Techniques and Technology Article

Data-Mining Discovery of Pattern and Process in Ecological Systems

WESLEY M. HOCHACHKA,[1] *Laboratory of Ornithology, Cornell University, Ithaca, NY 14850, USA*
RICH CARUANA, *Department of Computer Science, Cornell University, Ithaca, NY 14853, USA*
DANIEL FINK, *Laboratory of Ornithology, Cornell University, Ithaca, NY 14850, USA*
ART MUNSON, *Department of Computer Science, Cornell University, Ithaca, NY 14853, USA*
MIREK RIEDEWALD, *Department of Computer Science, Cornell University, Ithaca, NY 14853, USA*
DARIA SOROKINA, *Department of Computer Science, Cornell University, Ithaca, NY 14853, USA*
STEVE KELLING, *Laboratory of Ornithology, Cornell University, Ithaca, NY 14850, USA*

?1 **ABSTRACT** Most ecologists use statistical methods as their main analytical tools when analyzing data to identify relationships between a response and a set of predictors; thus, they treat all analyses as hypothesis tests or exercises in parameter estimation. However, little or no prior knowledge about a system can lead to creation of a statistical model or models that do not accurately describe major sources of variation in the response variable. We suggest that under such circumstances data mining is more appropriate for analysis. In this paper we: 1) present the distinctions between data-mining (usually exploratory) analyses, and parametric statistical (confirmatory) analyses; 2) illustrate 3 strengths of data-mining tools for generating hypotheses from data; and 3) suggest useful ways in which data mining and statistical analyses can be integrated into a thorough analysis of data to facilitate rapid creation of accurate models, and to guide further research. (JOURNAL OF WILDLIFE MANAGEMENT 71(7):000–000; 2007)

DOI: 10.2193/2006-503

KEY WORDS bagging, data mining, decision trees, exploratory data analysis, hypothesis generation, machine learning, prediction.

Ecologists mainly use statistical methods for analysis of the relationships between an observed response and a set of predictors in a data set, and typically the statistical techniques used are parametric. By parametric we mean statistical techniques that require the user to specify the predictor (i.e., independent) variables to include in analyses, the functional forms of relationships (e.g., categorical, linear, quadratic) between predictions and the response (i.e., dependent) variable, and a basic model of the underlying processes (e.g., unexplained variance is normally distributed). This approach to data analysis is appropriate, both for parameter estimation and hypothesis testing, as long as the analyst has sufficient prior knowledge to specify an appropriate parametric model.

If insufficient knowledge exists to specify a single parametric model with confidence, various statistical methods exist for exploring a wider model space, although not all such methods work well (see e.g., Burnham and Anderson 2002). One method currently popular among ecologists is multi-model inference based on use of Akaike's Information Criterion (AIC; Burnham and Anderson 2002), in which the support for a small number of models can be explored. Reversible-jump Markov-chain Monte Carlo analyses (e.g., King and Brooks 2004) are another technique for exploring a wider model set. Still another approach is to use generalized additive models (e.g., Hastie and Tibshirani 1995, Wood 2006), which are appropriate when predictor variables can be specified, but the functional forms of relationships between continuous predictors and response are unknown or extremely complex. However, even these types of flexibility may be insufficient to allow ecologists to extract biological insights from their data, if prior knowledge is minimal and hypotheses are not clearly developed. Under these circumstances, exploratory analyses (analyses useful for generating hypotheses) are more appropriate than the confirmatory analyses (analyses designed to test hypotheses or estimate model parameters) typically presented in ecological publications.

Data-mining techniques are often more powerful, flexible, and efficient for exploratory analysis than are statistical techniques. Data mining (e.g., Breiman 1996, Bauer and Kohavi 1999, Hastie et al. 2001) can be characterized as an analysis of data 1) which automatically makes accurate predictions from data, 2) with the ability to screen a large number of predictor variables and identify the most important predictors, 3) without requiring the user to make many assumptions about the forms of relationships between predictor variables and the response variable. That is, data-mining analyses are nonparametric analyses in the sense that the analysis identifies many components (important predictors, functional forms of relationships) that must be specified as parameters in parametric statistical analyses. While some basic features of data mining have been introduced to and used by ecologists (e.g., De'ath and Fabricius 2000, Elith et al. 2006), most ecological researchers have not availed themselves of all the strengths of data-mining techniques, or generally to the most recently developed and better-performing (e.g., Caruana and Niculescu-Mizil 2006) methods. ?2

We will highlight the strengths of data-mining analysis for

[1] *E-mail: wmh6@cornell.edu*

scientific discovery, and discuss the roles that data mining should play in a thorough analysis of ecological data. Our goals are to: 1) present the notion that exploratory and confirmatory analyses are both important parts of ecological research, and that differences in the objectives of exploratory and confirmatory analyses have led to the development of different sets of analytical tools, 2) highlight briefly 3 strengths of current data-mining tools for scientific discovery through exploration of data, and 3) discuss the roles that data mining can play at different stages in the progression of data analysis from initial exploration to confirmatory conclusion. We argue that data mining, in combination with statistical analyses, should be more commonly used to analyze ecological data in order for ecologists to extract as much insight from their data as possible.

PHILOSOPHY, GOALS, AND METHODS OF DATA MINING

Statistical techniques emphasize confirmatory analyses. Parametric statistical methods are developed around a theoretically based goal (e.g., maximizing likelihood) and assumption (e.g., normally distributed errors), with the process used to meet the assumption-constrained goal being a subsidiary issue. Thus, analysis presupposes sufficient knowledge to state the hypotheses and know the validity of assumptions. Because of the emphasis on accurate parameter estimation and interpretation of models, statistical data analysis often eschews the inclusion of correlated predictors or predictors with weak influences on the response variable—sometimes at the expense of the overall predictive performance of the model. Concerns about weak or confounding predictors are behind the emphasis on predictor selection and testing. All of the constraints resulting from a need to prespecify a model have the useful consequence that the end result of a parametric statistical analysis is a model that can be easily interpreted.

In contrast, data-mining methods have focused on detecting and describing patterns within data, possibly in the absence of any preconceived ideas of what these patterns may be. The major focus in data-mining research has been to develop methods for predictions of binary response. Data mining concentrates on predictive accuracy, and the modeling philosophy behind data mining stresses the inclusion of any predictor that is potentially informative. As a result, data-mining methods have been developed with the aim of analyzing large data sets, both large numbers of predictor variables as well as large numbers of data records. By focusing on predictive accuracy, as opposed to the goal in parametric statistical analysis of interpreting the effects of predictors on the response, many data-mining methods essentially produce black box models (models of which the user typically does not need to see or understand the structure). Additional work is required to open the black box and visualize the relationships between predictors and response. Note that AIC-based statistical model selection partially bridges this dichotomy, as it asymptotically (i.e., with large sample sizes; Burnham and Anderson [2002]) will maximize predictive accuracy; however, predictive accuracy is only maximized across a predefined set of parametric models.

Another importance difference between the 2 analytical philosophies is in their approach to model assessment and validation. In statistics, model fit is most often assessed by determining whether one model does better than some other model or than chance alone (e.g., likelihood-ratio tests of differences in residual deviance between models), or whether some model or models are better supported by data than others (comparisons of AIC values). Both of these assessments are usually based on the same data used to build the models. In contrast, data mining strongly emphasizes validating models and measuring model performance by assessing how well a model built with one set of data (known as the training set) can predict observations in a set of data that was not used to build the model (the test set). Approaches to this validation include cross-validation and the bootstrap, with accuracy assessed by metrics such as raw percentages of correct predictions, root mean squared error, cross-entropy, and area under the curve (AUC) values from response operative curves (ROCs).

Data mining encompasses diverse techniques, including decision trees (e.g., Breiman et al. 1984), neural nets (e.g., Mitchell 1997), and more recently, support vector machines (SVMs; e.g., Cristianini and Shawe-Taylor 2000), Bayes nets (e.g., Jensen 1996), and ensemble variants of tree-based methods (e.g., bagged and boosted decision trees, random forests; Breiman 1996, Breiman 2001). We provide a list of some software packages available for data mining in Appendix A. While data-mining techniques are very different in what they do, all share the feature that they are developed around a desire to place as few restrictions as possible on the models constructed in the analysis, leading to intuition-based algorithms (set of mechanistic steps and rules) through which data are processed to create a model that maximizes predictive performance. This is one reason that data-mining methods are assessed on empirical evidence of high predictive performance and not on theoretical grounds. The emphasis on algorithm creation and tuning means that even apparently identical software, when written by different programmers, can produce results that are not identical because of subtle differences in details of the algorithms.

All of the above is a caricature of distinctions between data mining and parametric statistics because the 2 fields and their methods are often interrelated. For example, the most simplistic neural nets are parametric logistic regressions, SVMs are a form of thin-plate spline, and the data-mining ensemble technique of boosting and statistic's generalized additive models have a theoretical connection (Friedman et al. 2000).

?3

For several reasons, we will use one method and its results to illustrate the utility of data mining: bagged (Breiman 2001) decision trees. First, while there is no single best method for all data, bagged decision trees consistently perform as well as the best other current classification

methods (Caruana and Niculescu-Mizil 2006). Second, no post–model-building calibration step is required to achieve highest predictive performance, unlike techniques such as boosted decision trees or SVMs for binary classification (Caruana and Niculescu-Mizil 2006); thus, bagged decision trees are relatively easy to implement and understand. Third, decision-tree analyses will automatically (i.e., no need for imputation) make use of data cases that have missing values for any one or several predictors, thus making decision-tree methods highly flexible with regard to the data that they can use. Fourth, decision-tree methods are also highly flexible because they automatically discover interactions (nonadditive relationships) among predictors. Fifth, decision-tree methods are relatively easy to explain and illustrate with this information already published in the ecological literature (De'ath and Fabricius 2000), thus allowing us to emphasize the results and not the methods themselves.

Briefly, one builds a decision-tree model by splitting a data set into groups, typically 2, and then recursively splitting each sub-group into smaller pieces. Each split partitions data based on values of one predictor variable, using a specific criterion and algorithm for identifying the rule that maximizes the information content of each split. Each predictor can appear multiple times within a decision tree in different decision rules. Decision trees can be used with response variables of a number of types including time-to-failure (survival) data, with decision-tree analysis of categorical (termed classification decision trees) and continuous (termed regression decision trees) responses being likely the most commonly used.

Like any highly flexible model, a single decision tree can over-fit data, producing a model that is too highly tailored to a specific sample of data. Early methods for dealing with over-fitting used various rules to "prune" back the branches of a decision tree (e.g., Breiman et al. 1984); however, pruning is not required using current ensemble decision-tree methods. Instead, ensemble methods combine information from multiple decision trees or tree components, which minimizes over-fitting. Bagging (Breiman 1996) is one such method. One builds bagged decision trees by taking multiple bootstrapped samples of data each the same size as the original data set. Multiple trees, each built from a separate bootstrapped data set, form the ensemble. The bagged prediction estimate for each data case is the average predicted value from all of the trees in an ensemble. Typically, the predictions that are averaged are the predictions for the data cases not included in the bootstrap sample used to generate each tree in the ensemble (termed out of bag predictions). Over-fitting decreases and out-of-bag predictions improve rapidly with increasing numbers of bootstrapped samples, with improvements typically reaching a plateau with $\leq$50 bootstrapped samples (Breiman 1996).

THREE STRENGTHS OF DATA-MINING TECHNIQUES

We use examples to illustrate 3 facets of working with data-mining tools to explore data: 1) generating predictions, 2) identifying important predictor variables, and 3) discovering the forms of relationships between predictors and response. Our examples are of analyses of data from Project FeederWatch (e.g., Lepage and Francis 2002), with different data sets used in different examples; Appendix B describes the data in more detail.

Making Accurate Predictions

While accurate predictions are the primary goal of data mining, a data-mining analysis will not necessarily produce more accurate predictions than a parametric statistical analysis. However, when little is known about a system—data are available but little else is known—the likelihood of rapidly producing accurate predictions from data mining is higher than from statistical analysis because errors in model misspecification are not an issue in building data-mining models. Data mining also is likely to perform better when the relationships between predictors and response are complex and not expressible as a simple combination of categorical and linear relationships. Additionally, when a large number, tens to hundreds, of potential predictors are available in the data, the automatic aspects of data mining will further increase the likelihood that data mining will outperform statistical analysis. However, high performance of data mining, as well as statistical models depends on the predictors having relevant information. Note that data mining does not require a data set with a large number of predictors (see e.g., Elith et al. 2006) or large numbers of data records, although data mining is likely better suited to exploring these types of data.

Many demonstrations exist of the potential for superior prediction accuracy of data mining compared to that of parametric statistical analysis (e.g., Caruana and Niculescu-Mizil 2006), including ecological examples (Elith et al. 2006). These comparisons have in common an artificially level playing field: the same set of predictor variables and the same cases of response data are used by all methods being compared. However, reality can be very different. A parametric statistical analysis will potentially only examine a relatively small number of predictors compared to a data-mining analysis, and the variables included in that small set will depend on the judgment of the analyst.

We contrasted the predictive performance of a naïve data-mining model with the performance that could be obtained from a statistical model guided by years of prior experience and data discovery. Our example analysis involves predicting presence and absence of house finches (*Carpodacus mexicanus*) across 11 winter seasons (1993–1994 to 2003–2004) and within a biogeographically consistent geographic area: the Appalachian mountains of the eastern United States with boundaries defined to be the Partners in Flight Bird Conservation Region 28 (see http://www.nabci-us.org/map.html). The response variable was binomial as observers recorded the presence or absence of reported house finches at a site during an individual observation period, with the multiple observations from each site being separate data points. We built an ensemble of 100 decision trees for the bagged decision-tree analysis, using the full set of 205

predictors available in our data set; see Appendix B. We used the accuracy of out-of-bag predictions to assess performance of the bagged decision-tree model. The statistical analysis used in the comparison was a single mixed-model logistic regression (PROC GLIMMIX in SAS Version 9.1.3). We chosen the terms in the statistical model (see Appendix C) based on >5 years of experience with statistical analysis of the data and of important predictors of presence and abundance of house finches (Hochachka and Dhondt 2000, 2006; Hosseini et al. 2006). The data set used in the logistic regression was a single random subset of the total data set, with each observation having a probability of 0.632 of being included in the data; the remaining data were withheld from model building, to be used in cross-validation. The proportion 0.632 is the theoretical expectation of an individual data point being included in a single bootstrap sample (Harrell 2001), and we used it to mimic the number of unique points that would appear in each decision tree within the bagged ensemble. We calculated performance metrics using the software PERF version 5.11 (http://kodiak.cs.cornell.edu/kddcup/software.html).

The bagged decision tree and logistic regression analyses had almost identical performance based on several metrics, with slightly higher accuracy for the data-mining analysis. Data mining predicted presence and absence with 85.4% accuracy, while accuracy was 84.0% for the logistic regression; accuracies were based on a 50% cut-off between presence and absence predictions. The AUC values were 0.917 for the bagged decision trees, and 0.907 for logistic regression. The AUC values of $\geq$0.9 are conventionally considered outstanding (Hosmer and Lemeshow 2000). Both model-building techniques produced highly accurate predictions; thus, there was no obvious advantage to using data mining based on measures of model accuracy.

However, note that with no more than the most basic ecological insights needed to determine the types of predictors entering the analysis (e.g., house finch presence is likely to vary through time and space, with habitat structure, and with human modification of habitat), we predicted presence of house finches using bagged decision trees with slightly better accuracy than was possible with a statistical model based on over a half decade of experience with the data and the biological system. For an ecologist needing, for example, to predict distribution of a species quickly and accurately, this is a compelling example of the benefits of adapting data-mining tools for data exploration. Conversely, these results also suggest that the expert's level of understanding of this system and the resultant ability to describe the system as a generalized linear mixed model was high, validating what was previously just a supposition of understanding.

Identifying Important Predictors

No single method is universally used to identify important predictor variables from data-mining models. One established method is a deviance-based method described by Breiman et al. (1984). Another widely accepted method compares predictive performance of a model and test data set with the predictive performance of this same model when the predictor variable of interest has been randomly shuffled among the cases in the data set (a randomization test known as a sensitivity analysis within the data-mining field; Breiman 2001). If a predictor variable is important, randomizing associations between predictor and response will decrease the accuracy of predictions relative to a model built with the true data. Larger declines in predictive performance indicate predictor variables that are more important. While this method is intuitively appealing, it is also computationally expensive and time-consuming because each of the many predictors must be permuted and run through the model, or models when an ensemble method is used.

When tree-based data-mining methods are used, a number of more rapidly computed measures of variable importance are also available (see Caruana et al. 2006). These faster methods summarize information on tree structure (e.g., No. of cases split on each predictor). While these more rapidly computed measures could compare favorably with Breiman's (2001) method in terms of similarity of rankings (Caruana et al. 2006), this result is not theoretically guaranteed and Breiman's method would be preferred unless severe computational constraints exist.

Continuing the comparison example from the last section, our choices of predictors for inclusion in the logistic regression were largely supported by the important predictor variables identified using bagged decision trees (Table 1). Of the 8 fixed-effect predictor variables in the logistic regression, 7 of these were directly identified as being important in the data-mining models using 2 criteria based on tree structure (Caruana et al. 2006). All 7 of these predictors were ranked within the top 11 variables by importance using both variable-importance measures shown in Table 1. The eighth fixed-effect predictor in the logistic regression was not directly indicated as being an important decision-tree predictor, but information represented by this predictor was present and important in the data-mining models. This last logistic-regression variable was an ordinal, 4-category description of urbanization, from rural to urban, provided by Project FeederWatch participants. This predictor was partially redundant in the logistic regression model because it was correlated with human population density data from the 2000 United States Census ($r = 0.36$). In the logistic regression only the ordinal variable and not the continuous, census-derived variable had estimated effects that differed from zero, based on confidence limits. In contrast, the data-mining model identified United States Census-based human population density as its important human-density predictor (Table 1). The decision-tree analysis identified several important predictors not included in the logistic regression as fixed effects (Table 1). We suspect that most if not all of these decision-tree predictors were functionally replaced in the logistic regression by the random effect accounting for site-to-site differences, and

Table 1. The most important predictor variables determining presence or absence of house finch reports in winter within Bird Conservation Region 28 (Appalachian Mountains, USA), as determined from a bagged decision-tree model. We list the 20 most important predicators based on each of 2 criteria for variable importance: an index of the number of data cases included at least once in a decision rule involving the predictor (the point count–no repeats criterion), and the reduction in deviance resulting from partitioning using decision rules involving the predictor (the deviance criterion; Breiman et al. 1984). Six predictors were unique to the top 20 set for each criterion. For both criteria, larger values indicate greater importance and predictors are sorted from highest to lowest importance based on the point count–no repeats criterion.

Name of predictor variable	Predictor importance: point count–no repeats criterion	Predictor importance: deviance criterion[a]
Proportion of human housing units vacant[b]	68.8	279
No. half-d observation[c,d]	61.1	234
D from season start[c,d]	49.6	1,465
Latitude[c,d]	46.4	268
FeederWatch season[c,d]	41.0	680
Human population density[b,d]	34.9	154
No. bird feeders, hanging[c]	33.3	191
Elevation, U.S. Geological Survey National Elevation dataset[e]	30.8	116
Elevation, GTOPO30 Digital Elevation model[d,e]	27.2	129
Hr of observation effort[c,d]	25.8	187
No. water sources for birds[c]	25.2	856
Human households density[b]	24.7	85
No. feeders, suet[c]	23.7	141
30-yr annual average of monthly snowfall amt[e]	21.1	65
Longitude[c]	20.7	134
Mean no. children/human household[b]	20.6	103
No. bird feeders, thistle[c]	19.9	80
No. human family households[b]	18.6	104
Proportion humans 30–39 yr old[b]	18.5	100
No. bird feeders, platform[c]	17.4	104
Density of humans with multiple races[b]	17.2	107
No. feeders, ground[c]	13.6	112
Max. temp during observation period[c]	11.6	203
Min. temp during observation period[c]	11.5	199
No. deciduous shrubs and trees within site[c]	11.5	140
Snow depth during observation period[c]	9.9	107

[a] The deviance measure of variable importance is expressed as deviance $\times 10^{-3}$.
[b] Predictor variable comes from U.S. 2000 census block level information.
[c] Predictor variable comes from Project FeederWatch participant-supplied data.
[d] Predictor used as a fixed-effect in the logistic regression model that is compared to the data-mining model.
[e] Predictor variable comes from one of several miscellaneous GIS data layers described in Appendix B.

that the data-mining analysis may be revealing the important underlying axes along which sites varied.

A caveat in the interpretation of any variable-importance measures is that they are sensitive to the presence of correlated predictors. When a set of correlated predictors is important for generation of accurate predictions, the variables in the set will potentially share the task of partitioning the data; thus, a variable-importance measure for any single variable within the set can underestimate the importance of the latent variable underlying the set. While interpretation of models with correlated predictors is problematic in both statistical and data-mining analyses, this issue is more likely to be present when larger numbers of predictors are used in the building of data-mining models.

Revealing Forms of Functional Relationships

Some additional computations are required in order for data-mining analyses to reveal the functional forms of relationships between predictors and response variables. The most widely used method is the computation of partial dependence values (Friedman 2001, Hastie et al. 2001), which describe the effect of the predictor on the modeled response after accounting for the average effect of all other predictors. Because these computations rely only on model predictions, partial dependence functions can be used to interpret any predictive model, statistical or from data mining.

When calculating partial dependence values, predictors may be focal predictors (whose effects we are investigating), otherwise they can be viewed as nonfocal, nuisance predictors. A partial dependence prediction is made by fixing each focal predictor at a single value for all cases in the data set, and averaging across the joint values of the nuisance predictors. For example, to compute the effect of a 1 January date on house finch prevalence, we first replace all actual date values in our data with the value for 1 January, while keeping all other predictors at their true values. Then we calculate the predicted probability of occurrence on 1 January by passing the synthetic data set through the data-mining model built using the real data. The average predicted response across all cases in the synthetic data set is the partial dependence value for the date of 1 January. This averaging procedure is repeated for each desired date (Fig. 1). By taking the mean we average out the variation in the modeled response due to all other predictors in the model, making it easier to uncover additive effects of date.

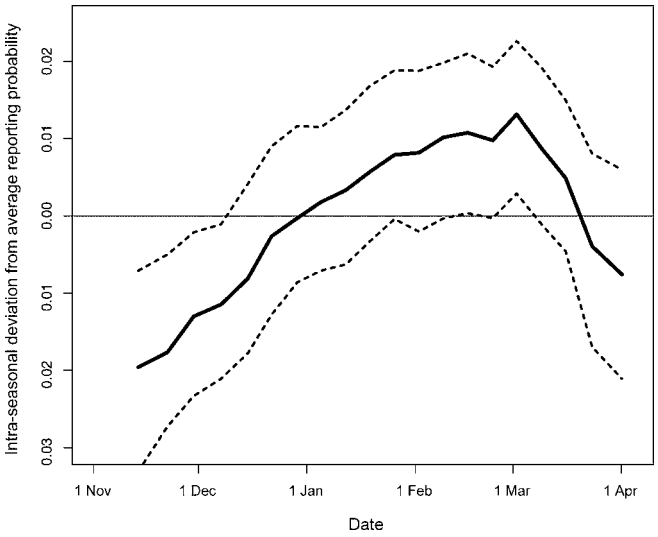

Figure 1. Apparent migration of house finches into Bird Conservation Region 27 (Southeastern Coastal Plain, USA) during the winter, illustrated by partial dependence values. Probabilities are expressed as deviations from the overall mean (horizontal line). We estimated confidence intervals by creating 100 bootstrap samples of the training data, building a bagged decision-tree model from each sample, and computing partial dependence values for each model within the bagged set. The overall trend is based on median partial dependence values (solid line), and we used inner 90% of values (dashed lines) to define the plotted confidence limits. We used Lowess smoothing for interpolation.

The partial effects (e.g., marginal means) commonly estimated in most statistical regression models achieve a similar type of control by conditioning on the values of the nonfocal predictors. Confidence limits can be bootstrapped around partial dependence values (Fig. 1), with each iteration of the bootstrap being the sampling with replacement of the original data, creation of a data-mining model on this sample, and calculation of partial dependence values. In practice, partial dependence calculations can be computationally expensive and are often approximated using Monte Carlo techniques.

Partial dependence values, just like marginal means from parametric analyses, will only be ecologically informative if the focal predictor or predictors that are systematically varied have effects on the response that can be logically isolated from the influence of all nuisance predictors. In statistical terms this means that no interaction exists. As an example of an interaction, many species of birds exhibit within-season variation in their winter distributions. We can calculate appropriate partial dependence values for known interactions, such as latitude $\times$ longitude $\times$ time interactions, by simultaneously and systematically varying the values of latitude, longitude, and time (Fig. 2). While the presence of interactions can be tested formally in statistical models, there are currently no general and implemented methods for identifying specific interactions in data-mining analyses, although methods exist that work under constrained circumstances (e.g., Hooker 2004). Interaction detection is an active area of research in data mining.

DISCUSSION

We believe that data mining can play an important role or roles in many analyses of ecological data. While data-mining techniques were developed with an eye to analysis of large data sets—thousands of cases of data and tens to hundreds of predictor variables—such large data sets are not a prerequisite for using data mining. Data-mining methods can be used effectively with a few hundred data cases and 10 predictors (e.g., see Elith et al. 2006), or even smaller data sets. In an extreme case, use of data-mining tools may even be appropriate with a single predictor, if the functional relationship between predictor and response is complex and unknown. Nevertheless, data mining is similar to statistical analysis in that more detail can be discovered and confidence in conclusions will be greater when more data are available. The exact roles of data mining will depend on the degree to which prior knowledge exists about an ecological system. This prior knowledge can range from little or no understanding of the factors that cause variation in the response variable in question to a very good understanding of factors that affect the response variable.

When little is known about a system, data-mining techniques provide a natural method for exploratory data analysis with all of the strengths that we discussed in the previous sections: 1) rapid production of predictions, 2) identification of the variables that are important in producing these predictions (Table 1), and 3) the ability to examine the forms of relationships between predictors and the response variable (Figs. 1, 2). Additionally, by bootstrapping data-mining estimates one can compute estimates of confidence that take into account the relatively large degree of model uncertainty associated with exploratory analyses. All of the output products from data mining would yield a rapid increase in understanding of a system, and allow informed decisions about further work, either confirmatory analyses or more focused collection of additional data. Further, the results from data-mining analyses set a benchmark for prediction accuracy, to which later statistical analyses can be compared.

Data-mining analysis is also useful when one has an intermediate level of knowledge about a system. In such cases, contrasting predictive performance of a data-mining and a statistical model would indicate the extent to which an existing statistical model produces accurate predictions and, thus, is an adequate description of the system. The flexibility and highly automated fitting of most data-mining models make them a good choice for use as objective benchmarks of the information contained in the predictor variables. Substantially lower accuracy of the statistical model would suggest that the analyst would need to explore whether functional forms of relationships in the statistical model are adequate or whether there are important predictors additional to those already present in the statistical model. For example, residuals from an existing statistical model can be used as the response variable in a data-mining analysis, allowing identification of predictors of just the unexplained variation. Data mining of residuals could also show that

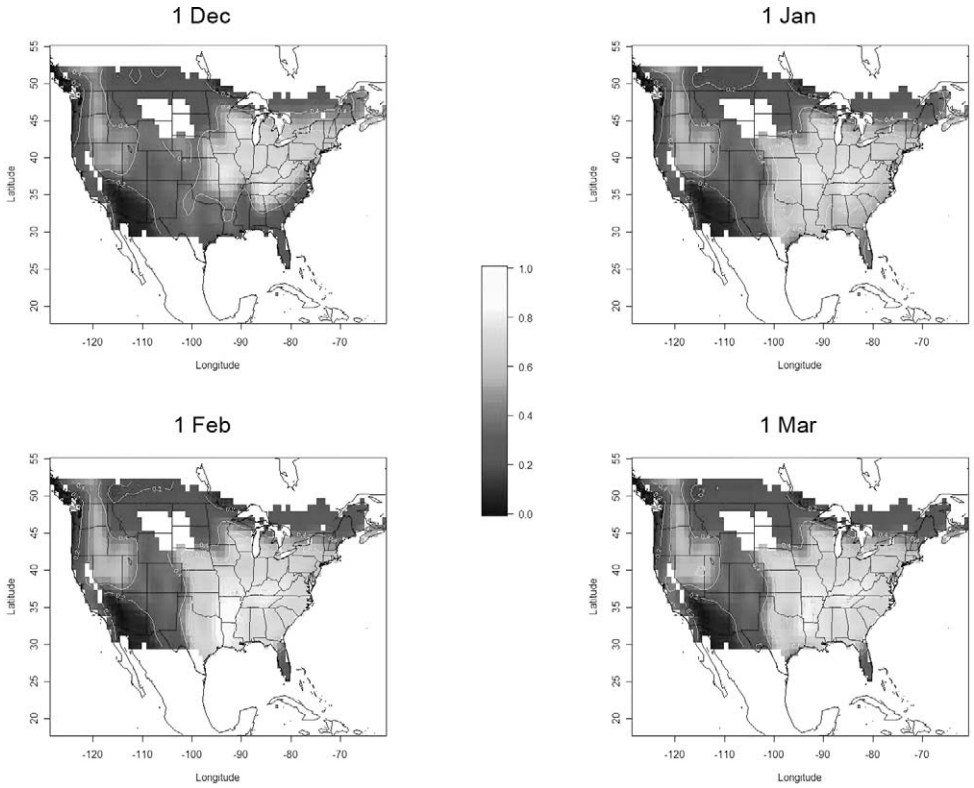

Figure 2. Within-winter variation in reported prevalence of American goldfinches (*Carduelis tristis*) across southern North America, used to illustrate the results of partial dependence calculation when interactions occur among the effects of predictor variables (latitude, longitude, and date-within-season). For each date, we calculated the partial dependence values at points on a systematic grid across much of North America, allowing visualization of within-winter changes in the geographic locations with the highest probability of reporting goldfinches. The white area in the Great Plains is a region with insufficient data for calculation of partial dependence value.

predictors currently in existing statistical models were still important in explaining residual variance, indicating that the form of the statistical model needs to be reevaluated; either the forms of relationships for continuous variables may be inappropriate, or statistical interactions may need to be added among predictors in the existing statistical model.

Another potential use of data mining, even when highly accurate statistical models exist, would be to increase the mechanistic accuracy of existing statistical models. For example, when we compared the accuracy of a logistic regression and bagged decision trees at predicting house finch presence (in the **Making Accurate Predictions** section), the statistical model that almost matched the performance of the decision trees made use of a random effect to account for consistent site-to-site variation in the probability of reporting house finches. As long as our goal is to account merely for this site-specific variation, then use of a random effect was appropriate. However, if we had wanted to understand the causes of consistent site-to-site differences, then data mining could be used to suggest the mechanisms behind the pattern.

We emphasize that regardless of the analysis techniques used to create or refine hypotheses, independent data need to be used for any confirmatory analyses of these hypotheses. Either data need to be withheld for confirmatory analysis before data mining is conducted, or additional data need to be collected. Unless this is done, parameter estimates and measure of performance of statistical models cannot be assumed to be valid. When large data sets are available it may be feasible to set aside subsets of data for training, validation, and finally, confirmatory statistical analysis. The only exception would be where data mining would be used after construction of a parametric statistical model, with the results from data mining used as a metric of suitability of the parametric model.

In most cases, the end product from a combination of data mining and statistical analysis will likely be a statistical model and not a data-mining model, for several reasons. First, a statistical model is an easier abstraction to understand than a data-mining model. Second, useful metrics such as confidence limits around predictions and parameter estimates and their confidence limits are easily generated in statistical analyses, but are computationally expensive (e.g., using bootstrapping) to generate using data-mining tools. Third, even if measures of confidence would be generated from a data-mining analysis, we suspect that intervals would be narrower from a comparable statistical analysis, a result that we have found in some preliminary experimentation. We suspect that our finding of narrower confidence intervals are the result of model-uncertainty enlarging confidence intervals from ensemble data-mining models, in the same way that confidence intervals are wider than single-model confidence intervals when using the AIC-based multimodel inference paradigm (Burnham and Anderson 2002).

Note, however, that the ability to explore a wider model set, which is the corollary of wider confidence limits, is desirable in many instances. A fourth advantage of statistical models is the ability to model multiple processes explicitly and separately as part of a single analysis. Most notably for ecologists, this is used in the separate modeling of a detection process and a biological process in capture–mark–recapture and similar (White and Burnham 1999) analyses. This or similar hierarchical processes cannot be readily modeled using current data-mining techniques, although extension of data mining for hierarchical model construction is an active area of research.

However, circumstances exist in which a data-mining model would be the desired end product of analysis of a data set. One instance is when the primary goal would be production of accurate predictions (e.g., Peters 1991), either as an end in itself or as part of an exploratory analysis that would guide the collection of additional data. Another instance is when important predictor variables are likely to have missing values for many cases, with the result that sample sizes from statistical models would be reduced greatly. While these missing cases could be imputed from the data, an appropriate data-mining technique, such as tree-based techniques, automatically can make use of data cases even with missing values.

MANAGEMENT IMPLICATIONS

The choice of appropriate technique for data analysis depends in part on the goals of the analysis, and statistical analysis techniques may not be appropriate for meeting management goals. Instead, data mining should be considered as the analysis tool when little prior knowledge exists of an ecological system or when accurate predictions are the desired product from an analysis. Both of these conditions are often met in reality, and we believe that wildlife managers, and ecologists in general, should make more use of data-mining techniques. When used appropriately (e.g., using cross-validation to assess performance, using independent data in confirmatory analyses following data exploration), data-mining analyses are not data dredging and their strengths in exploratory data analysis make data mining a logical component, or even end product, of a thorough analysis of data. Further, regardless of the degree of prior knowledge of a system, data-mining analyses can provide an objective benchmark against which to compare the performance of statistical analyses; such benchmarking of performance is difficult if not impossible using statistical techniques alone.

ACKNOWLEDGMENTS

We thank the participants in Project FeederWatch for providing the data used in the illustration of data-mining techniques, and the staff in the Information Sciences unit at the Cornell Laboratory of Ornithology for their work in managing these data. Work on this paper was funded under the National Science Foundation Information Technology Research (ITR) for National Priorities program (award No. EF-0427914). We thank the P. Hurtado, C. D. MacInnes, G. White, and one anonymous referee for their comments, which focused and improved the paper from its original form.

LITERATURE CITED

Bauer, E., and R. Kohavi. 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning 36:105–139.

Breiman, L. 1996. Bagging predictors. Machine Learning 24:123–140.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Chapman & Hall, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Caruana, R., M. Elhawary, A. Munson, M. Riedewald, D. Sorokina, D. Fink, W. M. Hochachka, and S. Kelling. 2006. Mining citizen science data to predict prevalence of wild bird species. Pages 909–915 *in* M. Craven and D. Gunopulos, editors. Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, New York, New York, USA. ?4

Caruana, R., and A. Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. Pages 161–168 *in* W. W. Cohen and A. Moore, editors. Proceedings of the 23rd international conference on machine learning. ACM Press, New York, New York, USA.

Cristianini, N., and J. Shawe-Taylor. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge, United Kingdom.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81: 3178–3192.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, G. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Freund, Y., and R. E. Schapire. 1996. Experiments with a new boosting algorithm. Pages 148–156 *in* L. Saitta, editor. Machine learning: proceedings of the thirteenth international conference. Morgan Kauffman, San Francisco, California, USA.

Friedman, J. 2001. Greedy function approximations: a gradient boosting machine. Annals of Statistics 29:1189–1232.

Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statistical view of boosting. Annals of Statistics 28:337–374.

Friedman, J. H., and B. E. Popescu. 2005. Predictive learning via rule ensembles. Technical Report. Stanford University Department of Statistics technical report, Palo Alto, CA, USA.

Harrell, F. E., Jr. 2001 Regression modeling strategies with applications to linear models, logistic regression and survival analysis. Springer Verlag, New York, New York, USA.

Hastie, T., and R. Tibshirani. 1990. Generalized additive models. Chapman and Hall, London, United Kingdom.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer Verlag, New York, New York, USA.

Hochachka, W. M., and A. A. Dhondt. 2000. Density-dependent decline of host abundance resulting from a new infectious disease. Proceedings of the National Academy of Sciences USA 97:5303–5306.

Hochachka, W. M., and A. A. Dhondt. 2006. House finch (*Carpodacus mexicanus)* population- and group-level responses to a bacterial disease. Ornithological Monographs 60:30–43.

Hooker, G. 2004. Discovering additive structure in black box functions. Pages 569–574 *in* W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors. Proceedings of the 10th SIGKDD international conference on

knowledge discovery and data mining. ACM Press, New York, New York, USA.
Hosmer, D. W., and S. Lemeshow. 2000. Applied logistic regression. Second edition. John Wiley and Sons, New York, New York, USA.
Hosseini, P. R., A. A. Dhondt, and A. P. Dobson. 2006. Spatial spread of an emerging infectious disease: conjunctivitis in house finches—seasonal rates and geographic barriers. Ecology 87:3037–3046.
Jensen, F. V. 1996. An introduction to Bayesian networks. Springer Verlag, New York, New York, USA.
King, R., and S. P. Brooks. 2004. Bayesian analysis of the Hector's dolphins. Animal Biodiversity and Conservation 27:343–354.
Lepage, D., and C. M. Francis. 2002. Do feeder counts reliably indicate bird population changes? 21 years of winter bird counts in Ontario, Canada. Condor 104:255–270.
Mitchell, T. 1997. Machine learning. McGraw-Hill, New York, New York, USA.
Peters, R. H. 1991. A critique for ecology. Cambridge University Press, Cambridge, United Kingdom.
Wells, J. V., K. V. Rosenberg, E. H. Dunn, D. L. Tessaglia-Hymes, and A. A. Dhondt. 1998. Feeder counts as indicators of spatial and temporal variation in winter abundance of resident birds. Journal of Field Ornithology 69:577–586.
White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked animals. Bird Study 46(Supplement):120–138.
Wood, S. N. 2006. Generalized additive models: an introduction with R. Chapman and Hall/CRC, Boca Raton, Florida, USA.

APPENDIX A: SOME AVAILABLE DATA-MINING SOFTWARE

A large number of software packages are available for conducting data-mining analyses. Below we list some packages with which the authors have experience, and no attempt has been made to be comprehensive. Instead, our goal was to give readers some suggestions on where to start with exploring data mining. Both freely available and commercial products are listed, and the basic abilities of each software package are described.

C5

C5 is a commercial decision-tree and rule-learning package developed by RuleQuest Research (St Ives, NSW, Australia). C5 is the latest release of the popular earlier C4 and C4.5 decision-tree software. C5 is a mature product with a sophisticated user interface and many advanced capabilities. One attractive feature of C5 is the ability to translate a learned decision tree into a set of rules, and to then further refine those rules using the training set. In some applications rules are easier to interpret than trees. C5 and other data-mining software from RuleQuest can be found at: http://www.rulequest.com/.

CART

CART (Classification and Regression Trees) is a commercial decision-tree package sold by Salford Systems (San Diego, CA). Like C5, CART is a mature product with a sophisticated user interface and capabilities. CART has been under constant development for 2 decades, and is one of the more refined decision-tree packages currently available. Users of CART also have the option of hiring consultants from Salford Systems or attending data-mining courses offered by Salford Systems. Salford Systems also sells other data-mining packages such as MARS, TreeNet, and RandomForests. Although IND (see below) has a CART emulation mode, this is a very limited emulation of some aspects of CART decision trees circa 1990. It is not a replacement for the full CART software. CART and other software distributed by Salford Systems can be found at: http://www.salford-systems.com/.

IND

IND (Induce) is a decision-tree package written by Wray Buntine and distributed by the NASA Ames Research Center (Moffett Field, CA). IND source code can be downloaded free of charge from: http://opensource.arc.nasa.gov/project.jsp?id=7. The package has been updated very little since its original release in 1992. The software runs only under Unix environments (including Linux and CYGWIN under Windows). The many options in IND are selected via command-line arguments; there is no GUI interface. IND can emulate a variety of decision-tree types such as CART, C4, ID3, as well as several types of tree specifically designed to predict probabilities. IND also has limited capabilities to create train and test sets and run experiments automatically. Although IND was written before methods such as bagging, boosting, and random forests were developed, it has limited capabilities to average multiple-tree models. Users interested in these ensemble methods, however, would need to write additional code. We used IND augmented with our own code for the examples in Table 1 and the 2 figures.

SAS Enterprise Miner

SAS Enterprise Miner provides a variety of data-mining capabilities, including decision trees, neural nets, and logistic regression. Enterprise Miner provides an exceptionally complete set of integrated tools for processing data, performing statistical analyses, data mining, and fielding solutions. Enterprise Miner can be used from a GUI, or called from the well-known SAS modeling language. As with CART, consultants and workshops are available for SAS users. More information about Enterprise Miner can be found at: http://www.sas.com/technologies/analytics/datamining/

R

The freely available R statistical programming language makes a number of types of data mining available to users. While the core software does not have any data-mining facilities, several libraries of routines have been written for R, or can easily be implemented in R. Among the data-mining methods that are or can be implemented in R are:

rpart.—The rpart library procedures build a single classification or regression-tree models of a very general structure, including the classification and regression trees of Breiman et al. (1984). The types of endpoints that rpart handles includes classifications (such as yes or no), continuous values (such as body mass), Poisson counts (such as counts of animals), and survival information (time to death). The rpart library includes tools to model, plot,

and summarize the end results. The rpart routines are stable, well-documented, and easy to modify and extend. Several extensions are available.

Bagging.—A bagging ensemble is constructed by bootstrapping the data, with the predictions from all bootstrap models averaged together. While no specific bagging libraries are present in R, it is straightforward to write bootstrapping code using the boot library and apply it to the rpart library (see below).

Boosting.—Boosting (Freund and Schapire 1996, Friedman 2001) adaptively creates new members of an ensemble, focusing effort on those parts of the data that are hardest to fit. The gbm (Gradient Boosting Machine) library in R allows one to fit boosting models for a variety of responses including continuous, binary classification, counts and survival events with a variety of losses. One of the advantages of this library is its tools for measuring the relative importance of the predictors and calculating partial dependence. The package is stable and well-documented. Friedman et al. (2000) have explored connections between GAM models and boosting.

RuleFit.—RuleFit (Friedman and Popescu 2005) is a 2-step procedure that works by first generating a large set of candidate rules, with each rule consisting of a conjunction of a small number of simple statements that are functions of individual input variables. In the second step a penalized criterion is used to construct an ensemble, or weighted average, of rules with good predictive performance. These rule ensembles can model nonlinear predictor effects and predictor interactions automatically. RuleFit achieves predictive accuracy comparable to the best current decision-tree methods for continuous responses. RuleFit has been implemented in the statistical computing language R by its authors, although it is not available as a standard R library, but must be downloaded from the RuleFit authors' own website. Routines are available to measure the relative importance of the predictors and describe their effects on the response as well as search for evidence of significant predictor interactions. The RuleFit is (at the time of writing) still in beta stage, which means that it requires a little patience and experimentation to get it to work. Our recent experimentation with RuleFit suggests that predictions from binomial responses may need to be calibrated to ensure accuracy.

WEKA

WEKA is an open source set of data-mining programs written in JAVA and distributed by the University of Waikato, New Zealand. WEKA has only been around for a few years, but already has become very popular. One of the key strengths of WEKA is the use of a uniform data format for all of the learning methods it incorporates. WEKA includes routines for several flavors of decision trees, as well as for neural nets, SVMs, logistic regression, random forests, bagging, boosting, and a variety of learning algorithms. Because WEKA is not a commercial package, the quality of the implementations is somewhat uneven. For example, we find that Breiman and Cutler's original Random Forest code sometimes learns better solutions than the WEKA Random Forests implementation. Despite these limitations, WEKA is a convenient and low-cost place to start for users who do not want to have to write their own code and also prefer to forego the expense and limitations of depending on commercial packages. WEKA is available at: http:///www.cs.waikato.ac.nz/ml/weka/ and as an R library. ?6

APPENDIX B: DATA USED IN EXAMPLES

To illustrate use of data mining we present examples based on data on presence or absence of wintering birds in North America. These data come from Project FeederWatch (PFW), a winter-long monitoring project in which members of the general public throughout the United States and Canada record the maximum number of birds seen together at one time, over 2-day observation periods, for each of the bird species that they see at their feeders. Observation periods are typically at weekly or biweekly intervals, over a season between mid-November and the very beginning of April. The maximum possible number of observation periods in one winter season is 21. In addition to recording the location, date, bird numbers, and effort expended in the observation process, participants are also asked to provide data describing the weather and the environments around their feeder locations, such as presence or absence of coniferous and deciduous trees, water bodies, and the degree to which landscapes are altered by humans. For more details see Wells et al. (1998), and Lepage and Francis (2002).

In addition to the information provided by PFW participants, we also collected several other descriptors of sites, describing the general biogeographic region, local habitat, climate, and human-related environmental features. These data came from existing GIS layers, and were extracted based on the latitudes and longitudes of the PFW sites. Of this larger list, we narrowed our data set to include only 205 predictors. This set included all of the data provided by PFW participants (77 predictors), each site's Bird Conservation Region (see http://www.nabci-us.org/map.html), as well as United States Census Bureau census block-level human demographic summaries (36 predictors; 2000 census), long-term climate descriptors (81 predictors; National Climatic Data Center's Climate Atlas of the United States), elevation (2 from different digital elevation data sources and resolutions: United States Geological Survey National Elevation dataset, 10-m resolution data http://www.mapmart.com/DEM/DEM.htm; and GTOPO30, 30-arc-second-resolution data http://edc.usgs.gov/products/elevation/gtopo30/gtopo30.html), and habitat type of the site from the United States National Land Cover Database (presence/absence of each of the 9 separate Anderson level 1 habitat classification categories within the grid block of the count site; U.S. National Landcover Data, 1992 version). ?7

APPENDIX C: DESCRIPTION OF EXAMPLE MIXED-MODEL LOGISTIC REGRESSION

Because of long past experience with these data, we created only a single parametric statistical model for predicting presence and absence of house finches. The statistical model contained fixed effects for: observer effort (2 categorical predictors: the No. of half-days within a 2-d observation period during which at least some observation time was spent, and a 4-category ordinal measure of the No. of hrs of observation), latitude (continuous linear predictor), elevation (continuous linear predictor), season (linear and quadratic predictors; each season was a winter season and not a calendar yr), day of season (continuous linear and quadratic predictors; 1 Nov was d 1 for a winter season), housing density (categorical predictor; a 4-category ordinal classification along a rural-to-urban gradient; provided by PFW participants), human population density (continuous linear predictor), and a number-of-half-days-observation × housing density interaction (categorical predictor). The logistic regression used a location identifier variable as a random effect with variance components covariance structure.

Associate Editor: White.