Scope and Vision Paper

Learning the Governing Equations of Ecological Dynamics

By
Markus Bauer, Samson Mont, Nour Rahal-Arabi,
Madelyn Smith, Jinshui Wang

Section 1: Background and Problem Statement

1.1. Background

Creating models to understand species abundance is a significant area of research in the field of Ecology. researchers must gain a deep and comprehensive understanding of population dynamics to explain interactions between predators and prey as well as predict population abundance over space and time.

Historically, ecological researchers have used statistical methods which rely on fitting a "correct" model to data by estimating parameters that are assumed to be significant. However, symbolic regression is a promising new method for producing descriptive models because it has the potential to derive the multiple significant factors and predator-prey interactions directly from species abundance data. In addition, symbolic regression does not assume a prior model which suggests that the one it extracts is the data's "true" model.

The models in this project will be derived from time-series data on a three-year field experiment that started in 2013. In this experiment, researchers collected data about population abundance for a variety of species that were submerged under water during high tide and out in open air during low tide in the Oregon marine intertidal. The primary predator species of interest is the whelk Nucella ostrina, with focus also on Nucella canaliculata. The primary prey species of interest is the acorn barnacle Balanus glandula and the gooseneck barnacle Pollicipes polymerus.

1.2. Problem Statement

This project aims to create an accurate and comprehensible model to describe and predict population dynamics for species in the Oregon marine intertidal using a novel machine learning technique known as symbolic regression. Ideally, this model can be applied to describe other ecosystems as an alternative to traditional methods of characterizing species interactions.

Section 2: Vision

The development team aims to produce a mathematical model for describing and predicting the population fluctuations in an Oregon coastal ecosystem resulting from predator-prey interactions. Our strategy utilizes symbolic regression, a machine learning algorithm recently introduced to the field of Ecology, to represent these relationships. The system will generate an equation or system of equations that will consider multiple variables and provide an accurate species abundance model for the Oregon marine intertidal region which was surveyed. The success of the project will provide an alternative to traditional modeling, which relies on statistical analysis, predetermined assumptions, and educated guesses about significant parameters.

2.1. Hypothesis

Growth Hypothesis:

If the resulting model from this research successfully describes how populations are influenced by intrinsic and extrinsic factors in the Oregon marine intertidal, then other researchers can apply this technique to model population dynamics in other ecosystems. For these users to discover the project, the stakeholder and team must write about their work to publish in ecological research journals.

Value Hypothesis:

The traditional approach to predict and describe population fluctuations in an ecological system employs statistical methods that assume a prior and fit the model to the data. The goal is to utilize a machine learning method that learns the true model from the data itself to better evaluate this relationship in a more time-effective manner while maximizing accuracy. If the model reaches an accuracy in the testing phase of at least 80% or outperforms traditional approaches, it will be considered a viable alternative for modeling population dynamics.

2.2.2 Requirements

Functional Requirements:

The functional requirements are as follows:

1. The system should import and read data from the four datasets gathered by the Project Partner. This data will train and test the symbolic regression algorithm.
2. The developed program will produce one model that describes population growth rates of the nucella ostrina over time, and one model that describes the feeding rates of nucella ostrina.

3. Each model produced by the system should describe and predict populations of interest to an accuracy of above 80%.

Non-Functional Requirements:

The non-functional requirements are as follows:

1. The system should use a symbolic regression approach to generate each model.
2. Each model produced by the system should be less complex than the previous models.
3. Each model describing population growth rates should be a function of their own abundance, other species abundance, temperature, and time with respect to significant variables.
4. Each model describing feeding rates of Nucella Ostrina should be a function of prey species abundance, their own abundance, temperature, and time.
5. The system should be created using regular communication between the project partner, Mark Novak, and the development team.
6. The model and any code developed to generate the model should be functional, well-documented, and free of software bugs.
7. The equation(s) derived from the system must be easily understandable to the audience (ecologists).

Section 3: Success Measure and Stakeholders

3.1. Success Measures

According to the project partner, the measure of success for this project is to make notable progress toward deriving equations from machine learning algorithms which accurately describe a population and are easy to understand. From this defined goal, the system must meet the following requirements.

1. The system must use Symbolic Regression to obtain 80% accuracy on the provided datasets by the end of the project.
2. The derived equation should be understandable to the general population, which will be evaluated using a comprehensibility metric which will be determined by user research and a literature review.
3. If any of the success measures listed above are not satisfied, the system is considered successful if ample documentation and software development are produced for future teams to expand upon.
4. If the team delivers the primary project goals before the conclusion of the project, an additional measure of success is to write and publish a scientific report on the system in an ecological research journal.

During the development process for this system, these measures of success are subject to change as more information is researched about existing systems and the provided dataset. For example, the team may opt to experiment with a machine learning technique which differs from symbolic regression if that strategy fails to provide accuracy or comprehensibility.

3.1.2 Success Measures Changes - 3/14/22

Instead of having the equation be understandable to the general population (requirement 2), the equation must be understandable to a technical audience that is concerned with our research. After generating some equations and discussing them with the stakeholder, it seems that a general audience would not understand what the equations reflected unless they were familiar with the area of research and the species involved. Therefore, a more technical audience makes more sense. As a result of this, user research to determine comprehensibility of the equations will not be necessary.

3.2. Stakeholders:

There are three stakeholders identified for this project. The first is the project partner, Mark Novak, a quantitative ecologist who seeks to use both empirical and theoretic methods to better understand the dynamic of populations and multispecies communities. As the main petitioner for the concept, he wishes

to utilize this new technology to provide better and more understandable models for the scientific community. The second stakeholder is identified as the ecologists and other scientists that are likely to utilize or read about the machine learning model through a scientific publication. Finally, the development team has a stake in the project because their capstone course grade and undergraduate graduation depend on the success of their research and development for the system.

Section 4: Project Constraints and Risks

4.1. Constraints:

Time

This research and development must be completed by the end of the 2021-2022 academic year. Due to the pioneering nature of the methodology (symbolic regression for modeling) in this field, general progress toward an understanding of the ecological species in the experiment and the tradeoffs of using machine learning to model abundance data is an indication of success. To specify time constraints for deliverables, Mark and the development team must learn more about the development process for machine-learning symbolic regression models. Thus, primary time constraints are the deadlines for documentation assignments for the Capstone class. The R&D team expects to spend a collective 10-15 hours each week researching, developing, and evaluating the models. This estimate also includes weekly meetings with the assigned TA and primary stakeholder and team Scrum meetings to discuss individual progress. The team will formalize the timeline for deliverables as familiarity with the dataset and symbolic regression increases.

Resources

The only resources available are the resources that each member possesses and free services that Oregon State University provides to students. Mark Novak will be consulted for clarification on project requirements. The team will also refer to free online forums and professors who can provide support for model development. The National Science Foundation is no longer funding this project, so the team has no budget. Given the scope of this project, the lack of funding is not expected to be a problem.

Scope

The scope for this project is flexible. Mark defined his measure of success to be any progress made toward an understandable equation to define population dynamics as observed by a three-year field experiment. A few stretch goals were mentioned in the case that a functional solution is successfully

created using symbolic regression. The first is to test the produced model against existing solutions and similar datasets to see how the equation and ML algorithm compare. The second is to collaboratively write a paper to be published in a scientific journal detailing the work from the course of the project.

4.2. Risk Management

Risk	Likeli-hood	Impact	Mitigation Strategy	Early Detection	Consequence
The deadline may not be possible with the size of the minimal required scope of the project.	Unlikely	High	To mitigate this, a review of current progress and an update to the plan will be done on a weekly basis so more resources can be brought in if required to accomplish the minimal required scope on time. Each update will provide an opportunity to assess current progress and decide if the project should move forward or if it should be scrapped or postponed	Weekly plan updates result in an iteration plan that goes beyond the deadline.	Should the mitigation strategy fail to prevent/avoid the risk, the project may be scrapped without any solution being deployed.
Team member gets covid	Unlikely	Low	Discord will serve as a hub for remote communication, so it should be easy to stay in touch and ensure deliverables are met	Frequent communication is expected from each member. Members will notify each other of any illness.	Productivity may be slightly lower during this time
The model is less than 80% accuracy	Unlikely	Medium	In the early design model, the method	For the establishment of	Results predicted by the model may

			of improving model accuracy will be considered.	the model, output (print) the corresponding precision results	be largely inconsistent with the actual situation, so the project could not be used for experimental analysis

Section 5: Iteration Plan and Estimate

Our individual iteration plans can be found in our previously submitted WIC's. Submitted with this file is a copy of our spreadsheet containing a week by week task list in the form of a Gantt Chart. We have structured this model off of our current understanding of the project, which is likely to change over the course of the next term. We have this time allocated in two-week long sprints instead of the assumed three week style; we chose this to better align with a 10 week long term and allow for more adaptability in case we need to make changes over the course of the project. We aim to be testing models during much of the Winter term, nearing completion during weeks 9 and 10. We have included both a link to the spreadsheet in this document as well as an image of the Gantt Chart.

Gantt Chart Link: Gantt Chart Section 5

Learning the Governing Equations of Ecological Dynamics		Sprint 4 *Research*		Sprint 5 *Research*		Sprint 6 *Initial Prototyping*		Sprint 7 *Refining Models*		Sprint 8 *Refining Models*		Sprint 9 *Refining Models*		Sprint 10 *Model Selection*	
Task #	Task Name	*Week 7*	*Week 8*	*Week 9*	*Week 10*	*Week 11*	*Week 12*	*Week 13*	*Week 14*	*Week 15*	*Week 16*	*Week 17*	*Week 18*	*Week 19*	*Week 20*
	Understand Data														
1	Research Datasets														
1.1	Research factors that impact Nucella ostrina														
1.2	Conduct user research														
1.3	Conduct literature analysis														
1.4	Identify factors of comprehensibility														
1.5	Create comprehensibility metric														
1.6	Share findings with developers														
1.7	Evaluate each model														
1.8	Compare all models														
1.9	Select best model														
2	Existing Solutions														
2.1	Collecting information for model														
2.2	Using model from research results														
2.3	Code implementation														
2.4	Training and testing model														
2.5	Optimizing model hyperparameter														
2.6	Integrate code and write a report														
	Documentation and Communication														
1	Documentation														
1.1	Weekly Group Assignments														
1.2	Individual Class Assignments														
1.3	Document Results														
2	Communication														
2.1	Team/TA Meetings														
2.2	Project Partner Meetings														
2.3	Communicate Findings														
	Select Final ML Model														
1	Model 1														
1.1	Planning														
1.2	Data Manipulation														
1.3	Code Implementation														
1.4	Testing/Cross Validation														
1.5	Visualizations?														
1.6	Analysis														
2	Model 2														
2.1	Conduct Preliminary Research														
2.2	Data Analysis														
2.3	Design and Implement Model														
2.4	Train/Test Symbolic Regression														
2.5	Validate/Evaluate Accuracy														
2.6	Refinement Iterations														
3	Model 3														
3.1	Learning Symbollic Regression														
3.2	Planning first model														
3.3	Code Model 1														
3.4	Train/Test Model														
3.5	Visualize/Evaluate Success														
3.6	Refine Model														

5.2 Iteration Plan and Estimate Changes - 3/14/22

One major change we will be making to this section is that model development will continue for the first four weeks of spring term rather than completing at the end of winter like we originally planned for. This derivation from the initial draft should not harm our ability to complete deliverables by the required deadlines as we designed our project timeline with flexibility at the forefront.

A small change that was made is that the project partner updated new temperature data with daily average and we only needed to update the new data set.