

Beyond Mechanistic Models: The case for machine learning in ecological management

Carl Boettiger^{a,*}, Marc Mangel^a, Stephan B. Munch^b

^a*Center for Stock Assessment Research, Department of Applied Math and Statistics, University of California, Mail Stop SOE-2, Santa Cruz, CA 95064, USA*

^b*Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, 110 Shaffer Road, Santa Cruz, CA 95060, USA*

Abstract

Introduction

Process-based models built on modern machine learning approaches offer a powerful and underappreciated approach for ecological forecasting and management. We present a novel example of how one such example - non-parametric Gaussian processes, can be adapted to the context of management and decision-making.

While effective machine learning algorithms have speckled the literature, most theoretical ecologists in particular have been wary of these ‘black-box’ approaches in favor of simple, mechanistic models. While a blind application of machine learning approaches would indeed be dangerous grounds to stake either ecological understanding or management practice, such concerns are not grounds to reject such methods out of hand. Rather, further attention is warranted to identify the circumstances in which machine-learning approaches will most likely outperform existing more mechanistic approaches, or help crack difficult management questions where ecological or dynamical complexity has caused traditional approaches to founder.

Here we illustrate how a machine learning model can be implemented in place of simple mechanistic model in a decision theory (optimal control) problem, and highlight the features of this system that make the nonparametric approach particularly effective.

Mechanistic models have long been the gold standard of theoretical modeling in ecology (e.g. see @Geritz2012 or @Cuddington2013). Only by understanding the processes involved can we make reliable long term predictions and build an knowledge of cause and effect that guides the hypotheses we make, the data we collect, and the management decisions we make. Process-based models, whether expressed in the language of mathematics or English, identify the connection between mosquitoes and the spread of malaria, or greenhouses gases and climate change, guiding our approach to understand and manage these threats. Despite this central importance, we

*Corresponding author, cboettig@ucsc.edu
Email address: cboettig@ucsc.edu (Stephan B. Munch)

argue here that ecologists would do well to give greater attention to the role non-mechanistic models can play in ecological management and decision making. The value of these approaches is greatest in a context where decisions are made over short time horizons, and updated as new data becomes available.

The alternative to process-based models: pattern-based modeling

Before we proceed further it would be useful to define our terms and lay some greater context for the discussion. Throughout, we will distinguish “mechanistic” or “process-based” models from models that are merely “correlative” or “pattern-based”. Historically pattern-based modeling meant regression (usually linear regression) and resided in methodology rooted in statistics departments and was the primary focus of empirical ecologists, while mechanistic modeling meant dynamical systems (ODEs, PDEs, SDEs and their discrete kin), residing in methodology from the mathematics department and was the primary focus of theoretical ecologists.

We believe these historical divides continue to color much of the literature today, with theorists more skeptical than empiricists of statistical methodology and vice versa. Meanwhile, the ground underneath has shifted. Originally pattern-matching approaches could be critiqued on the grounds of their simplicity (not everything is linear) while mechanistic approaches could be critiqued on their tenuous connections to data – model parameter values such as death rates, birth rates, etc. would be estimated in advance and then stuck into the model, rather than estimated in the context of the model itself. Today both of these critiques are outdated. Computational power and hierarchical statistical methods (particularly approximate likelihood or simulation techniques such as particle filters and approximate Bayesian computing) have brought the inference richer dynamical systems into their fold, while pattern-based modeling has spawned an entirely new approach under the banner of machine learning that has divided the statistics community (See @Bremman2001, “The Two Cultures”). Machine learning models can represent almost arbitrarily complex patterns and incorporate learning and decision making strategies that have earned them the name ‘artificial intelligence.’ Yet at the same time they can be far less transparent their linear regression predecessors, making both astounding successes and startling biases and failures like their biological namesake. More ecologists, particularly theorists and modelers, would do well to learn both how to take advantage of these strengths and recognize their weaknesses.

Mechanistic modeling emphasizes the importance of capturing the correct gross properties of a system over tracking minute fluctuations. For instance, in selecting and parameterizing model of a population of conservation concern, we may be most interested in getting the long-term behavior correct – such as identifying if the dynamics support persistence of the population – rather than worrying how well they reflect the year-to-year fluctuations. We would certainly have good reason to prefer such a model over alternatives which are irreconcilable to the most basic biological processes, such as unbounded growth, or growth curves that do not pass through the origin in a closed system. So it may come as a surprise to realize that such obviously wrong models can perform as well or even better than reasonable mechanistic models in guiding ecological management and decision making.

A simple example comparing the performance of the two modeling approaches in a quantitative decision problem will help to introduce these issues.

Models & Methods

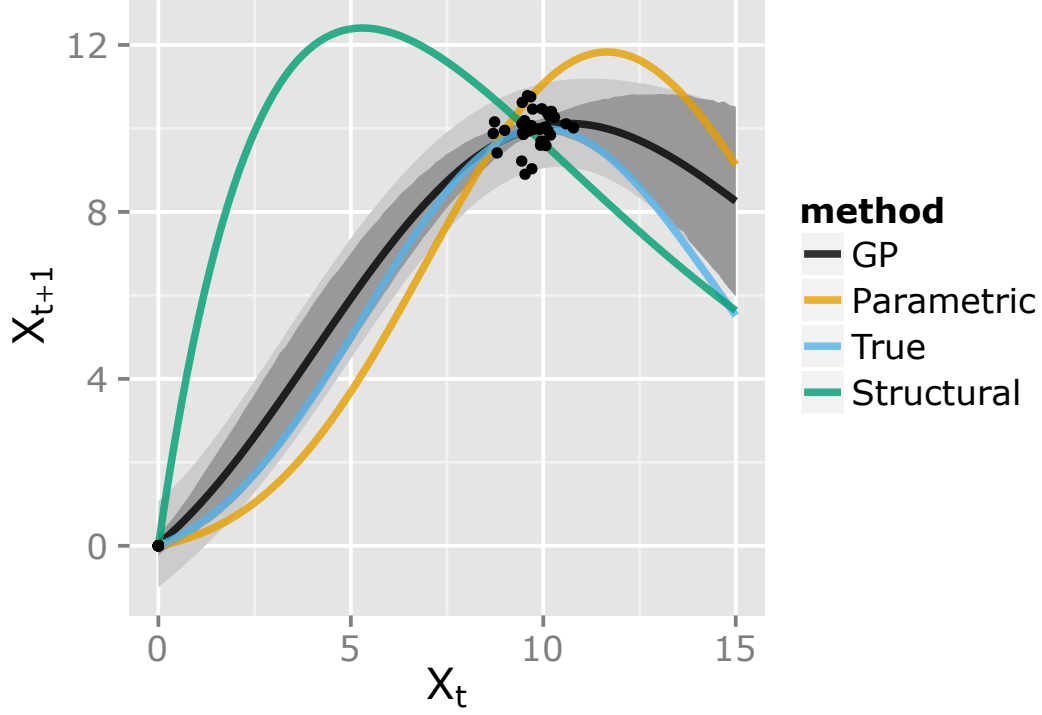


Figure 1: Graph of the inferred Gaussian process compared to the true process and maximum-likelihood estimated process. Graph shows the expected value for the function f under each model. Two standard deviations from the estimated Gaussian process covariance with (light grey) and without (darker grey) measurement error are also shown. The training data is also shown as black points. (The GP is conditioned on 0,0, shown as a pseudo-data point).

An example of such a comparison is shown in Figure 1, in which the optimal management decision is determined by stochastic dynamic programming algorithm using the biologically (a) plausible and (b) the implausible model of the population growth.

Discussion

Three elements contribute to the biologically implausible phenomenological model performing better in this context:

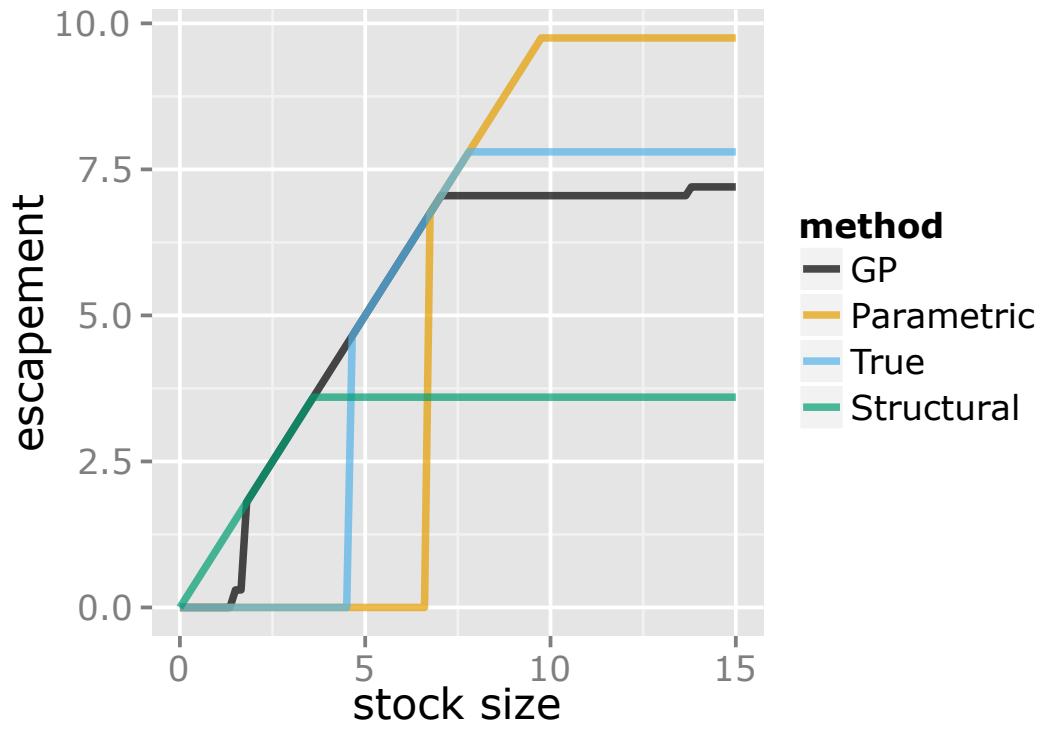


Figure 2: The steady-state optimal policy (infinite boundary) calculated under each model. Policies are shown in terms of target escapement, S_t , as under models such as this a constant escapement policy is expected to be optimal (Reed 1979).

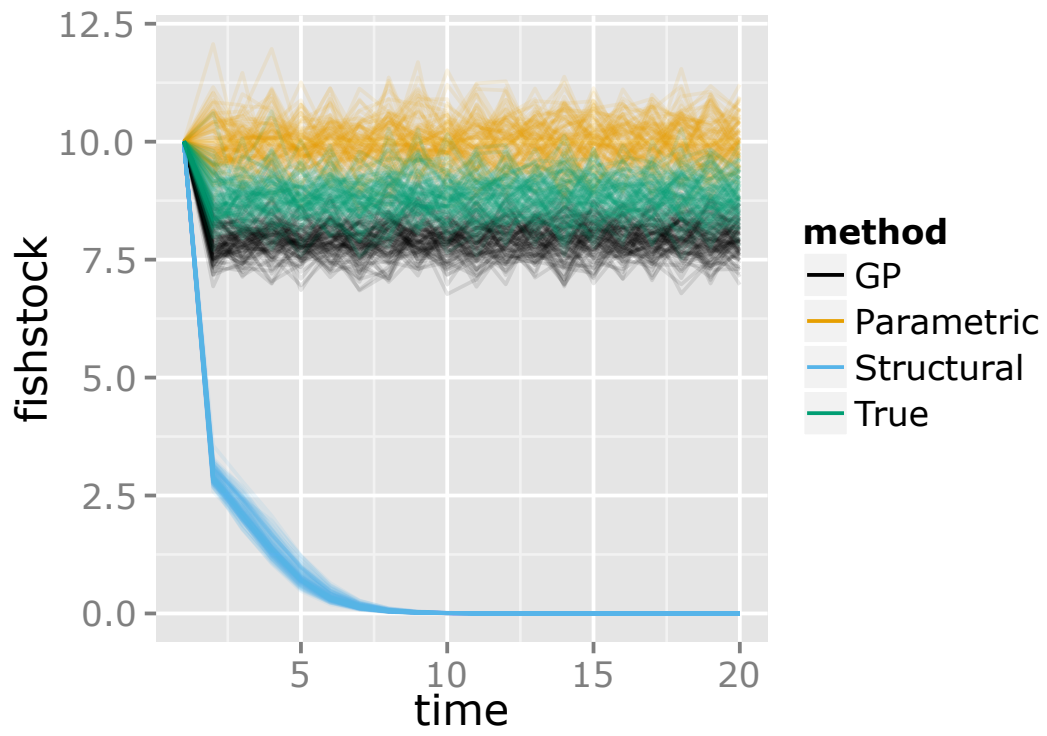


Figure 3: Gaussian process inference outperforms parametric estimates. Shown are 100 replicate simulations of the stock dynamics (eq 1) under the policies derived from each of the estimated models, as well as the policy based on the exact underlying model.

1. Relevant state space

The dynamics occur over a range of state-space in which the biologically implausible model performs as well or better than the more plausible model.

This is the most obvious and immediate reason why the shortcomings that appear to make the model implausible do not make it useless. This effect is enhanced in our example because the management actions help drive the system towards rather than away from this region of state-space.

2. Predictive accuracy

This application relies only on the predictive accuracy of the model, not an interpretation of the parameters. Predictive accuracy is not the goal of all modeling, as ecologists have been observing for as long as they made models (perhaps none more memorably than @Levins1969).

Mechanistic modeling is at its most powerful not when it is used to directly forecast future states but when it provides an understanding of how to approach a problem. SIR-type models from the epidemiological literature are a good example. While the simplest SIR models have little predictive power over the outbreak intensity or timing at a particular location, they provide a powerful understanding of the spread of an infection in terms of a single, biologically meaningful parameter: R_0 , the basic reproductive number. From the model, it becomes clear that management need not vaccinate every member of the population to stop the spread, but rather it suffices to vaccinate a sufficient fraction of the population to reduce R_0 below 1.

3. Time scale for new data

In the sequential decision making problem we considered, we are presented with new data after each action. The relevant timescale for the prediction is thus not the long-term dynamics, which would be wildly divergent, but the dynamics over this much shorter interval. While ecologists may be hesitant to base continual management on a model with obviously inaccurate long-term behavior, engineers tend to consider the problem in frequency space and gravitate to the opposite position – a good control model should prioritize high-frequency accuracy over low frequency accuracy. The differences in intuition may arise from the timescales at which each profession can typically adjust their control variables – much faster for a control system of a chemical plant than state policy for a natural resource. Still, the lesson is clear: when facing repeated management decisions over a short timescale, such as setting annual harvests of a fishery, it may be more valuable to use a machine learning algorithm that makes accurate year-ahead predictions that capture some of the high-frequency fluctuations that appear only as noise in a mechanistic model of the long-term population dynamics.

Advantages of non-mechanistic models

1. Can better use all of the data available.

Incorporating various sources of information into mechanistic models can be an immensely difficult due to the increased complexity involved. Only thanks to tandem advances in increasing computational power and hierarchical statistical methodology have we been able to tackle such intuitively important complexity (and the potentially new available data that accompanies it) such as spatial distribution, heterogeneities of space, time, and individuals, to shift to ecosystem-based approaches from single-species based approaches. Without the need to formulate mechanisms, many modern machine learning algorithms can leverage potential information from all available sources of data directly. The algorithms can recognize unanticipated or subtle patterns in large data sets that enable more accurate predictions than mechanistic models that are formulated at a more macroscopic level. Such approaches depend critically

2. Better express ambiguity in regions where data is not available

One of the greatest strengths of mechanistic models is their greatest weakness as well.

Better the devil you know? The dangers of pattern-based approaches

When all we have identified is a pattern rather than a mechanism, there is little way to guarantee that the pattern will not change in the future. Few examples illustrate this danger more broadly than the sub-prime mortgage crisis of 2008. Simplifying greatly, the unquenchable demand for mortgage backed securities had shifted the underlying dynamics on which the phenomenological models were based. Historical data indicated that mortgage-backed securities involved little risk. As demand exhausted the supply of traditional mortgages, it became profitable to invent new mortgages that could be sold to higher risk customers, such as those without proof of income. The pattern-based models did not have enough time or data to learn about the higher default rate occurring as a result, but remained anchored too in the old patterns where default rates were low. While a mechanistic formulation may have anticipated the resulting higher risk, the pattern based model must learn by seeing it. The economy still struggles to pay off that lesson.

Of course mechanistic models may also fail to predict sudden structural changes. It is all too easy to overlook some slowly changing environmental variable, particularly if temporal changes in a parameter must be modeled explicitly if they are to be included at all. Machine learning approaches are frequently less rigid in their assumptions, but can also less transparent about them. This creates a second weakness in the machine learning approach.

Mechanistic models are not free from this danger either. In fact they can be more susceptible to them for the very same reasons. A mechanistic model has the ability to more convincingly produce out-of-sample predictions or extrapolation. This may allow us to determine that a virus will spread or an endangered species will go extinct even before we have observed

Dangers of biased estimation.

Parametric models are frequently used phenomenologically, rather than being derived by plausible underlying mechanisms. @Geritz2012 condemn this practice on the basis of the common misinterpretations that result. These

mistakes arise from ignoring or misinterpreting the biological meaning of the model parameters. The use of delay equations is a frequent offender – for instance, one is hard-pressed to describe a mechanistic model consistent with Hutchinson’s 1948 delay-logistic. Assuming either scramble competition or interference competition gives rise to two alternative delay equations, both structurally different from the original.

since the late 1980s there has been some consensus amongst ecologists that management decisions are best guided by models which are grounded in ecological theory, and which strike a balance between too much or too little detail describing the relevant processes (e.g., DeAngelis 1988, Starfield 1997, Jackson et al. 2000, Carpenter 2003, Nelson et al. 2008)

@Geritz2012 @Cuddington2013

Managing by rule-of-thumb

Managing by algorithm

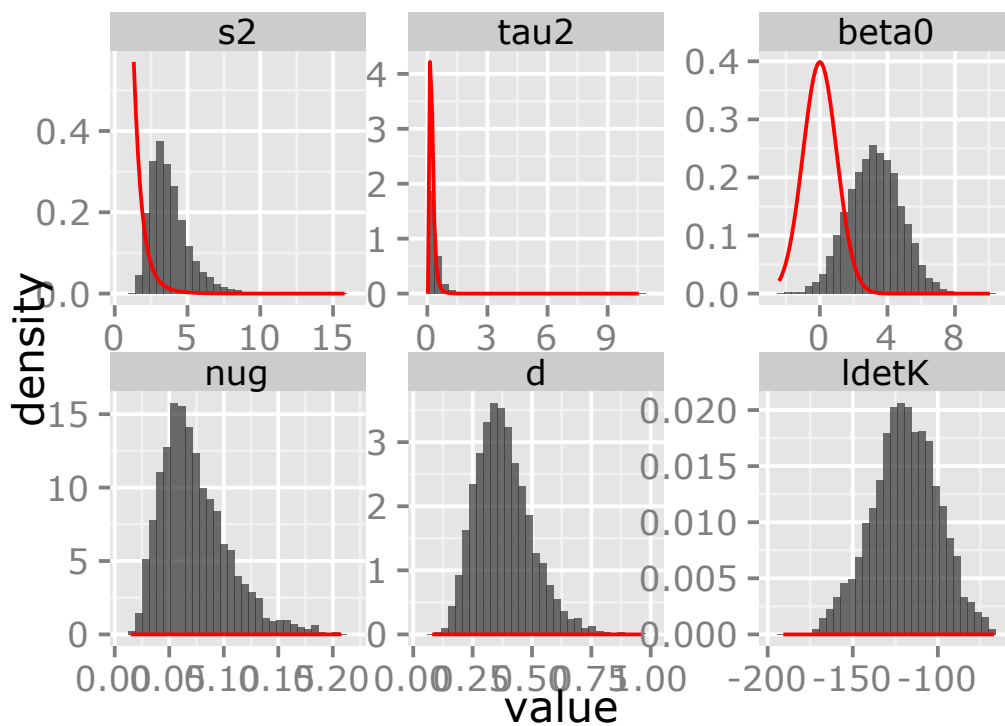


Figure 4: Histogram of posterior distributions for the estimated Gaussian Process shown in Figure 1. Prior distributions overlaid.

Reed, William J. 1979. “Optimal escapement levels in stochastic and deterministic harvesting models.” *Journal of Environmental Economics and Management* 6 (4) (dec): 350–363. doi:10.1016/0095-0696(79)90014-7.