 1997. p. 555–86.
11. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci U S A. 2016;113(15):3932–7. https://doi.org/10.1073/pnas.1517384113 PMID: 27035946
12. Rudy SH, Brunton SL, Proctor JL, Kutz JN. Data-driven discovery of partial differential equations. Sci Adv. 2017;3(4):e1602614. https://doi.org/10.1126/sciadv.1602614 PMID: 28508044
13. Messenger DA, Bortz DM. Weak sindy for partial differential equations. J Comput Phys. 2021;443:110525. https://doi.org/10.1016/j.jcp.2021.110525 PMID: 34744183
14. Messenger DA, Bortz DM. Weak SINDy: Galerkin-based data-driven model selection. Multiscale Model Simul. 2021;19(3):1474–97. https://doi.org/10.1137/20m1343166 PMID: 38239761
15. Bortz DM, Messenger DA, Dukic V. Direct estimation of parameters in ODE models Using WENDy: weak-form estimation of nonlinear dynamics. Bull Math Biol. 2023;85(11):110. https://doi.org/10.1007/s11538-023-01208-6 PMID: 37796411
16. Rummel N, Messenger DA, Becker S, Dukic V, Bortz DM. WENDy for nonlinear-in-parameter ODEs. arXiv preprint 2025. https://doi.org/arXiv:250208881
17. Bortz DM, Messenger DA, Tran A. Weak form-based data-driven modeling: computationally efficient and noise robust equation learning and parameter inference. In: Mishra S, Townsend A, editors. Numerical Analysis Meets Machine Learning. Elsevier; 2024. p. 54–82.
18. Messenger DA, Tran A, Dukic V, Bortz DM. The weak form is stronger than you think. SIAM News. 2024;57(8).
19. Ackleh AS, Ito K. Measure-valued solutions for a hierarchically size-structured population. Journal of Differential Equations. 2005;217(2):431–55. https://doi.org/10.1016/j.jde.2004.12.013

20. Ackleh AS, Lyons R, Saintier N. A structured coagulation-fragmentation equation in the space of radon measures: unifying discrete and continuous models. ESAIM: M2AN. 2021;55(5):2473–501. https://doi.org/10.1051/m2an/2021061
21. Düll C, Gwiazda P, Marciniak-Czochra A, Skrzeczkowski J. Spaces of measures and their applications to structured population models. Cambridge University Press; 2021.
22. Getz WM. The ultimate-sustainable-yield problem in nonlinear age-structured populations. Mathematical Biosciences. 1980;48(3–4):279–92. https://doi.org/10.1016/0025-5564(80)90062-0
23. Perthame B. Transport equations in biology. Basel: Birkhäuser Basel;2007.
24. Falster DS, Brännström Å, Dieckmann U, Westoby M. Influence of four major plant traits on average height, leaf□area cover, net primary productivity, and biomass density in single-species forests: a theoretical investigation. Journal of Ecology. 2010;99(1):148–64. https://doi.org/10.1111/j.1365-2745.2010.01735.x
25. Kooijman SA, Metz JA. On the dynamics of chemically stressed populations: the deduction of population consequences from effects on individuals. Ecotoxicol Environ Saf. 1984;8(3):254–74. https://doi.org/10.1016/0147-6513(84)90029-0 PMID: 6734503
26. Tran A, Bortz D. Weak form scientific machine learning: test function construction for system identification. arXiv preprint 2025. https://doi.org/arXiv:250703206
27. Minor S, Messenger DA, Dukic V, Bortz DM. Learning physically interpretable atmospheric models from data with WSINDy. Journal of Geophysical Research: Machine Learning and Computation. 2025;2(3):e2025jh000602. https://doi.org/10.1029/2025jh000602
28. Messenger D, Dwyer G, Dukic V. Weak-form inference for hybrid dynamical systems in ecology. J R Soc Interface. 2024;21(221):20240376. https://doi.org/10.1098/rsif.2024.0376 PMID: 39689846
29. Fasel U, Kutz JN, Brunton BW, Brunton SL. Ensemble-SINDy: robust sparse model discovery in the low-data, high-noise limit, with active learning and control. Proc Math Phys Eng Sci. 2022;478(2260):20210904. https://doi.org/10.1098/rspa.2021.0904 PMID: 35450025
30. Ackleh AS, Chellamuthu VK, Ito K. Finite difference approximations for measure-valued solutions of a hierarchicallysize-structured population model. Math Biosci Eng. 2015;12(2):233–58. https://doi.org/10.3934/mbe.2015.12.233
31. Ackleh AS, Ma B. A second-order high-resolution scheme for a juvenile-adult model of amphibians. Numer Funct Anal Optim. 2013;34(4):365–403. https://doi.org/10.1080/01630563.2012.730595
32. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. In: Springer Series in Statistics. Cham: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-19425-7
33. Avraam D, Arnold S, Vasieva O, Vasiev B. On the heterogeneity of human populations as reflected by mortality dynamics. Aging (Albany NY). 2016;8(11):3045–64. https://doi.org/10.18632/aging.101112 PMID: 27875807
34. Ducci G, Kouyate M, Reuter K, Scheurer C. Pareto-based optimization of sparse dynamical systems. J Chem Phys. 2025;162(11):114118. https://doi.org/10.1063/5.0249780 PMID: 40105141
35. Ackleh AS, Lyons R, Saintier N. Finite difference schemes for a structured population model in the space of measures. Math Biosci Eng. 2019;17(1):747–75. https://doi.org/10.3934/mbe.2020039 PMID: 31731375
36. Bajwa WU, Calderbank R, Jafarpour S. Model selection: two fundamental measures of coherence and their algorithmic significance. In: 2010 IEEE International Symposium on Information Theory. 2010. p. 1568–72. https://doi.org/10.1109/isit.2010.5513474
37. Zörlein H, Akram F, Bossert M. Dictionary adaptation in sparse recovery based on different types of coherence. 2013.
38. Gwiazda P, Lorenz T, Marciniak-Czochra A. A nonlinear structured population model: lipschitz continuity of measure-valued solutions with respect to model ingredients. Journal of Differential Equations. 2010;248(11):2703–35. https://doi.org/10.1016/j.jde.2010.02.010
39. Ackleh AS, Lyons R, Saintier N. High resolution finite difference schemes for a size structured coagulation-fragmentation model in the space of radon measures. Math Biosci Eng. 2023;20(7):11805–20. https://doi.org/10.3934/mbe.2023525 PMID: 37501421
40. Hille SC, Lyons R, Muntean A. Invariance properties of the solution operator for measure-valued semilinear transport equations. AnalMathPhys. 2025;15(4). https://doi.org/10.1007/s13324-025-01093-3
41. Jackson J, Mar K, Htut W, Childs D, Lummaa V. Data from: changes in age-structure over four decades were a key determinant of population growth rate in a long-lived mammal. 2020.
42. Jackson J, Mar KU, Htut W, Childs DZ, Lummaa V. Changes in age-structure over four decades were a key determinant of population growth rate in a long-lived mammal. J Anim Ecol. 2020;89(10):2268–78. https://doi.org/10.1111/1365-2656.13290 PMID: 32592591
43. Rockwood LL. Introduction to population ecology. Malden, MA: Blackwell Publication; 2006.
44. Messenger DA, Bortz DM. Asymptotic consistency of the WSINDy algorithm in the limit of continuum data. IMA Journal of Numerical Analysis. 2024. https://doi.org/10.1093/imanum/drae086
45. Vasey G, Messenger D, Bortz D, Christlieb A, O'Shea B. Influence of initial conditions on data-driven model identification and information entropy for ideal mhd problems. Journal of Computational Physics. 2025;524:113719. https://doi.org/10.1016/j.jcp.2025.113719
46. Lyu W, Galvanin F. DoE-integrated sparse identification of nonlinear dynamics for automated model generation and parameter estimation in kinetic studies. In: Computer Aided Chemical Engineering. vol. 53. Elsevier; 2024. p. 169–74.
47. Tucker SL, Zimmerman SO. A nonlinear model of population dynamics containing an arbitrary number of continuous structure variables. SIAM J Appl Math. 1988;48(3):549–91. https://doi.org/10.1137/0148032

48. Banks HT, Botsford LW, Kappel F, Wane C. Estimation of growth and survival in size-structured cohort data: an application to larval striped bass (Morone saxatilis). J Math Biol. 1991;30(2):125–50. https://doi.org/10.1007/bf00160331

49. Banks HT, Sutton KL, Thompson WC, Bocharov G, Roose D, Schenkel T, et al. Estimation of cell proliferation dynamics using CFSE data. Bull Math Biol. 2011;73(1):116–50. https://doi.org/10.1007/s11538-010-9524-5 PMID: 20195910