PeerJ Preprints NOT PEER-REVIEWED

Adoption of Machine Learning Techniques in Ecology and Earth Science

Anne E Thessen
The Ronin Institute for Independent Scholarship, Montclair, NJ, USA
The Data Detektiv, Waltham, MA, USA

ABSTRACT

The natural sciences, such as ecology and earth science, study complex interactions between biotic and abiotic systems in order to infer understanding and make predictions. Machine-learning-based methods have an advantage over traditional statistical methods in studying these systems because the former do not impose unrealistic assumptions (such as linearity), are capable of inferring missing data, and can reduce long-term expert annotation burden. Thus, a wider adoption of machine learning methods in ecology and earth science has the potential to greatly accelerate the pace and quality of science. Despite these advantages, machine learning techniques have not had wide spread adoption in ecology and earth science. This is largely due to 1) a lack of communication and collaboration between the machine learning research community and natural scientists, 2) a lack of easily accessible tools and services, and 3) the requirement for a robust training and test data set. These impediments can be overcome through financial support for collaborative work and the development of tools and services facilitating ML use. Natural scientists who have not yet used machine learning methods can be introduced to these techniques through Random Forest, a method that is easy to implement and performs well. This manuscript will 1) briefly describe several popular ML methods and their application to ecology and earth science, 2) discuss why ML methods are underutilized in natural science, and 3) propose solutions for barriers preventing wider ML adoption.

INTRODUCTION

Machine Learning (ML) is a discipline of computer science that develops dynamic algorithms capable of data-driven decisions, in contrast to models that follow static programming instructions. In 1959, Arthur Samuel first defined ML as a “Field of study that gives computers the ability to learn without being explicitly programmed”. The very first mention of ‘machine learning’ in the literature occurred in 1930 and use of the term has been growing steadily since 1980 (Fig. 1). While discussion of ML is likely to recall scenes from popular science-fiction books and movies, there are many practical applications of ML in a wide variety of disciplines from medicine to finance. Part of what makes ML so broadly applicable is the diversity of ML algorithms capable of performing very well under messy, real-world conditions. Despite, and perhaps because of this versatility, uptake of ML applications have lagged behind traditional statistical techniques in the natural sciences.

The advantage of ML over traditional statistical techniques, especially in earth science and ecology, is the ability to model highly dimensional and non-linear data with complex interactions and missing values (De’ath & Fabricius 2000; Recknagel 2001; Olden et al. 2008; Haupt, Pasini, et al. 2009; Knudby, Brenning, et al. 2010). Ecological data specifically are known to be non-linear and highly dimensional with intense interaction effects; yet, methods that

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

assume linearity and are unable to cope with interaction effects are still being used (Knudby, Brenning, et al. 2010; Olden et al. 2008). To make these methods work, researchers cope in various ways, including 1) data transformations, which can limit the interpretability of the final results (Knudby, Brenning, et al. 2010) 2) decompose/recompose methods to break up the system into bits with fewer complicated dynamics (Pasini 2009) or 3) assuming linearity without any modification of the data (Džeroski 2001). Several comparative studies have already shown that ML techniques can outperform traditional statistical approaches in a wide variety of problems in earth science and ecology (Lek, Delacoste, et al. 1996; Levine et al. 1996; Lawler et al. 2006; Prasad et al. 2006; Cutler et al. 2007; Olden et al. 2008; Zhao et al. 2011; Bhattacharya 2013; Manel et al. 2001; Segurado & Araújo 2004; Elith et al. 2006); however, comparing techniques can be difficult and requires careful consideration (Fielding 2007).

The exact division between ML methods and traditional statistical techniques is not always clear and ML methods are not always better than traditional statistics. For example, a system may not be linear, but a linear approximation of that system may still yield the best predictor. The exact method(s) must be chosen based on the problem at hand and a meta approach that considers the results of multiple algorithms may be best. This manuscript will discuss six types of ML methods and their relative strengths and weaknesses in ecology and earth science. Specific applications of ML in ecology and earth science will be briefly reviewed with the reasons ML methods are underutilized in natural sciences. Potential solutions will be proposed.

BACKGROUND

The basic premise of ML is that a machine (i.e., algorithm or model) is able to make new predictions based on data. Some algorithms are supervised, meaning they are shown data *a priori* and then make predictions about new data based on the previous data. Some are unsupervised, meaning they can make predictions with no *a priori* data. Some are a combination of the two, (i.e., semi-supervised). The basic technique behind all ML methods is an iterative combination of statistics and error minimization or reward maximization, applied and combined in varying degrees. Many ML algorithms iteratively check all or a very high number of possible outcomes to find the best result, with “best” defined by the user for the problem at hand. The potentially high number of iterations is prohibitive of manual calculations and is a large part of why these methods are only now widely available to individual researchers.

Computing power has increased such that ML methods can be implemented with a desktop or even a laptop. The availability of high-performance computing and the maturation of the internet has opened up an even broader array of possibilities for individuals. Before the current availability of computing power, ecologists and earth scientists had to settle for statistical methods that assumed linearity (Knudby, Brenning, et al. 2010) and limited, controlled experiments (Fielding 1999a). Both of these restrictions limit scale of studies and accuracy of results. A similar acceleration has been observed for numerical modeling of natural systems, where model predictions have improved because increased computing power has allowed for the inclusion of more parameters and, more importantly, finer granularity (see Semtner 1995; Forget et al. 2015 for examples in oceanography).

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

The first step in applying ML is teaching the algorithm using a training data set. The training data set is a collection independent variables with the corresponding dependent variables. The machine uses the training data to "learn" how the independent variables (input) relate to the dependent variable (output). Later, when the algorithm is applied to new input data, it can apply that relationship and return a prediction. After the algorithm is trained, it needs to be tested to get a measure of how well it can make predictions from new data. This requires another data set with independent and dependent variables, but the dependent variables are not provided to the learner. The algorithm predictions are compared to the withheld data to determine the quality of the predictions. This process requires a data set that is large enough to be split in two for training and testing. The type of ML method, the size and nature of the training and test data set, and the evaluation method should be chosen to optimize the trade-off between bias and accuracy to give a meaningful result for the problem at hand.

Tree-based methods

Tree-based ML methods include decision trees, classification trees, and regression trees (Olden et al. 2008; Hsieh 2009; Kampichler et al. 2010). For these methods, a tree is built by iteratively splitting the data set based on a rule that results in the divided groups being more homogeneous than the group before (Fig. 2). The rules used to split the tree are identified by an exhaustive search algorithm and give insight into the workings of the modeled system. A single decision tree can give vastly different results depending on the training data and typically has low predictive power (Iverson et al. 2004; Olden et al. 2008; Breiman 2001b). Several ensemble-tree methods have been developed to improve predictive power by combining the results of multiple trees, including boosted trees and bagged trees (Breiman 1996; De'ath 2007). A boosted tree results from a pool of trees created by iteratively fitting new trees to minimize the residual errors of the existing pool (De'ath 2007). The final boosted tree is a linear combination of all the trees (Elith et al. 2008). Bagging is a method that builds multiple trees on subsamples of the training data (bootstrap with replacement) and then averages the predictions from each tree to get the bagged predictions (Breiman 1996; Knudby, Brenning, et al. 2010).

Random Forest is a relatively new tree-based method that fits a user-selected number of trees to a data set and then combines the predictions from all trees (Breiman 2001a). The Random Forest algorithm creates a tree for a subsample of the data set. At every decision only a randomly selected subset of variables are used for the partitioning. The predicted class of an observation in the final tree is calculated by majority vote of the predictions for that observation in all trees with ties split randomly.

Ensemble tree-based methods, especially Random Forest, have been demonstrated to outperform traditional statistical methods and other ML methods in earth science and ecology applications (Cutler et al. 2007; Kampichler et al. 2010; Knudby, Brenning, et al. 2010). They can cope with small sample sizes, mixed data types, and missing data (Cutler et al. 2007; Olden et al. 2008). The single-tree methods are fast to calculate and the results are easy to interpret (Kampichler et al. 2010), but they are susceptible to overfitting (Olden et al. 2008) and frequently require "pruning" of terminal nodes that do not give enough additional accuracy to justify the increased complexity caused by its presence (Breiman et al. 1984; Garzón et al. 2006; Cutler et al. 2007; Olden et al. 2008; Džeroski 2009). The ensemble-tree methods can be

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

computationally expensive (Olden et al. 2008; Džeroski 2009; Cutler et al. 2007), but resist overfitting (Breiman 2001a). Random Forest algorithms can provide measures of relative variable importance and data point similarity that can be useful in other analyses (Cutler et al. 2007), but can be clouded by correlations between independent variables (Olden et al. 2008). Implementing Random Forest is relatively straightforward. Only a few, easy-to-understand parameters need to be provided by the user (Kampichler et al. 2010), but the final Random Forest does not have a simple representation that characterizes the whole function (Cutler et al. 2007). Tree methods also do not give probabilities for results, which means that data are classified into categories, but the probability that the classification is correct is not given.

Artificial Neural Networks

An Artificial Neural Network (ANN) is a ML approach inspired by the way neurological systems process information (Recknagel 2001; Olden et al. 2008; Boddy & Morris 1999; Hsieh 2009). There are many types of ANNs that can be supervised or unsupervised learners, but only a few are typically used in earth science and ecology (Pineda 1987; Kohonen 1989; Chon et al. 1996; Recknagel 2001; Lek & Guégan 2000). An ANN has three parts: 1) the input layer, 2) the hidden layer, and 3) the output layer (Fig. 3). Each layer is made up of several "neurons". Each neuron is connected to all the other neurons in the neighboring layer, but not the neurons in the same layer or in non-adjacent layers. The input layer contains one neuron for every independent variable. The output layer can have one neuron (for binary or continuous output) or more (for categorical output). The number of neurons in the hidden layer can be changed by the user to optimize the trade-off between overfitting and variance (Geman et al. 1992). Too many neurons in this layer can lead to overfitting. Each neuron has an activity level and each connection has a weight. The activity level of the input neurons are set by the value of the independent variable. Training the ANN involves an algorithmic search for an optimal set of connection weights that produces an output value with a small error relative to the observed value. Performance can be sensitive to initial connection weights and the number of hidden neurons, so multiple networks should be processed while varying these parameters (Olden et al. 2008).

ANN can be a powerful modeling tool when the underlying relationships are unknown and the data are imprecise and noisy (Lek & Guégan 1999). Interpretation of the ANN can be difficult and neural networks are often referred to as a "black box" method (Lek & Guégan 1999; Olden et al. 2008; Wieland & Mirschel 2008; Kampichler et al. 2010). ANNs can be more complicated to implement and are more computationally expensive than tree-based ML methods (Olden et al. 2008), but ANNs can accommodate a major gain in computational speed with a minor sacrifice in accuracy. For example, an ANN with one fourth the computational cost of a traditional satellite data retrieval algorithm (that uses an iterative method) can come to within 1/10 of the traditional algorithm accuracy (Young 2009). The user-defined parameters, such as the number of hidden neurons and the initial connection weights, can be complicated and overfitting can be a problem (Kampichler et al. 2010). Many ANNs mimic standard statistical methods (A. Fielding pers. comm.), so a good practice while using ANNs is to also include a rigorous suite of validation tests and a general linear model for comparison (Özesmi et al. 2006).

Support Vector Machines

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

A Support Vector Machine (SVM) finds a plane that divides two classes in a feature space. It does this by finding the plane giving the largest gap between two classes, the Maximal Margin Classifier (MMC), which is typically the plane that is equidistant from the points in each class closest to the boundary (Fig. 4A; Moguerza & Muñoz 2006; Rasmussen & Williams 2006; Zhao et al. 2008; Hsieh 2009; Kampichler et al. 2010; Zhao et al. 2011). A new data point would be classified according to which side of the decision boundary it fell. This MMC works very well on data that are easily separated by a straight line, but data are often noisy and cannot be separated by a straight plane. In this case, a Support Vector Classifier can be used to create a "buffer zone" around the hard decision boundary (Fig. 4B). Some data sets cannot be divided by a linear decision boundary. In this case a type of algorithm known as a kernel computes the mean squared error of the product of a pairwise multiplication of every data point and uses that to draw decision boundaries around groups of data.

SVMs perform very well on binary classification tasks, especially when classifying images or sounds (Durbha et al. 2007; Acevedo et al. 2009; Zhao et al. 2011; Duro et al. 2012). They can distinguish more classes and cope with non-linear decision boundaries with additional modification (Kreßel 1999; Hsieh 2009), but with these modifications are not guaranteed to converge to the optimal classifier and can be difficult to interpret (Lee et al. 2004). Like tree methods, SVM is not a probability model and does not assign probabilities to its classifications. When there is some overlap in the data and a support vector classifier becomes necessary, logistic regression can be more useful than SVM because it calculates probabilities. When the decision boundaries are non-linear SVM performs better than logistic regression and linear discriminant analysis because of the kernels.

Genetic Algorithms

Genetic Algorithms (GA) are based on the process of evolution in natural systems in that a population of competing solutions evolves over time to converge on an optimal solution (Holland 1975; Goldberg & Holland 1988; Olden et al. 2008; Koza 1992; Haupt & Haupt 2004). Solutions are represented as "chromosomes" and model parameters are represented as "genes" on those chromosomes (Fig. 5). Training a GA has four steps: 1) random potential solutions are generated (chromosomes), 2) potential solutions are altered using "mutation", and "recombination", 3) solutions are evaluated to determine fitness (minimizing error), and 4) the best solutions cycle back to step 2 (Holland 1975; Mitchell 1998; Haefner 2005). Each cycle represents a "generation". Depending on the nature of the problem the GA is trying to solve, the chromosome can be strings of bits, real values, rules, or permutations of elements (Recknagel 2001).

An advantage of GA is the removal of the often arbitrary process of choosing a model to apply to the data (Jeffers 1999). In a GA, multiple models are compared. GAs have seen a rise in popularity due to development of the Genetic Algorithm for Rule-Set Prediction (GARP) used to predict species distributions (Stockwell & Noble 1992). GAs are very popular in hydrology (see Mulligan & Brown 1998 for description of how GA was used to find the Pareto Front) and meteorology (Haupt 2009). GAs are able to cope with uneven sampling and small sample sizes (Olden et al. 2008). GAs were developed with broad application in mind and can use a wide range of model structures and model-fitting approaches (Olden et al. 2008). As a result, a larger

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

burden is placed on the user to select complicated model parameters with little guidance, and the fixed-length "chromosomes" can limit the potential range of solutions (Olden et al. 2008). GAs are not best for all problems and many traditional statistical techniques can perform just as well or better (Olden et al. 2008). GARP, in particular, can be susceptible to overfitting (Elith et al. 2008; Lawler et al. 2006).

Logic-Based Methods

Logic-based methods, such as Fuzzy Logic and Inductive Logic Programming, (Fig. 6) provide a practical approach to automating complex analysis and inference in a long workflow (Williams et al. 2009). They represent and process knowledge in terms of natural language in a set of "if/then" rules (Wieland 2008). The if/then rules are created through an algorithm that iteratively selects each class and refines the if-statement until only the selected class remains (Džeroski 2009). The National Center for Atmospheric Research (NCAR) has developed three logic-based algorithms to address a complex problem in meteorology (Williams et al. 2009) and logic-based methods have been widely used in ecology for the induction of rules for classification problems (Bouchon-Meunier et al. 2007).

Logic-based algorithms and the resulting rules can be easy to understand and interpret, as long as the rule sets are not too large, but overfitting can be a problem (Kampichler et al. 2010).

Bayesian Classifier

Bayesian ML methods are based on Bayesian statistical inference, which started in the 18th century with the development of Bayes' theorem (Laplace 1986). These methods are based on expressing the true state of the world in terms of probabilities and then updating the probabilities as evidence is acquired (Bishop 2006). In most cases, it is important to know the probability that a new datum belongs to a given class, not just the class. A Bayesian classifier calculates a probability density for each class (Fig. 7A). The probability density is a curve showing, for any given value of the independent variable, the likelihood of being a member of that class (Fig. 7). The new datum is assigned to the class with the highest probability. The values of the independent variable that have an equal probability of being in either class are known as the decision boundary and this marks the dividing line between the classes. In the real world, it can be difficult to calculate these *a priori* probabilities and the user must often make a best-guess.

A Bayesian classifier gives good results in most cases and requires fewer training data compared to other ML methods. It is useful when there are more than two distinct classes. The disadvantage is that it can be very hard to specify prior probabilities and results can be quite sensitive to the selected prior. This method does assume that variables are independent, which is not always true. Some Baysian classifiers have Gaussian assumptions which may not be reasonable for the problem at hand. Another issue is that if a specific feature never appears in a class, the resulting zero probability will complicate calculations; therefore, a small probability must often be added, even if the feature does not appear in the class.

Using ML in Earth Science and Ecology

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

For many researchers, machine learning is a relatively new paradigm that has only recently become accessible with the development of modern computing. While the adoption of ML methods in earth science and ecology has been slow, there are several published studies using ML in these disciplines (e.g., Park & Chon 2007). The following is a brief review of the different published applications of ML in earth science and ecology.

Habitat Modeling and Species Distribution

Understanding the habitat requirements of a species is important for understanding its ecology and managing its conservation. Habitat modelers are interested in using multiple data sets to make predictions and classifications about habitat characteristics and where taxa are likely to be located or engaging in a specific behavior (e.g., nesting Cutler et al. 2007; Fielding 2009). The rule-sets developed are referred to as Species Distribution Models (SDM) and can use a wide variety of ML methods or none at all (Guisan & Thuiller 2005). Typically, an algorithm would be trained using a data set matching environmental variables to taxon abundance or presence/absence data. If the algorithm tests well, it can be given a suite of environmental variables from a different location to make predictions about what taxa are present. This technique has been used to identify current suitable habitat for specific taxa, model future species distributions including predicting invasive and rare species presence, and predict biodiversity of an area (Tan & Smeins 1996; Kampichler et al. 2000; Cutler et al. 2007; Olden et al. 2008; Knudby, Brenning, et al. 2010). Common tools include Random Forest (Cutler et al. 2007; Peters et al. 2007), classification and decision trees (Ribic & Ainley 1997; Kobler & Adamic 2000; Bell 1999; Vayssières et al. 2000; Debeljak et al. 2001; Miller & Franklin 2002), neural networks (Mastrorillo et al. 1997; Guégan et al. 1998; Özesmi et al. 2006; Brosse et al. 2001; Thuiller 2003; Fielding 1999a; Dedecker et al. 2004; Manel et al. 2001; Segurado & Araújo 2004), genetic algorithms (D'Angelo et al. 1995; Stockwell & Peters 1999; McKay 2001; Wiley et al. 2003; Termansen et al. 2006; Stockwell 1999; Peterson et al. 2002), and Bayesian classifiers (Fischer 1990; Brzeziecki et al. 1993; Guisan & Zimmermann 2000).

Species Identification

Identifying taxa can require specialized knowledge only possessed by a very few and the data set requiring expert curation can be large (e.g., automated collection of images and sounds). Thus, this step is a major bottleneck in biodiversity studies. In order to increase throughput, algorithms are trained on images, sounds, and other types of data labeled with taxon names. (For more information about automated taxon identification specifically, see Edwards et al. (1987) and MacLeod (2007).) Then, algorithms are shown new data and asked to classify them to genus or species. This technique has been used to identify plankton, spiders, and shellfish larvae from images (Boddy & Morris 1999; Sosik & Olson 2007; Goodwin et al. 2014; Do et al. 1999). Audio files of amphibian, bird, bat, insect, elephant, cetacean, and deer sounds have been classified to species using ML techniques (Acevedo et al. 2009; Armitage & Ober 2010; Kasten et al. 2010; Parsons & Jones 2000; Jennings et al. 2008; Chesmore 2004). Fish and algal species have been identified using acoustic (Simmonds et al. 1996) and optical characteristics (Boddy et al. 1994; Balfoort et al. 1992). ML has been used to differentiate between the radar signals of birds and abiotic objects (Rosa et al. 2016). In some cases, individuals of the same species can be

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

distinguished even if the individuals themselves are unknown *a priori* (Reby et al. 1998; Fielding 1999a). Common tools include support vector machines (Acevedo et al. 2009; Armitage & Ober 2010; Goodwin et al. 2014; Sosik & Olson 2007; Fagerlund 2007; Rosa et al. 2016), Random Forest (Armitage & Ober 2010; Rosa et al. 2016), Bayesian classifiers (Fielding 1999a), genetic algorithms (Jeffers 1999), and neural networks (Armitage & Ober 2010; Parsons & Jones 2000; Simmonds et al. 1996; Jennings et al. 2008; Do et al. 1999; Boddy et al. 1994; Balfoort et al. 1992; Rosa et al. 2016).

Remote Sensing

Satellite images and other data gathered from sensors at great elevation (e.g., LIDAR) are an excellent way to gather large amounts of data about Earth over broad spatial scales. In order to be useful, these data must go through some minimum level of processing (Atkinson & Tatnall 1997) and are often classified into land cover or land use categories (Guisan & Zimmermann 2000). ML methods have been developed to automate these laborious processes (Fitzgerald & Lees 1992; Lees 1996; Lees & Ritman 1991; Atkinson & Tatnall 1997; Pal 2005; Ham et al. 2005; Gislason et al. 2006; Guisan & Zimmermann 2000; Lakshmanan 2009). After processing, the data are available for research and ML can be used here too. ML methods can be used to infer geophysical parameters from remote sensing data, such as inferring the Leaf Area Index from Moderate Resolution Imaging Spectrometer data (Rumelhart et al. 1986; Krasnopolsky 2009; Hsieh 2009). Sometimes remote sensing data and the parameters inferred from them can require spatial interpolation in the vertical or horizontal dimension, which is often performed using ML methods (Li et al. 2011; Krasnopolsky 2009). Common tools for classifying remote sensing images include Random Forest (Knudby, LeDrew, et al. 2010; Duro et al. 2012), support vector machines (Durbha et al. 2007; Knudby, LeDrew, et al. 2010; Zhao et al. 2011; Duro et al. 2012), neural networks (Rogan et al. 2008), genetic algorithms (Haupt 2009), and decision trees (Huang & Jensen 1997).

Resource Management

Making decisions about conservation and resource management can be very difficult because there is often not enough data for certainty and the consequences of being wrong can be disastrous. ML methods can provide a means of increasing certainty and improving results, despite data gaps. Several algorithms have been applied to water (Maier & Dandy 2000; Haupt 2009), soil (Henderson et al. 2005; Tscherko et al. 2007), and biodiversity/wildlife management (Baran et al. 1996; Lek, Delacoste, et al. 1996; Lek, Belaud, et al. 1996; Giske et al. 1998; Guégan et al. 1998; Spitz & Lek 1999; Chen et al. 2000; Vander Zanden et al. 2004; Sarkar et al. 2006; Worner & Gevrey 2006; Cutler et al. 2007; Quintero et al. 2014; Bland et al. 2014; Jones et al. 2006). ML methods have been used to model population dynamics, production, and biomass in terrestrial, aquatic, marine, and agricultural systems (Scardi 1996; Recknagel 1997; Scardi & Harding Jr. 1999; Recknagel et al. 2000; McKenna Jr. 2005; Muttil & Lee 2005; Recknagel et al. 2002; Džeroski 2001; Schultz et al. 2000). Some specific examples of ML applications in resource management and conservation include 1) inference of IUCN (International Union for Conservation of Nature) conservation status of Data Deficient species using intrinsic and extrinsic characters (Quintero et al. 2014; Bland et al. 2014), 2) predicting

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

farmer risk preferences (Kastens & Featherstone 1996), 3) predicting the production and biomass of various animal populations (Brey et al. 1996), 4) examining the effect of urbanization on bird breeding (Lee et al. 20007), 5) predicting disease risk (Guo et al. 2005; Furlanello et al. 2003), and 6) modeling ecological niches (Drake et al. 2006). Being able to make these types of predictions and inferences can help focus conservation efforts for maximum impact (Knudby, Brenning, et al. 2010; Guisan et al. 2013). Common ML methods for resource management include genetic algorithms (Haupt 2009), neural networks (Brey et al. 1996; Kastens & Featherstone 1996; Recknagel 1997; Giske et al. 1998; Guégan et al. 1998; Schultz et al. 2000; Lee et al. 20007), support vector machines (Guo et al. 2005; Drake et al. 2006), fuzzy logic (Tscherko et al. 2007), decision trees (Henderson et al. 2005; Jones et al. 2006), and Random Forest (Cutler et al. 2007; Quintero et al. 2014; Furlanello et al. 2003).

Forecasting

Discovery of deterministic chaos in meteorological models (Lorenz 1963) led to reconsideration of the use of traditional statistical methods in forecasting (Pasini 2009). Today, predictions about weather are often made using ML methods. The most common ML methods used in meteorological forecasting are genetic algorithms, which have been used to model rainy vs non-rainy days (Haupt 2009) and severe weather (Hsieh 2009). Forecasting can be important for applications other than weather prediction. In atmospheric science, neural networks are able to find dynamics hidden in noise and successfully forecast important variables in the atmospheric boundary layer (Pasini 2009). The oceanography community makes extensive use of neural networks for forecasting sea level, waves, and sea surface temperature (Wu et al. 2006; Hsieh 2009). In addition to being directly used for forecasting, neural networks are commonly used for downscaling environmental and model output data sets used in making forecasts (Casaioli et al. 2003; Marzban 2003; Hsieh & Hsieh 2003).

Environmental Protection and Safety

Just as ML can help resource managers make important decisions with or without adequate data coverage, environmental protection and safety decisions can be aided with ML methods when data are sparse. ML has been used to classify environmental samples into inferred quality classes in situations where direct analyses are too costly (Džeroski 2001). The mutagenicity, carcinogenicity, and biodegradability of chemicals have been predicted based on structure without lengthy lab work (Džeroski 2001). Sources of air contaminants have been identified and characterized in spite of lack of *a priori* knowledge about source location, emission rate, and time of release (Haupt, Allen, et al. 2009). ML can relate pollution exposure to human health outcomes (Džeroski 2001). Common ML methods for environmental protection include genetic algorithms (Haupt, Allen, et al. 2009), Bayesian classifiers (Walley et al. 1992), neural networks (Ruck et al. 1993; Walley & Džeroski 1996; Walley et al. 2000), and fuzzy logic (Srinivasan et al. 1997; Džeroski 2001; Džeroski et al. 1999).

Climate Change Studies

One of the more pressing societal problems is the mitigation of and adaptation to climate change. Policy-makers require well-formed predictions in order to make decisions, but the

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

complexity of the climate system, the interdisciplinary nature of the problem, and the data structures prevents the effective use of linear modeling techniques. ML is used to study important processes such as El Niño, the Quasi Biennial Oscillation, the Madden-Julian Oscillation, and monsoon modes (Cavazos et al. 2002; Pasini 2009; Krasnopolsky 2009; Hsieh 2009), and to predict climate change itself (Casaioli et al. 2003; Marzban 2003; Hsieh & Hsieh 2003; Pasini 2009). Predictions about the greenhouse effect (Seginer et al. 1994) and environmental change (Guisan & Zimmermann 2000) have also been made using ML. A very common use of ML in climate science is downscaling and post processing data from General Circulation Models (refs in Pasini 2009; Hsieh 2009). Ecological niche modeling and predictive vegetation mapping (as discussed above) can help predict adaptation to climate change (Wiley et al. 2003; Iverson et al. 2004). The most commonly used ML method in climate change studies is the neural network (Guisan & Zimmermann 2000; Pasini 2009).

DISCUSSION

How can ML advance ecology and earth science?

The application of ML methods in ecology and earth science has already demonstrated the potential for increasing the quality and accelerating the pace of science. One of the more obvious ways ML does this is by coping with data gaps. The Earth is under-sampled, despite spending hundreds of millions of dollars on earth and environmental science (e.g., Webb et al. 2010). Where possible, ML allows a researcher to use data that are plentiful or easy to collect to infer data that are scarce or hard to collect (e.g., Wiley et al. 2003; Edwards Jr. et al. 2005; Buddemeier et al. 2008). Conservation managers are particularly well positioned to take advantage of ML via SDMs in invasive species management, critical habitat identification, and reserve selection (Guisan et al. 2013). Depending on the ML method used, one can also learn more about how a system works, for example through the Random Forest Variable Importance analysis. Another important way ML can fill in data gaps is through downscaling and performing spatial interpolation (Li et al. 2011). There will never be enough research funding to sample everything all of the time. ML can be a tractable method for addressing the data gaps that prevent scientific progress.

ML can accelerate the pace of science by quickly performing complex classification tasks normally performed by a human. A bottleneck in many ecology and earth science workflows are the manual steps performed by an expert, usually a classification task such as identifying a species. Rather than having all of the data classified by an expert, the expert only needs to review enough data to train and test an algorithm. Expert annotation can be even more time consuming when the expert must search through a large volume of data, like a sensor stream, for a desired signal (Kasten et al. 2010). This bottleneck has been addressed for some types of taxon identification (Cornuet et al. 1996; Acevedo et al. 2009; Armitage & Ober 2010; Sosik & Olson 2007), finding relevant data in sensor streams (Kasten et al. 2010), and building a reference knowledgebase (Huang & Jensen 1997). In addition to relieving a bottleneck, ML methods can sometimes perform tasks more consistently than experts, especially when there are many categories and the task continues over a long period of time (Culverhouse et al. 2003; Jennings et al. 2008). In these cases, ML methods can improve the quality of science by providing more quantitative and consistent data (Olden et al. 2008; Acevedo et al. 2009; Sutherland et al. 2004).

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

As discussed above, ML techniques can perform better than traditional statistical methods in systems that are poorly represented by linear models, but a direct comparison of performance between ML techniques and traditional statistical methods is difficult because there is no universal measure of performance and results can be very situation-specific (Fielding 2007). The true measure of the utility of a tool is how well it can make predictions from new data and how well it can be generalized to new situations. Highly significant p-values, R^2 values, and accuracy measurements may not reflect this. A study comparing 33 methods with 32 data sets found no real difference in performance and suggested that choice of algorithm be driven by factors other than accuracy, such as the characteristics of the data set (Lim et al. 2000). If the accuracy is not significantly improved using ML, it may be better to use a traditional method that is more familiar and accepted by peers and managers. Best practice is to test multiple methods (including traditional statistics) while probing the trade-off between bias and accuracy and choose the tool that is most useful. In many natural systems, where non-linear and interaction effects are common, a ML-based model is more useful can improve science by building better models. Individual researchers need to select a method based on the specific problem and the data at hand.

Why don't more people use ML?

Even though ML can outperform traditional statistics in some applications (Kampichler et al. 2000; Peters et al. 2007; Pasini 2009; Armitage & Ober 2010; Knudby, Brenning, et al. 2010; Li et al. 2011; Zhao et al. 2011; Bhattacharya 2013; Manel et al. 2001; Segurado & Araújo 2004; Elith et al. 2006), wide acceptance of ML methods in ecology and earth science has not yet happened (Olden et al. 2008). The reasons for this seem to be more social than technical. New methods can be resisted by established scientists, which can delay wide-spread use (Azoulay et al. 2015). ML methods (as well as some more complex statistical models) can require a high degree of math skill to understand in detail, which means either a long familiarization phase or an acceptance of the algorithm as a "black box" (Kampichler et al. 2010). Even some of the names of these methods, such as support vector machine, sound very foreign in the natural sciences. ML methods are highly configurable; thus, it can be overwhelming for researchers to choose the proper test for the job (Kampichler et al. 2010). Many of them need to be run in a "command line" style environment, such as R or MatLab and many ecologists lack the familiarity with command-line-style interfaces (Olden et al. 2008). Alternatively, many of the traditional statistical methods are fast to calculate and give easy-to-interpret metrics, like p-values (Olden et al. 2008; Kampichler et al. 2010). Traditional statistical methods are easier to find as a part of an off-the-shelf software package with a user interface and much of the complicated inner workings pleasantly hidden. All of these make ML methods less attractive than traditional statistical methods.

Another barrier to using ML techniques is the need for adequate training and test data; however, it could be argued that traditional statistical methods should also be subject to validation with test data. It can be hard to have enough quality data to properly train and test an algorithm, especially for automated taxon identification (Lek & Guégan 1999; Wiley et al. 2003; Acevedo et al. 2009; Kampichler et al. 2010). There are techniques available for developing a model when the data set is too small to split into training and test sets (cross-validation,

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Bayesian), but these methods give a weaker estimate of model error (Guisan & Zimmermann 2000; Hsieh 2009).

Finally, the ML research community has done a poor job of communicating the relevance of their discoveries to the natural science research community, largely because there is no professional reward for following through on the application of their results in this domain (Wagstaff 2012). The financial sector is applying ML, suggesting that communication is possible when the potential monetary reward is great enough. There is too much reliance on abstract metrics in the ML research community and not enough consideration of whether or not a particular ML advance will result in a real-world impact (Wagstaff 2012). The small community of ecologists using ML to develop SDMs are not communicating the value of their research to decision-makers and accounts of SDMs being used successfully in conservation are hidden in grey literature (Guisan et al. 2013). A search of 'machine learning' in the Web of Science database returns 104642 results: 46664 from computer science, 20264 from engineering, and 11988 from computational biology. The highest ranking ecology topic in this search was 'environmental sciences ecology' with 1003 results. The highest ranking earth science topic was 'geology' with 677 results. Thus, the vast majority of publications available about machine learning are not connected to ecology or earth science disciplines. Searching for 'machine learning ecology' and 'machine learning earth science' gives many fewer results (144 and 32, respectively, according to Web of Science). As a proportion of total publication output of a discipline, ML has very low representation in ecology (0.11%) and earth science (0.14%), but some sub-disciplines, such as oceanography (1.19%), have a higher proportion of ML-related publications (Fig. 8A). In contrast, a traditional statistical method, linear regression, has a higher representation in discipline output (0.70% in ecology and 0.21% in earth science) with the exception of oceanography (0.54%; Fig. 8B). Communication and collaboration between the ML community, the ecology community, and the earth science community is poor.

Next Steps

How can the use of ML methods in ecology and earth science be encouraged? One barrier that can be lowered is the lack of tools and services to support the application of ML in these domains. The higher use of ML algorithms built with user infrastructure, such as GRASS-GIS (Garzón et al. 2006) and GARP (Stockwell & Noble 1992), argues that if more user-friendly interfaces were available, ML would be more popular. As it is, many ML techniques have to be developed and run via the command line, but the use of ML packages in R (a statistics software package with an interface much like the command line) has been gaining acceptance in ecology (Kangas 2004). Programming skills have become more common in the natural sciences, but user interfaces are still very important for adoption of techniques.

Research scientists want to have a good understanding of the algorithms they use, which makes adoption of a new method a non-trivial investment. Reducing the cost of this investment for ML techniques is an important part of encouraging adoption. One way to do this is through a trusted collaborator who can simultaneously guide the research and transfer skills. Not everyone can find such a collaborator, so a useful tool would be a publicly-available repository of annotated data sets to act as a sandbox for researchers wanting to learn and experiment with these methods. Psychological barriers can be reduced by using alternative names for ML

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

techniques that use familiar terms instead of the computer science names given to algorithms when they were first developed. Random Forest is easier for a beginner to implement, gives easy to interpret results, and has high performance on ecology and earth science classification problems (Prasad et al. 2006; Kampichler et al. 2010); thus, Random Forest would be a good starting point for a ML novice. Students can be exposed to ML and command line programming through their graduate education, eliminating the need for a costly time investment during their research career. In addition, an improved statistical education for students would make them more aware of the limitations imposed by rigid models and thus more open to trying ML for some problems. An important part of promoting new techniques is recognizing the practical needs of researchers and working within those boundaries to facilitate change.

Finally, ML successes and impacts in ecology and earth science need to be more effectively communicated and the results from ML analyses need to be easily interpreted for decision-makers (Guisan et al. 2013). The ML research community needs to do a better job of communicating the impact of their results for specific communities (Wagstaff 2012). For best communication between experts, collaborations should begin during and even before algorithm development to help properly define the problem being addressed, instead of developing an algorithm in isolation (Guisan et al. 2013). Once an algorithm has been successfully used in a decision-making process, the results need to be reported as a part of the published literature in addition to the grey literature.

Funding agencies can facilitate this process by specifically soliciting new collaborative projects (research projects, workshops, hack-a-thons, conference sessions) that apply ML methods to ecology and earth science in innovative ways. Proper implementation of ML methods requires an understanding of the data science and the discipline that can best be achieved through interdisciplinary collaboration.

CONCLUSION

ML methods offer a diverse array of techniques, now accessible to individual researchers, that are well suited to the complex data sets coming from ecology and earth science. These methods have the potential to improve the quality of scientific research by providing more accurate models and accelerate progress in science by widening bottlenecks and filling data gaps. Application of these methods within the ecology and earth science domain needs to increase if society is to see the benefit. Adoption can be promoted through interdisciplinary collaboration, increased communication, and financial support for ML research. A good introductory ML method is Random Forest, which is easy to implement and gives good results. However, ML methods are not the answer to all problems, and in some cases traditional statistical approaches are more appropriate (Olden et al. 2008; Meynard & Quinn 2007); thus, these methods should be used with discretion.

There are many more types of ML methods and subtly different techniques than what has been discussed in this paper. Implementing these ML effectively requires additional background knowledge. A very helpful series of lectures by Stanford Professors Trevor Hastie and Rob Tibshirani called “An Introduction to Statistical Learning with Applications in R” can be accessed online for free and gives a general introduction to traditional statistics and some ML methods. A suggested introductory text is "Machine Learning Methods in the Environmental

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Sciences", by William Hsieh (Hsieh 2009), written at the graduate student level. A useful paper and book written for ecologists is "Machine learning methods without tears: A primer for ecologists" by Olden (Olden et al. 2008) and “Machine Learning Methods for Ecological Applications” by Fielding (Fielding 1999b). ML can be mastered by natural scientists and the time invested in learning it can have significant reward.

ACKNOWLEDGEMENTS
The author would like to acknowledge NASA for financial support and the Boston Machine Learning Meetup Group for inspiration. This paper was greatly improved by comments from Christopher W. Lloyd, Holly A. Bowers, and Alan H. Fielding.

REFERENCES

Acevedo, M. et al., 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4), pp.206–214.

Armitage, D. & Ober, H., 2010. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informaticsq*, 5(6), pp.465–473.

Atkinson, P. & Tatnall, A., 1997. Neural networks in remote sensing - introduction. *International Journal of Remote Sensing*, 18(4), pp.699–709.

Azoulay, P., Fons-Rosen, C. & Graff Zivin, J., 2015. *Does science advance one funeral at a time?*, Available at: http://www.nber.org/papers/w21788.

Balfoort, H. et al., 1992. Automatic identification of algae: Neural network analysis of flow cytometric data. *Journal of Plankton Research*, 14(4), pp.575–589.

Baran, P. et al., 1996. Stochastic models that predict trouts population densities or biomass on macrohabitat scale. *Hydrobiologia*, 337, pp.1–9.

Bell, J., 1999. Tree-based methods. In A. Fielding, ed. *Machine Learning Methods for Ecological Applications*. New York: Springer US, pp. 89–106.

Bhattacharya, M., 2013. Machine learning for bioclimatic modelling. *International Journal of Advanced Computer Science and Applications*, 4(2), p.8.

Bishop, C., 2006. *Pattern Recognition and Machine Learning*, New York: Springer-Verlag.

Bland, L. et al., 2014. Predicting the Conservation Status of Data Deficient Species. *Conservation Biology*, 29(1), pp.250–259.

Boddy, L. et al., 1994. Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry*, 15(4), pp.283–293.

Boddy, L. & Morris, C., 1999. Artificial neural networks for pattern recognition. In A. Fielding, ed. *Machine Learning Methods for Ecological Applications*. New York: Springer US, pp. 37–88.

Bouchon-Meunier, B. et al., 2007. Real-world fuzzy logic applications in data mining and information retrieval. In P. Wang, D. Ruan, & E. Kerre, eds. *Fuzzy Logic - A Spectrum of Theoretical & Practical Issues*. Berlin: Springer, pp. 219–247.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24(2), pp.123–140.

Breiman, L. et al., 1984. *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Breiman, L., 2001a. Random forests. *Machine Learning*, 45(1), pp.5–32.

Breiman, L., 2001b. Statistical modeling: The two cultures. *Statistical Science*, 16(3), pp.199–231.

Brey, T., Jarre-Teichmann, A. & Borlich, O., 1996. Artificial neural network versus multiple linear regression: Predicting P/B ratios from empirical data. *Marine Ecology Progress*

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Series, 140(1-3), pp.251–256.
Brosse, S., Lek, S. & Townsend, C., 2001. Abundance, diversity, and structure of freshwater invertebrates and fish communities: An artificial neural network approach. *New Zealand Journal of Marine and Freshwater Research*, 35(1), pp.135–145.
Brzeziecki, B., Kienast, F. & Wildi, O., 1993. A simulated map of the potential natural forest vegetation of Switzerland. *Journal of Vegetation Science*, 4(4), pp.499–508. Available at: http://www.jstor.org/stable/3236077.
Buddemeier, R. et al., 2008. Coastal typology: an integrative neutral technique - Google Search. *Estuarine and Coastal Shelf Science*, pp.197–205.
Casaioli, M. et al., 2003. Linear and nonlinear postprocessing of numerical forecasted surface temperature. *Nonlinear Processes in Geophysics*, 10(4-5), pp.373–383.
Cavazos, T., Comrie, A. & Liverman, D., 2002. Intraseasonal variability associated with wet monsoons in southeast Arizona. *Journal of Climate*, 15(17), pp.2477–2490.
Chen, D. et al., 2000. A fuzzy logic model with genetic algorithm for analyzing fish stock-recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 57(9), pp.1878–1887.
Chesmore, D., 2004. Automated bioacoustic identification of species. *Annals of the Brazilian Academy of Sciences*, 76(2), pp.435–440.
Chon, T. et al., 1996. Patternizing commuinities by using an artificial neural network. *Ecological Modelling*, 90(1), pp.69–78.
Cornuet, J. et al., 1996. Classifying individuals among infraspecific taxa using microsatellite data and neural networks. *Comptes Rendus de l'Académie des sciences, Série III, Sciences de la vie*, 319(12), pp.1167–1177.
Culverhouse, P. et al., 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247, pp.17–25.
Cutler, D. et al., 2007. Random forest for classification in ecology. *Ecology*, 88(11), pp.2783–2792.
D'Angelo, D. et al., 1995. Ecological uses for genetic algorithms: Predicting fish distributions in complex physical habitats. *Canadian Journal of Fisheries and Aquatic Sciences*, 52(9), pp.1893–1908.
De'ath, G., 2007. Boosted trees for ecological modeling and prediction. *Ecology*, 81(1), pp.243–251.
De'ath, G. & Fabricius, K., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), pp.3178–3192.
Debeljak, M. et al., 2001. Habitat suitability modelling for red deer (Cervus elaphus L.) in South-central Slovenia with classification trees. *Ecological Modelling*, 138(1-3), pp.321–330.
Dedecker, A. et al., 2004. Optimization of artificial neural network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). *Ecological Modelling*, 174(1-2), pp.161–173.
Do, M., Harp, J. & Norris, K., 1999. A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research*, 89, pp.217–224.
Drake, J., Randin, C. & Guisan, A., 2006. Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, 43(3), pp.424–432.
Durbha, S., King, R. & Younan, N., 2007. Support vector machines regression for retrieval of leaf area index from multi-angle imaging spectroradiometer. *Remote Sensing of Environment*, 107(1-2), pp.348–361.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Duro, D., Franklin, S. & Dubé, M., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, 118, pp.259–272.

Džeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. *Ecological Modelling*, 146(1-3), pp.263–273.

Džeroski, S. et al., 1999. Experiments in predicting biodegradability. In *Proceedings of the Ninth International Conference on Inductive Logic Programming*. Berlin: Springer, pp. 80–91.

Džeroski, S., 2009. Machine learning applications in habitat suitability modeling. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 397–412.

Edwards Jr., T. et al., 2005. Model-based stratification for enhancing the detection of rare ecological events. *Ecology*, 86(5), pp.1081–1090.

Edwards, M., Morse, D. & Fielding, A., 1987. Expert systems: Frames, rules or logic for species identification? *Computer Applications in the Biosciences*, 3(1), pp.1–7.

Elith, J. et al., 2006. Novel methods improve prediction of species occurrence data. *Ecography*, 29(2), pp.129–151.

Elith, J., Leathwick, J. & Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), pp.802–813.

Fagerlund, S., 2007. Bird species recognition using support vector machines. *EURASIP Journal of Advances in Signal Processing*, p.038637.

Fielding, A., 1999a. An introduction to machine learning methods. In A. Fielding, ed. *Machine Learning Methods for Ecological Applications*. New York: Springer US, pp. 1–36.

Fielding, A., 2007. *Cluster and Classification Techniques in the BioSciences*, Cambridge, UK: Cambridge University Press.

Fielding, A., 1999b. *Machine Learning Methods for Ecological Applications*, New York: Springer US.

Fischer, H., 1990. Simulating the distribution of plant communities in an alpine landscape. *Coenoses*, 5, pp.37–43.

Fitzgerald, R. & Lees, B., 1992. The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain. In American Society of Photogrametry and Remote Sensing, ed. *Proceedings of the XVII Congress ASPRS*. Bethesda, pp. 570–573.

Forget, G. et al., 2015. ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development Discussions*, 8, pp.3653–3743.

Furlanello, C. et al., 2003. GIS and the random forest predictor: Integration in R for tick-borne disease risk assessment. In K. Hornik, F. Leisch, & A. Zeileis, eds. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Garzón, M. et al., 2006. Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological Modelling*, 197(3-4), pp.383–393.

Geman, S., Bienenstock, E. & Doursat, R., 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), pp.1–58.

Giske, J., Huse, G. & Fiksen, O., 1998. Modelling spatial dynamics of fish. *Reviews in Fish Biology and Fisheries*, 8(1), pp.57–91.

Gislason, P., Benediktsson, J. & Sveinsson, J., 2006. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4), pp.294–300.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Goldberg, D. & Holland, J., 1988. Genetic algorithms and machine learning. *Machine Learning*, 3(2), pp.95–99.
Goodwin, J., North, E. & Thompson, C., 2014. Evaluating and improving a semi-automated image analysis technique for identifying bivalve larvae. *Limnology and Oceanography Methods*, 12(8), pp.548–562.
Guégan, J., Lek, S. & Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. *Nature*, 391, pp.382–384.
Guisan, A. et al., 2013. Predicting species distribution for conservation decisions. *Ecology Letters*, 16(12), pp.1424–1435.
Guisan, A. & Thuiller, W., 2005. Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8(9), pp.993–1009.
Guisan, A. & Zimmermann, N., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3), pp.147–186.
Guo, Q., Kelly, M. & Graham, C., 2005. Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling*, 182(1), pp.75–90.
Haefner, J., 2005. *Modeling Biological Systems: Principles and Applications* 2nd ed., New York: Springer US.
Ham, J. et al., 2005. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience Remote Sensing*, 43(3), pp.492–501.
Haupt, R. & Haupt, S., 2004. *Practical Genetic Algorithms* 2nd ed., John Wiley & Sons, Inc.
Haupt, S., 2009. Environmental optimization: Applications of genetic algorithms. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 379–396.
Haupt, S., Allen, C. & Young, G., 2009. Addressing air quality problems with genetic algorithms: A detailed analysis of source characterization. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 269–296.
Haupt, S., Pasini, A. & Marzban, C., 2009. *Artificial Intelligence Methods in the Environmental Sciences*, Amsterdam: Springer Netherlands.
Henderson, B. et al., 2005. Australia-wide predictions of soil properties using decision trees. *Geoderma*, 124(3-4), pp.383–398.
Holland, J., 1975. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, Cambridge, MA, USA: MIT Press.
Hsieh, W., 2009. *Machine Learning Methods in the Environmental Sciences*, Cambridge: Cambridge University Press.
Hsieh, Y. & Hsieh, W., 2003. An adaptive nonlinear MOS scheme for precipitation forecasts using neural networks. *Weather and Forecasting*, 18(2), pp.303–310.
Huang, X. & Jensen, J., 1997. A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. *Photogrammetric Engineering & Remote Sensing*, 63(10), pp.1185–1194.
Iverson, L., Prasad, A. & Liaw, A., 2004. New machine learning tools for predictive vegetation mapping after climate change: Bagging and random forest perform better than regression tree analysis. In R. Smithers, ed. *Landscape Ecology of Trees and Forests, Proceedings of the Twelfth Annual IALE(UK) Conference*. Cirencester, UK: International Association for Landscape Ecology, pp. 317–320.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Jeffers, J., 1999. Genetic Algorithms I. In *Machine Learning Methods for Ecological Applications*. Amsterdam: Springer Netherlands, pp. 107–122.

Jennings, N., Parsons, S. & Pocock, M., 2008. Human vs. machine: Identification of bat species from their echolocation calls by humans and by artificial neural networks. *Canadian Journal of Zoology*, 86(5), pp.371–377.

Jones, M., Fielding, A. & Sullivan, M., 2006. Analysing extinction risk in parrots using decision trees. *Biodiversity and Conservation*, 15(6), pp.1993–2007.

Kampichler, C. et al., 2010. Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics*, 5(6), pp.441–450.

Kampichler, C., Džeroski, S. & Wieland, R., 2000. Application of machine learning techniques to the analysis of soil ecological data bases: Relationships between habitat features and Collembolan community characteristics. *Soil Biology & Biochemistry*, 32(2), pp.197–209.

Kangas, M., 2004. R: A computational and graphics resource for ecologists. *Frontiers in Ecology and the Environment*, 2(5), pp.277–277.

Kasten, E., McKinley, P. & Gage, S., 2010. Ensemble extraction for classification and detection of bird species. *Ecological Informatics*, 5(3), pp.153–166.

Kastens, T. & Featherstone, A., 1996. Feedforward backpropagation neural networks in prediction of farmer risk preference. *American Journal of Agricultural Economics*, 78(2), pp.400–415. Available at: http://www.jstor.org/stable/1243712.

Knudby, A., Brenning, A. & LeDrew, E., 2010. New approaches to modelling fish-habitat relationships. *Ecological Modelling*, 221(3), pp.503–511.

Knudby, A., LeDrew, E. & Brenning, A., 2010. Predictive mapping of reef fishes specie richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment*, 114(6), pp.1230–1241.

Kobler, A. & Adamic, M., 2000. Identifying brown bear habitat by a combined GIS and machine learning method. *Ecological Modelling*, 135(2-3), pp.291–300.

Kohonen, T., 1989. *Self-Organization and Associative Memory*, Berlin: Springer-Verlag.

Koza, J., 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, MA, USA: MIT Press.

Krasnopolsky, V., 2009. Neural network applications to solve forward and inverse problems in atmospheric and oceanic satellite remote sensing. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 191–206.

Kreßel, U., 1999. Pairwise classification and support vector machines. In B. Schölkopf, C. Burges, & A. Smola, eds. *Advances in Kernel Methods*. Cambridge, MA, USA: MIT Press, pp. 255–268.

Lakshmanan, V., 2009. Automated analysis of spatial grids. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 329–346.

Laplace, P., 1986. Memoir on the probability of the causes of events. *Statistical Science*, 1(3), pp.364–378.

Lawler, J. et al., 2006. Predicting climate-induced range shifts: Model differences and model reliability. *Global Change Biology*, 12(8), pp.1568–1584.

Lee, J. et al., 20007. Classification of breeding bird communities along an urbanization gradient using an unsupervised artificial neural network. *Ecological Modelling*, 203(1-2), pp.62–71.

Lee, Y., Lin, Y. & Wahba, G., 2004. Multicategory support vector machines: Theory and

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99, pp.67–81.

Lees, B., 1996. Sampling strategies for machine learning using GIS. In M. Goodchild et al., eds. *GIS and Environmental Modeling: Progress and Issues*. Fort Collins: GIS World Inc., pp. 39–42.

Lees, B. & Ritman, K., 1991. Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environment. *Environmental Management*, 15(6), pp.823–831.

Lek, S., Delacoste, M., et al., 1996. Application of neural networks to modelling non linear relationships in ecology. *Ecological Modelling*, 90(1), pp.39–52.

Lek, S., Belaud, A., et al., 1996. Role of some environmental variables in trout abundance models using neural networks. *Aquatic Living Resources*, 9(1), pp.23–29.

Lek, S. & Guégan, J., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120(2-3), pp.65–73.

Lek, S. & Guégan, J., 2000. *Artificial Neuronal Networks: Application to Ecology and Evolution*, Berlin: Springer-Verlag.

Levine, E., Kimes, D. & Sigillito, V., 1996. Classifying soil-structure using neural networks. *Ecological Modelling*, 92, pp.101–108.

Li, J. et al., 2011. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26(12), pp.1647–1659.

Lim, T., Loh, W. & Shih, Y., 2000. A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms. *Machine Learning*, 40(3), pp.203–229.

Lorenz, E., 1963. Deterministic non-periodic flow. *Journal of Atmospheric Sciences*, 20, pp.130–141.

MacLeod, N., 2007. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, CRC Press.

Maier, H. & Dandy, G., 2000. Neural networks for the prediction and forecasting of water resource variables: A review of modelling issues and applications. *Environmental Modelling & Software*, 15(1), pp.101–124.

Manel, S. et al., 2001. Alternative methods for predicting species distribution: An illustration with Himalayan river birds. *Journal of Applied Ecology*, 36(5), pp.1999–2010.

Marzban, C., 2003. A neural network for post-processing model output: ARPS. *Monthly Weather Review*, 131, pp.1103–1111.

Mastrorillo, S. et al., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology*, 38(2), pp.237–246.

McKay, R., 2001. Variants of genetic programming for species distribution modelling - fitness sharing, partial functions, population evaluation. *Ecological Modelling*, 146(1-3), pp.231–241.

McKenna Jr., J., 2005. Application of neural networks to prediction of fish diversity and salmonid production in the Lake Ontario basin. *Transactions of the American Fisheries Society*, 134(1), pp.28–43.

Meynard, C. & Quinn, J., 2007. Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, 34(8), pp.1455–1469.

Miller, J. & Franklin, J., 2002. Modelling distribution of four vegetation alliances using

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

generalized linear models and classification trees with spatial dependence. *Ecological Modelling*, 157(2-3), pp.227–247.
Mitchell, M., 1998. *An Introduction to Genetic Algorithms*, Cambridge, MA, USA: MIT Press.
Moguerza, J. & Muñoz, A., 2006. Support vector machines with applications. *Statistical Science*, 21(3), pp.322–336.
Mulligan, A. & Brown, L., 1998. Genetic algorithms for calibrating water quality models. *Journal of Environmental Engineering*, 124(3), pp.202–211.
Muttil, N. & Lee, J., 2005. Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecological Modelling*, 189(3-4), pp.363–376.
Olden, J., Lawler, J. & LeRoy Poff, N., 2008. Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology*, 83(2), pp.171–193. Available at: http://www.jstor.org/stable/10.1086/587826.
Özesmi, S., Tan, C. & Özesmi, U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecological Modelling*, 195(1-2), pp.83–93.
Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), pp.217–222.
Park, Y. & Chon, T., 2007. Biologically-inspired machine learning implemented to ecological informatics. *Ecological Modelling*, 203(1), pp.1–7.
Parsons, S. & Jones, G., 2000. Acoustic identification of twelve species of echoloating bat by discriminant function analysis and artificial neural networks. *Journal of Experimental Biology*, 203, pp.2641–2656.
Pasini, A., 2009. Neural network modeling in climate change studies. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 235–254.
Peters, J. et al., 2007. Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, 207(2-4), pp.304–318.
Peterson, A. et al., 2002. Future projections for Mexican faunas under global climate scenarios. *Nature*, 416, pp.626–629.
Pineda, F., 1987. Generalization of backpropagation to recurrent neural networks. *Physical Review Letters*, 19(59), pp.2229–2232.
Prasad, A., Iverson, L. & Liaw, A., 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9(2), pp.181–199.
Quintero, E. et al., 2014. A statistical assessment of population trends for data deficient Mexican amphibians. *PeerJ*, 2, p.e703.
Rasmussen, C. & Williams, C., 2006. *Gaussian Processes for Machine Learning*, Cambridge, MA, USA: MIT Press.
Reby, D. et al., 1998. Individuality in the groans of fallow deer (Dama dama) bucks. *Journal of the Zoological Society of London*, 245(1), pp.78–84.
Recknagel, F., 1997. ANNA - artificial neural network model for predicting species abundance and succession of blue-green algae. *Hydrobiologia*, 349, pp.47–57.
Recknagel, F., 2001. Applications of machine learning to ecological modelling. *Ecological Modelling*, 146(1-3), pp.303–310.
Recknagel, F. et al., 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. *Journal of Hydroinformatics*, 4(2), pp.125–133.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Recknagel, F. et al., 2000. Multivariate time-series modelling of algal blooms in freshwater lakes by machine learning. In *Proceedings of the 5th International Symposium WATERMATEX'2000 on Systems Analysis and Computing in Water Quality Management*. Ghent, Belgium.

Ribic, C. & Ainley, D., 1997. The relationships of seabird assemblages to physical habitat features in Pacific equatorial waters during spring 1984-1991. *ICES Journal of Marine Science*, 54(4), pp.593–599.

Rogan, J. et al., 2008. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, 112(5), pp.2272–2283.

Rosa, I. et al., 2016. Classification success of six machine learning algorithms in radar ornithology. *Ibis*, 158(1), pp.28–42.

Ruck, B., Walley, W. & Hawkes, H., 1993. Biological classification of river water quality using neural networks. In G. Rzevski, J. Pastor, & R. Adey, eds. *Applications of Artificial Intelligence in Engineering VIII*. Essex, UK: Elsevier Applied Science, pp. 361–372.

Rumelhart, D., Hinton, G. & Williams, R., 1986. Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, & P. Group, eds. *Parallel distributed processing*. Cambridge, MA, USA: MIT Press, pp. 318–362.

Sarkar, S. et al., 2006. Biodiversity conservation planning tools: Present status and challenges for the future. *Annual Review of Environment and Resources*, 31, pp.123–159.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. *Marine Ecology Progress Series*, 139, pp.289–299.

Scardi, M. & Harding Jr., L., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling*, 120(2-3), pp.213–223.

Schultz, A., Wieland, R. & Lutze, G., 2000. Neural networks in agroecological modelling - stylish application or helpful tool? *Computers and Electronics in Agriculture*, 29(1-2), pp.73–97.

Seginer, I., Boulard, T. & Bailey, B., 1994. Neural network models of the greenhouse climate. *Journal of Agricultural Engineering Research*, 59(3), pp.203–216.

Segurado, P. & Araújo, M., 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography*, 31(10), pp.1555–1568.

Semtner, A., 1995. Modeling Ocean Circulation. *Science*, 269, pp.1379–1380.

Simmonds, E., Armstrong, F. & Copland, P., 1996. Species identification using wideband backscatter with neural network and discriminant analysis. *ICES Journal of Marine Science*, 53, pp.189–195.

Sosik, H. & Olson, R., 2007. Automated taxonomic classification of phytoplankton sampled with imaging in flow cytometry. *Limnology and Oceanography: Methods*.

Spitz, F. & Lek, S., 1999. Environmental impact prediction using neural network modelling: An example in wildlife damage. *Journal of Applied Ecology*, 36(2), pp.317–326.

Srinivasan, A. et al., 1997. Carcinogenesis prediction using inductive logic programming. In N. Lavrač, E. Keravnou, & B. Zupan, eds. *Intelligent Data Analysis in Medicine and Pharmacology*. New York: Springer US, pp. 243–260.

Stockwell, D., 1999. Genetic Algorithms II. In A. Fielding, ed. *Machine Learning Methods for Ecological Applications*. New York: Springer US, pp. 123–144.

Stockwell, D. & Noble, I., 1992. Induction of sets of rules from animal distribution data: A robust and informative method of analysis. *Mathematics and Computers in Simulation*, 33(5-6), pp.385–390.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Stockwell, D. & Peters, D., 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, 13(2), pp.143–158.

Sutherland, W. et al., 2004. The need for evidence-based conservation. *Trends in Ecology and Evolution*, 19(6), pp.305–308.

Tan, S. & Smeins, F., 1996. Predicting grassland community changes with an artificial neural network model. *Ecological Modelling*, 84(1-3), pp.91–97.

Termansen, M., McClean, C. & Preston, D., 2006. The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling*, 192(3-4), pp.410–424.

Thuiller, W., 2003. BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, 9(10), pp.1353–1362.

Tscherko, D., Kandeler, E. & Bárdossy, A., 2007. Fuzzy classification of microbial biomass and enzyme activities in grassland soils. *Soil Biology & Biochemistry*, 39(7), pp.1799–1808.

Vayssières, M., Richard, R. & Allen-Diaz, B., 2000. Classification trees: an alternative non-parametric approach for predicting species distribution. *Journal of Vegetation Science*, 11(5), pp.679–694.

Wagstaff, K., 2012. Machine learning that matters. In *Proceedings of the 29th International Conferene on Machine Learning*. Edinburgh: California Institute of Technology, pp. 298–303.

Walley, W. & Džeroski, S., 1996. Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification. In R. Denzer, G. Schimak, & D. Russell, eds. *Environmental Software Systems: Proceedings of the International Symposium on Environmental Software Systems*. New York: Springer US, pp. 229–240.

Walley, W., Hawkes, H. & Boyd, M., 1992. Application of Bayesian inference to river water quality surveillance. In D. Grierson, G. Rzevski, & R. Adey, eds. *Applications of Artificial Intelligence in Engineering VII*. Essex, UK: Elsevier, pp. 1030–1047.

Walley, W., Martin, R. & O'Connor, M., 2000. Self-organising maps for the classification and diagnosis of river quality from biological and environmental data. In R. Denzer et al., eds. *Environmental Software Systems: Environmental Information and Decision Support*. New York: Spinger US, pp. 27–41.

Webb, T., Berghe, E. & O'Dor, R., 2010. Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. *PLoS One*, 5(8), p.e10223.

Wieland, R., 2008. Fuzzy models. In S. Jørgensen & B. Fath, eds. *Encyclopedia of Ecology*. Amsterdam: Elsevier, pp. 1717–1726.

Wieland, R. & Mirschel, M., 2008. Adaptive fuzzy modeling versus artificial neural networks. *Environmental Modelling & Software*, 23(2), pp.215–224.

Wiley, E. et al., 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, 16(4), pp.120–127.

Williams, J. et al., 2009. Fuzzy logic applications. In S. Haupt, A. Pasini, & C. Marzban, eds. *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 347–378.

Worner, S. & Gevrey, M., 2006. Modelling global insect pest species assemblages to determine risk of invasion. *Journal of Applied Ecology*, 43(5), pp.858–867.

Wu, A., Hsieh, W. & Tang, B., 2006. Neural networks forecasts of the tropical Pacific sea

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

surface temperatures. *Neural Networks*, 19, pp.145–154.
Young, G., 2009. Implementing a neural network emulation of a satellite retrieval algorithm. In *Artificial Intelligence Methods in the Environmental Sciences*. Amsterdam: Springer Netherlands, pp. 207–216.
Vander Zanden, M. et al., 2004. Predicting occurrences and impacts of smallmouth bass introductions in north temperate lakes. *Ecological Applications*, 14(1), pp.132–148.
Zhao, K. et al., 2011. Characterizing forest canopy structure with lidar composite metrics and machine learning. *Remote Sensing of Environment*, 115(8), pp.1978–1996.
Zhao, K., Popescu, S. & Zhang, X., 2008. Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. *Photogrammetric Engineering & Remote Sensing*, 74(10), pp.1223–1234.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

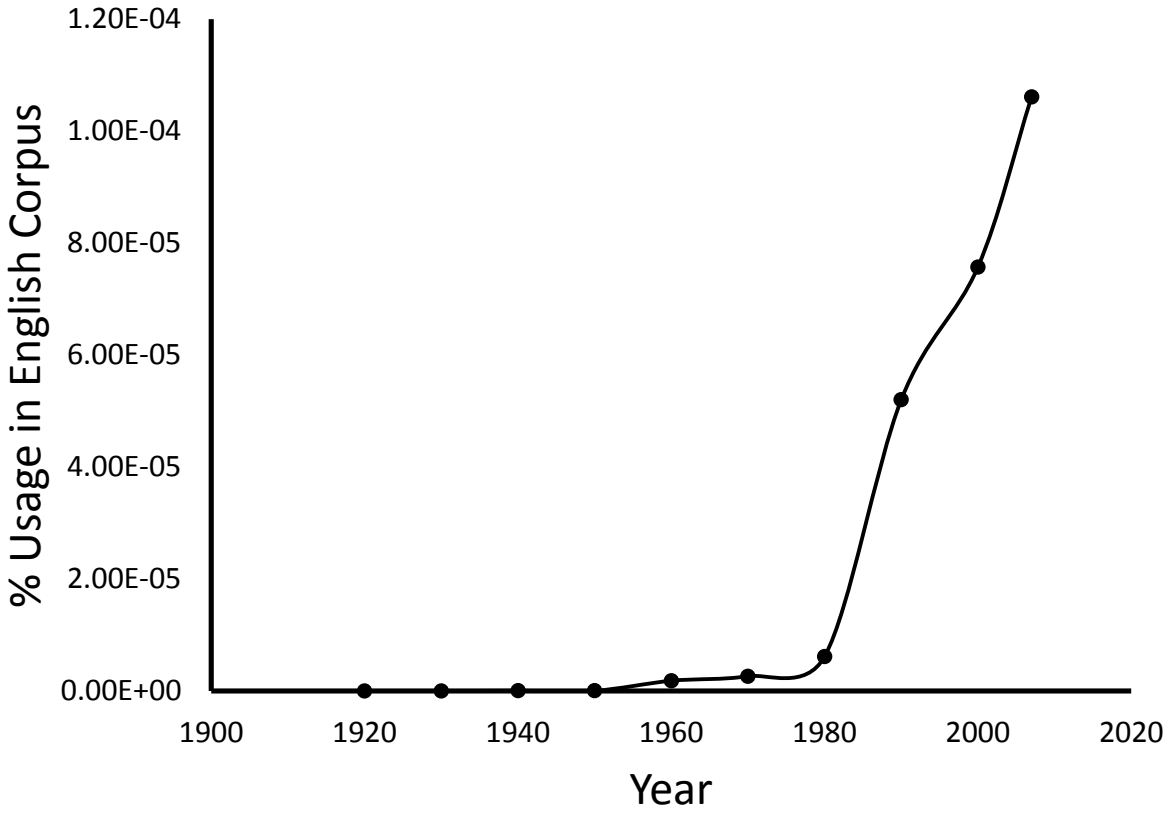

Figure 1: **Use of the phrase 'machine learning' in the Google Books Ngram Viewer**: This plot shows the use of the phrase 'machine learning' by decade as percentage of total words in the Google English Corpus. http://books.google.com/ngrams

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

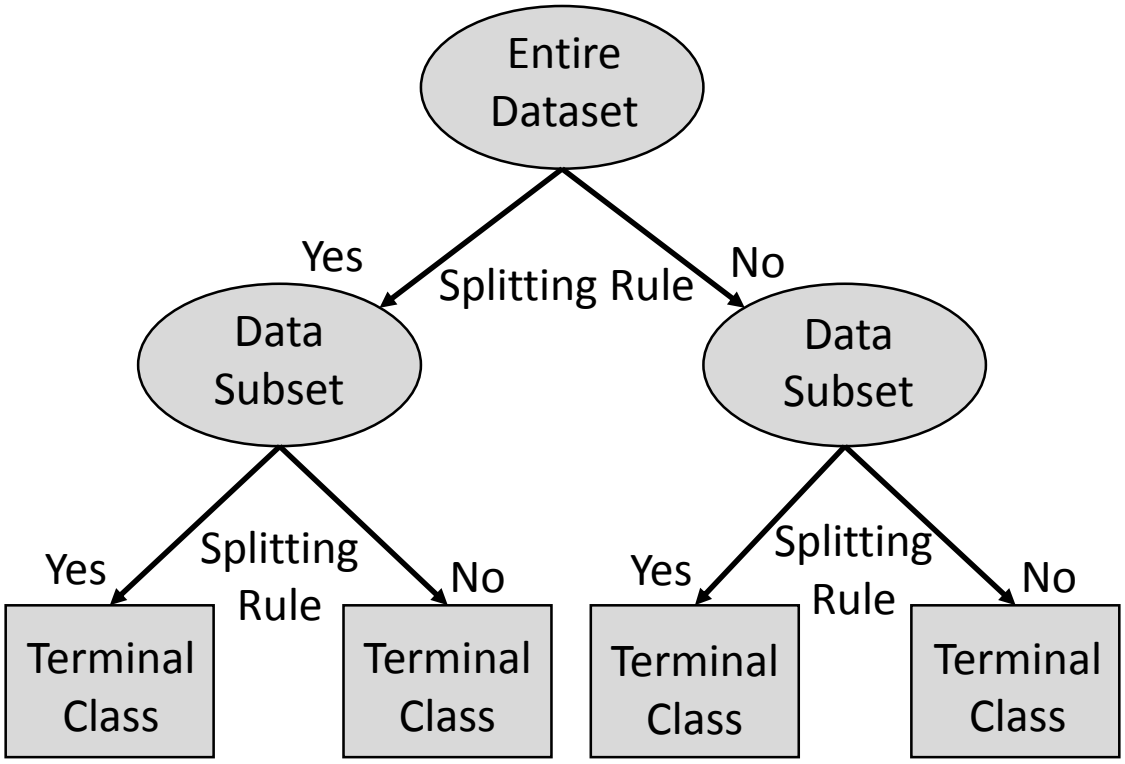

Figure 2: **Decision and Classification Tree Schematic**: Tree-based machine learning methods infer rules for splitting a data set into more homogeneous data sets until a specified number of terminal classes or maximum variance within the terminal classes is reached. The inferred splitting rules can give additional information about the system being studied.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

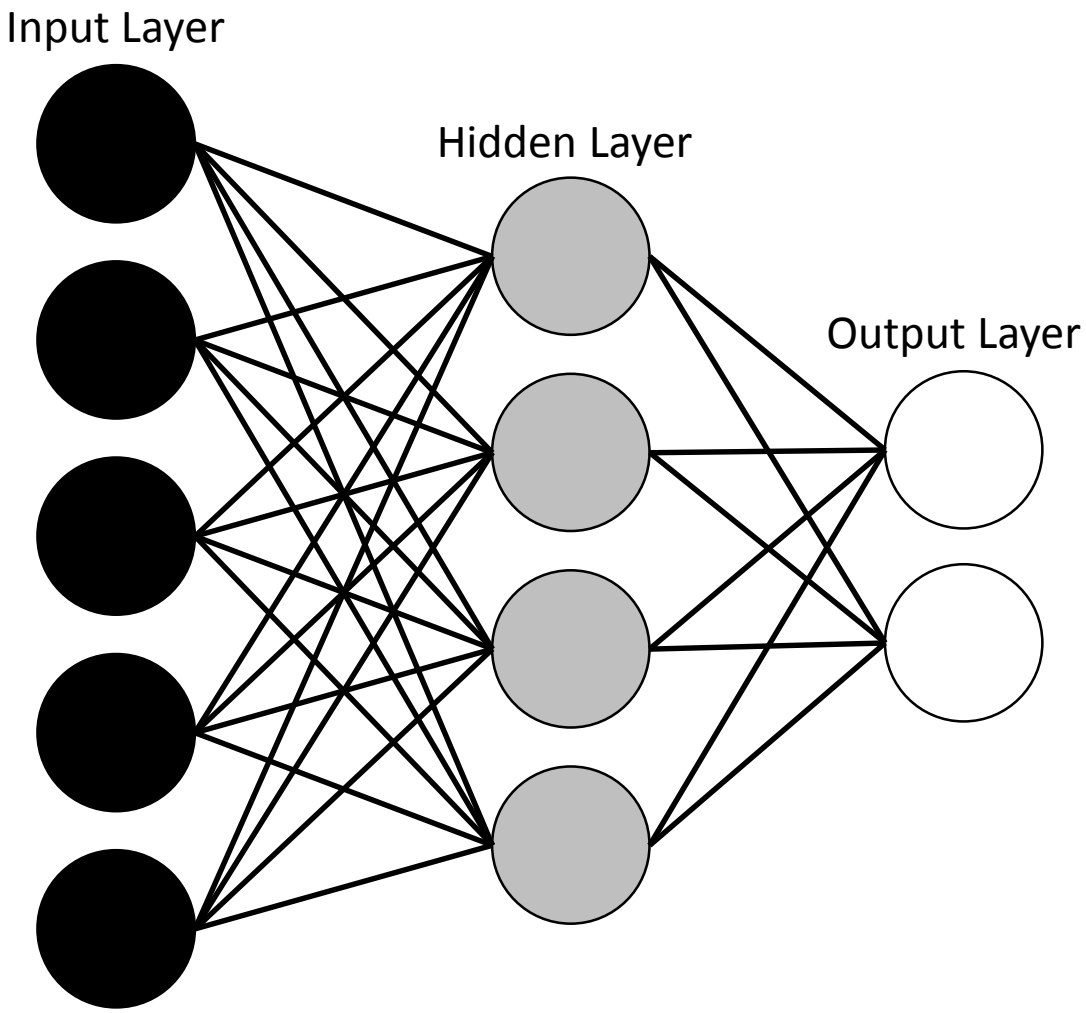

Figure 3: **Artificial Neural Network Schematic**: A neural network is made up of three layers (input, hidden, output). Each layer contains "neurons" with an assigned activity level. Each neuron has a connection to every other neuron in adjacent layers with an assigned connection weight. The neurons in the input layer correspond to the independent variables and the neurons in the output layer correspond to the dependent variables. The number and activity level of hidden neurons and the connection weights are varied to minimize the error in the output layer.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

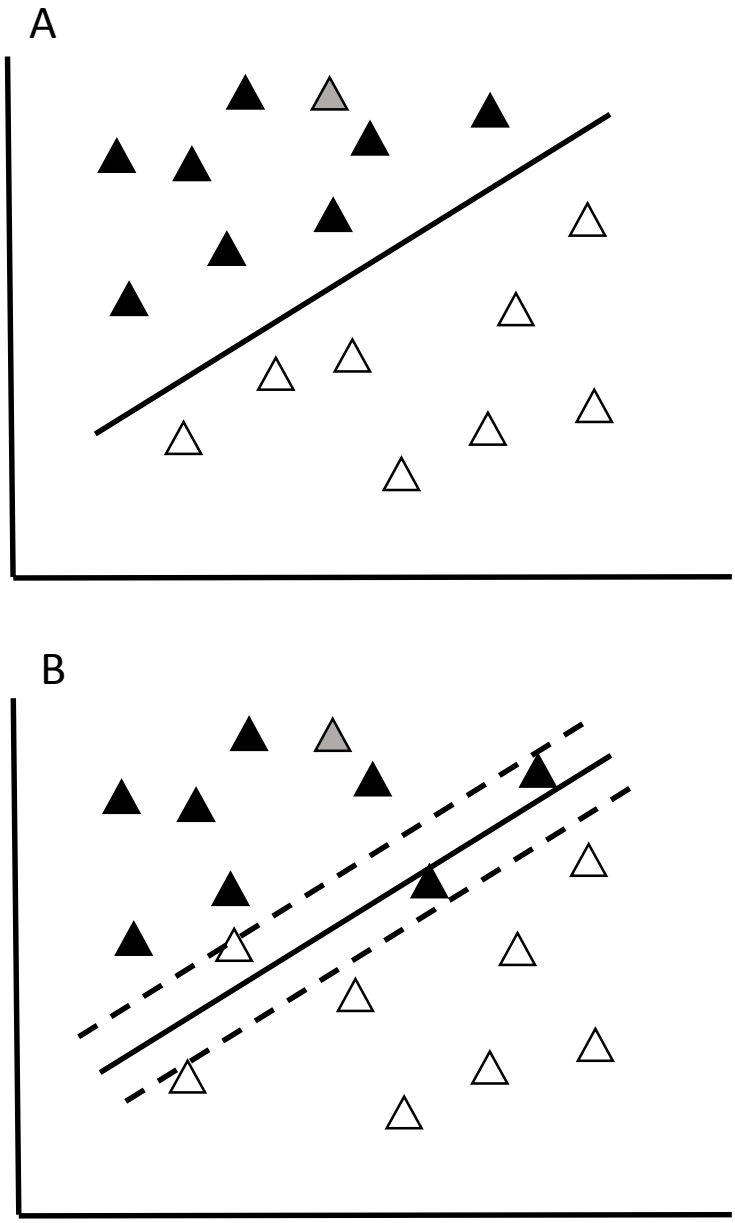

Figure 4: **Support Vector Machine Schematic**: A) This simplified schematic shows the plane inferred by the maximal margin classifier (black line) dividing the data set into two classes. A new datum (grey), will be classified according to its position relative to the hyperplane. B) If the data are noisy, and not easily separated, a "buffer zone" (dotted lines) can be used to separate the two classes.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

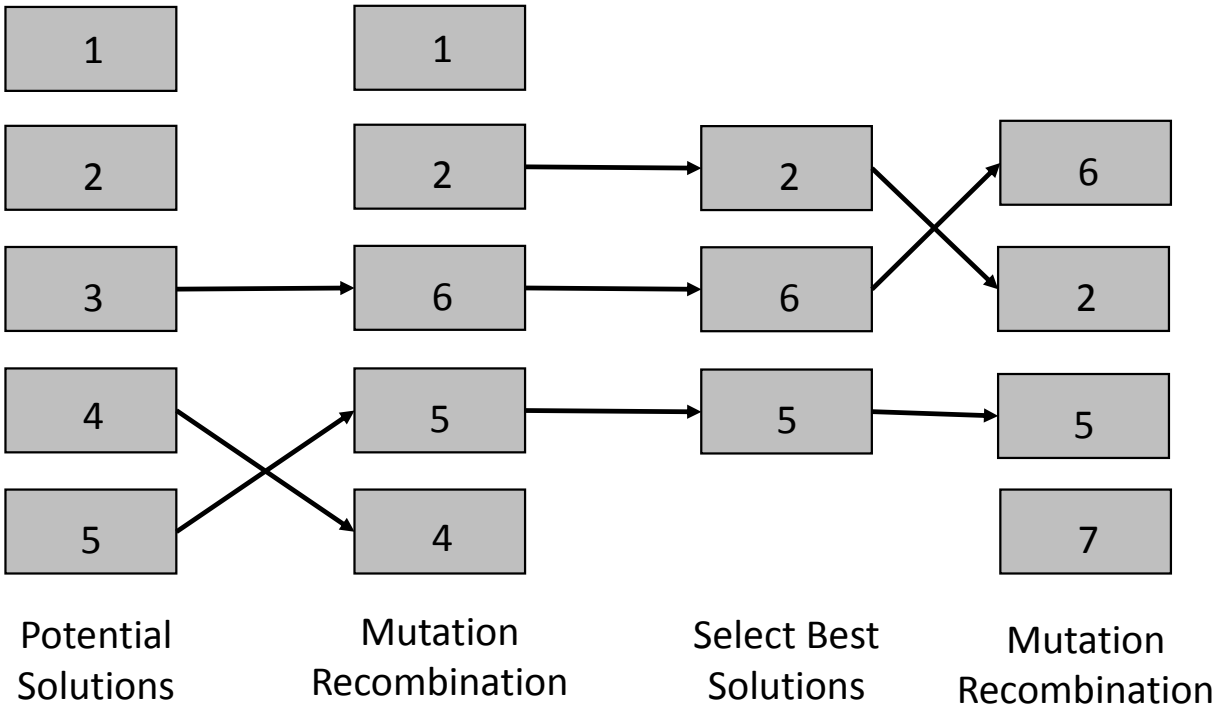

Figure 5: **Genetic Algorithm Schematic**: In this simplified schematic of a genetic algorithm, the five potential solutions, or "chromosomes", undergo mutation and recombination. Then the best performing solutions are selected for another iteration of mutation and recombination. This cycle is repeated until an optimal solution is found.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

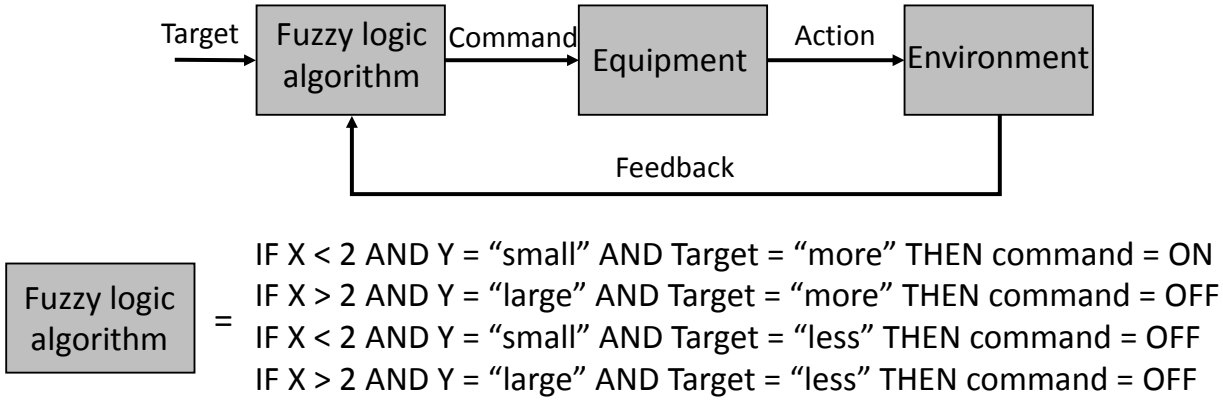

Figure 6: **Logic-Based Algorithm Schematic**. This diagram shows a simplified schematic of a fuzzy logic algorithm that controls a piece of equipment and responds to environmental variables according to user-defined settings. X and Y are environmental variables and the Target is user-defined.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

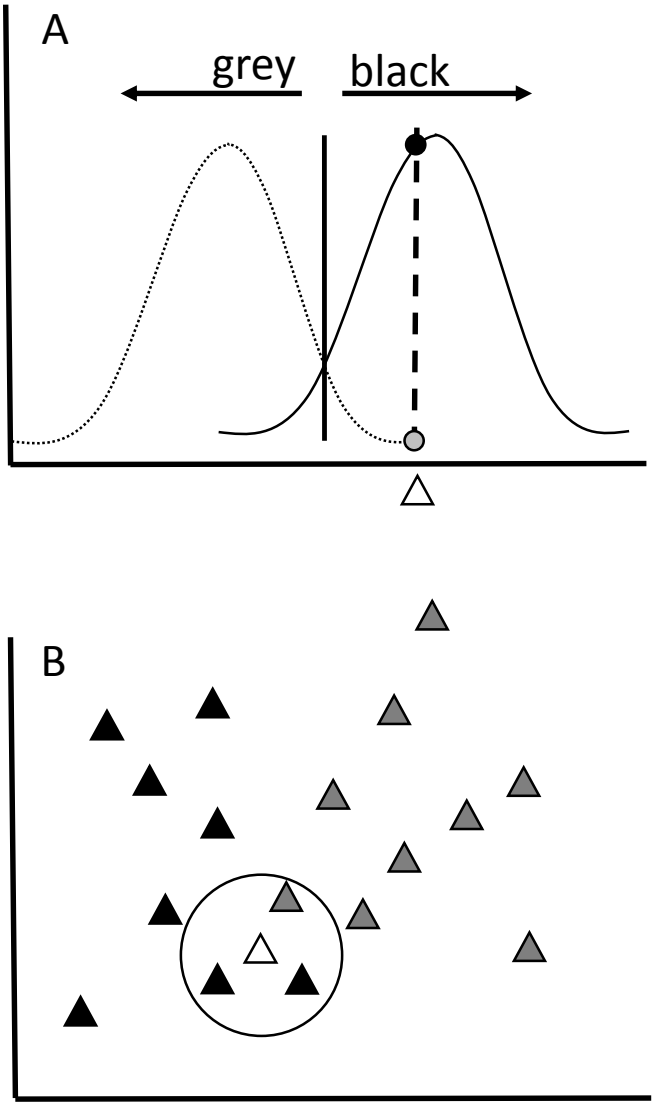

Figure 7: **Bayesian Classifier Schematic**: This diagram shows a simplified schematic of a Bayesian classifier working to assign a new datum (white triangle) to one of two classes (grey and black). A) Probability Density Plot: A Bayesian classifier calculates a probability density for each class (solid and dotted curve) across a range of values for the new datum (white triangle), which is classified according to which probability is highest at its value (black). The value for which the datum has an equal probability of being in both classes is called the decision boundary (black line). B) Data Plot: An object to be classified (white) can belong to one of two groups (grey or black). This method would classify the object within the group with the highest probability of being correct. In this example, the white item would be classified as a member of the black group because the probability is higher (Black = 8/17 * 2/8 and Grey = 9/17 *1/9)

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

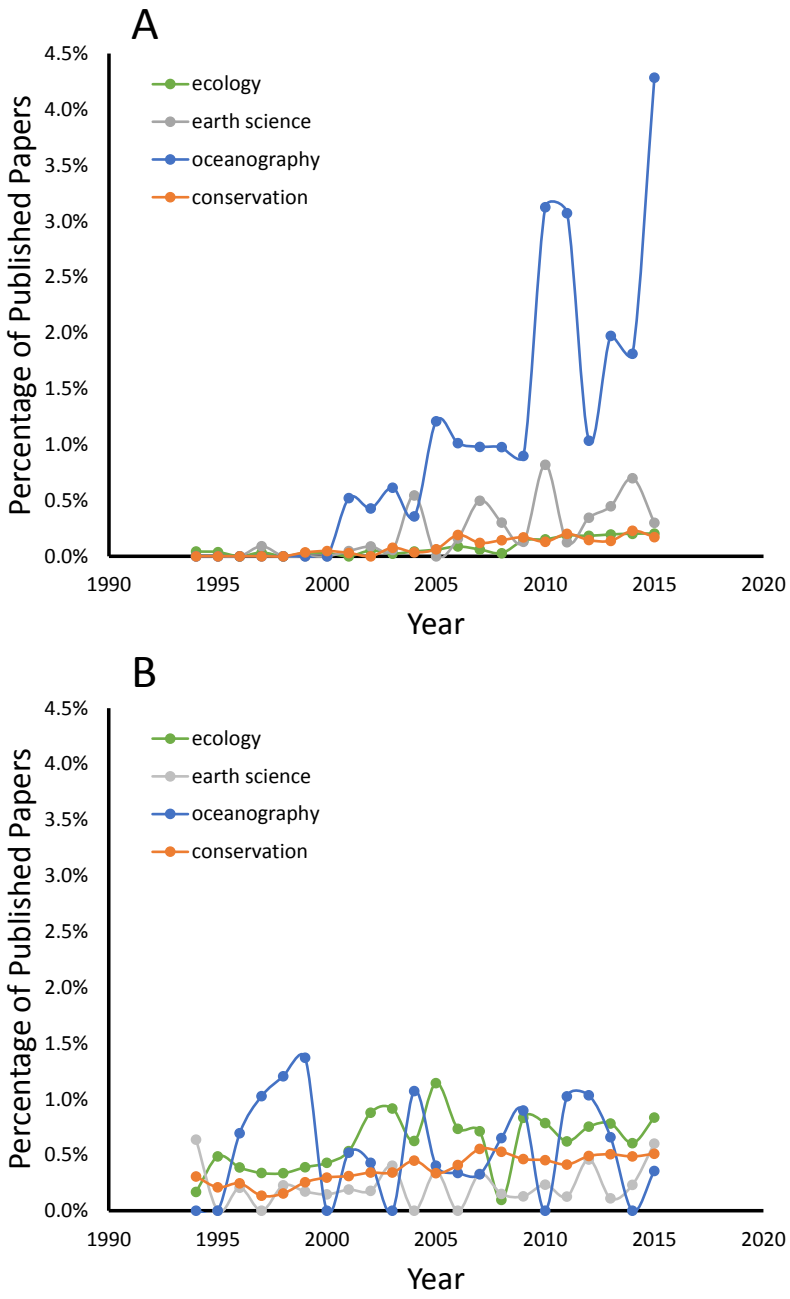

Figure 8: **Proportion of Machine Learning Research Articles in Earth Science and Ecology**: This plot shows the proportion of articles about machine learning (A) and linear regression (B) in four natural science disciplines from 1994 to 2015. Data were collected from Web of Science using the discipline name (ecology, earth science, oceanography, conservation) and "machine learning" + discipline name or "linear regression" + discipline name as search terms.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1720v1 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016