LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Steffen Oppel, Carolin Strobl and Falk Huettmann

Alternative Methods to Quantify Variable Importance in Ecology

Technical Report Number 65, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Alternative methods to quantify variable importance in ecology

STEFFEN OPPEL[1,4], CAROLIN STROBL[2], AND FALK HUETTMANN[1,3]

[1] *Department of Biology and Wildlife, 211 Irving 1, University of Alaska, Fairbanks, Alaska 99775-6100, USA*
[2] *Department of Statistics, Ludwig-Maximilians-Universität, Ludwigstraße 33, D-80539 München, Germany*
[3] *EWHALE lab, Institute of Arctic Biology, 419 Irving 1, University of Alaska, Fairbanks, Alaska 99775-6020, USA*
[4] Corresponding Author. E-mail: steffen.oppel@gmail.com

Ecological data are highly complex, often with a large array of variables interacting and explaining different components of the dependent variable of interest. Because nature itself is very complex, most of the variables measured by ecologists do not operate independently, so that interactions and correlations between variables need to be accounted for. Moreover, the number of variables influencing an ecological phenomenon may be very large, and the use of simplifying indices in ecology and wildlife management has been criticized for many years (Anderson et al. 2003).

Determining which variables have the greatest effect on a response variable can be a formidable challenge in many ecological data sets. Traditional linear regression models, which are widely used as tools to quantify and understand the ecological relationship between several explanatory variables and a dependent, reach their limitations when the number of predictor variables is large (Fielding 1999, Breiman 2001b, Burnham and Anderson 2002).

Despite the recognized shortcomings of generalized linear regression models (GLM), such approaches are still widely used and promoted in recent literature (Planque and Buffaz 2008, Yee et al. 2008, Bolker et al. 2009, Gompert and Buerkle 2009, Koper and Manseau 2009). Many ecologists are hesitant to use Bayesian approaches (Cressie et al. 2009), or machine learning methods (Cutler et al. 2007, Olden et al. 2008, Cutler et al. 2009), both of which have existed for several decades and have become popular among ecologists over the past 10 years (Fielding 1999, De'ath and Fabricius 2000). Poor understanding of advanced techniques and an inherent reluctance to try novel approaches are likely causes for the slow adoption of Bayesian and machine learning techniques (Bolker 2009, Uriarte and Yackulic 2009).

Recently, in a paper published in *Ecology* (Murray and Conner 2009), the presentation of methods to quantify variable importance was limited to standard parametric methods and the results appeared to contradict the statistical literature. This paper prompted us to (*i*) remind ecologists of the great utility of machine learning methods, which can provide enhanced and reliable measures of variable importance even in situations with a large number of predictor variables containing potentially complex interactions, and (*ii*) emphasize the importance of using correct terminology when evaluating statistical approaches. Specifically, we clarify the identification and simulation of spurious variables.

Alternative approaches to modeling high dimensional data

For modeling high dimensional data containing potentially complex interactions, a variety of new methods adopted from machine learning have become popular in many disciplines such as genetics and, more recently, ecology. Some of these so called algorithmic models can incorporate many predictor variables, and methods exist to reliably identify the most important predictors (Strobl et al. 2007, Archer and Kimes 2008, Strobl et al. 2008). Algorithmic models encompass a suite of analytical approaches (Elith et al. 2006); for simplicity and brevity we here focus on a widespread ensemble method, the Random Forest (Breiman 2001a), to demonstrate that the concept of algorithmic models is easy to understand and implement with freely available software solutions for virtually any ecologist worldwide.

In contrast to traditional data models like GLMs, algorithmic models do not require the *a priori* specification of a model to relate explanatory and dependent variables, but rather use an algorithm to learn the form of those relationships (Breiman 2001b). The basic decision tree algorithm underlying many modern algorithmic models was developed in the 1980s (Breiman et al. 1984), and was introduced to ecologists more than a decade ago (O'Connor and Jones 1997, Fielding 1999, De'ath and Fabricius 2000). Since then, classification trees and more advanced techniques have increased in popularity among ecologists due to their high classification accuracy, ability to incorporate a large number of predictor variables, ability to handle missing data, ability to characterize complex interactions among variables, and comparatively easy application and interpretation (Cutler et al. 2007, De'ath 2007, Hochachka et al. 2007, Elith et al. 2008, Olden et al. 2008, Elith and Graham 2009).

Algorithmic models have been used successfully in many ecological applications, such as analyses of species range shifts (Hill et al. 1998, Lawler et al. 2009), species richness patterns (Leathwick et al. 2006), species presence and distribution patterns (O'Connor et al. 2004, Peters et al. 2005, Elith et al. 2008, Elith and Graham 2009), identification of populations (Perdiguero-Alonso et al. 2008), and analyses of behavioral patterns (Grubb and King 1991, Low et al. 2006, Oppel et al. 2009). When compared to traditional statistical models such as GLMs, algorithmic models provide convenience, speed, and most importantly superior model fit and prediction (Elith et al. 2006, Prasad et al. 2006, Peters et al. 2007). They require substantially less prior knowledge about the study system to achieve the same accuracy as, for example, logistic regression models (Hochachka et al. 2007, Ritter 2007).

To understand how algorithmic models determine the importance of a variable it is necessary to briefly review the structure of algorithmic models. We emphasize that our comment is not designed to serve as a manual for the successful application of algorithmic models; such introductions already exist in the ecological literature (Cutler et al. 2007, De'ath 2007, Hochachka et al. 2007, Elith et al. 2008, Olden et al. 2008). Here we merely provide an exemplary, brief description of one particular algorithmic modeling technique, the Random Forest (Breiman 2001a, Cutler et al. 2007), and its extensions designed to reliably estimate variable importance in highly complex data sets (Hothorn et al. 2006, van der Laan 2006, Strobl et al. 2007).

A brief overview of Random Forests

The Random Forest algorithm is based on classification and regression tree analysis (Breiman et al. 1984, De'ath and Fabricius 2000). A classification or regression tree uses a series of rules to recursively split the data set into binary groups by identifying regions with the most homogenous set of a response to predictor variables. For each node the predictor variable and the split point are chosen to maximize the homogeneity of the data set along each of the two branches. Each branch can then be split again, either until a stopping criterion is reached, or until a user-specified number of terminal nodes is reached. The two main advantages of trees are that predictor variables can be both, categorical and continuous, and that irrelevant predictors are seldom selected for a split. Thus, there is no cost to including a large number of predictor variables. In contrast to GLMs, trees also incorporate, and benefit from, interactions due to the hierarchical structure within the tree. At each split the response depends not only on the value of the predictor at that split, but also on the predictors at all splits that occurred higher up in the tree. Further, trees are insensitive to outliers or missing values in a data set, which is a common occurrence in large spatial data sets (Craig and Huettmann 2008).

A Random Forest is an assemblage of a large number of classification or regression trees using two levels of randomisation in the construction of every tree in the Random Forest (Breiman 2001a). First, each tree is constructed from a random subset of the original data, either taken with a bootstrap sample with replacement or sampled randomly to a specified proportion of the entire data set. The data not chosen to construct the tree (termed 'out-of-bag' data, oob) are used to assess the predictive ability of that tree. Each tree thus provides both an algorithm to classify the data and an error estimate of predictive ability based on the oob data. Second, at each split within each tree a random subset *m* of the available predictor variables is used to partition the data set into two groups with minimal heterogeneity. Each tree recursively partitions the data using a random subset of predictor variables until homogeneity of the data in each terminal node cannot be increased by a further subdivision. After a user-specified number of trees (100s – 1000s) have been constructed, each data point is run down every single tree in the Random Forest. Different trees may predict different outcomes for the same data point, and the most common classification across all trees is used to determine the predicted outcome of a data point.

Variable importance estimation

To estimate the importance of predictor variables, Random Forests use a specific permutation procedure. In this procedure, the values for a given variable are randomly permuted over the oob data set and the resulting reduction in model accuracy is assessed. Variable importance is inversely related to the reduction in model accuracy after permutation (Strobl et al. 2007). For easier interpretation, the variable importance can be standardized, with the most important variable being assigned a relative variable importance of 100%. A Random Forest provided a reliable method to identify the most important predictor variables in a large simulation study including 100 variables (Archer and Kimes 2008). Hence, algorithmic models usually provide a simpler, more accurate and more widely applicable approach to determine variable importance in ecology than approaches that rely on correlations among variables or models with different subsets of the full suite of predictor variables (Burnham and Anderson 2002, Murray and Conner 2009). Advanced algorithms based on a conditional inference framework (Hothorn et al. 2006) are able to reliably identify the most important predictor variables even when continuous and categorical variables are used simultaneously (Strobl et al. 2007), or when variables are correlated (Strobl et al. 2008).

Conclusion

The continued use and promotion of simple linear techniques in ecology is troublesome because such models require a higher level of statistical knowledge to adequately describe the complexities of many large ecological data sets (Hochachka et al. 2007). Despite many recent advances, such as information-theoretic approaches to model selection (Burnham and Anderson 2002), the use of Bayesian approaches (Stauffer 2008, Cressie et al. 2009), and attempts to overcome the problem of spurious and correlated variables (Murray and Conner 2009), traditional regression models will rarely be able to match algorithmic models in situations where a large number of explanatory variables need to be included in a model that is based on a limited number of observations (Breiman 2001b). We therefore recommend that ecologists that are challenged by large data sets, interactions, and correlated variables consider the application of algorithmic models to improve the explanatory power and robustness of their analysis.

We realize that the widespread adoption of algorithmic approaches faces similar challenges as the adoption of hierarchical Bayesian modeling techniques (Uriarte and Yackulic 2009): ecologists struggle to use new analysis tools that go beyond their original training and education. We encourage ecologists to broaden their analytical horizons with existing literature (Cutler et al. 2007, Hochachka et al. 2007, Elith et al. 2008, Olden et al. 2008) and make use of powerful techniques that have been developed in fields outside of ecology in order to better understand ecological patterns and processes.

A note about spurious vs. suppressor variables

A spurious variable, for example in a multivariate regression problem with predictor variables x_1 through x_p and response variable y, is a variable that has no effect on the response, but is highly correlated with another predictor variable that does have an effect on the response (Burnham and Anderson 2002, Brett 2004). Consider, for example, that the number of storks that occur in an area (x_1) is correlated with the number of newborn infants in that area (y). However, stork abundance has no biological influence on the number of newborns; in fact, a third variable, for example a low degree of environmental pollution (x_2), positively influences both the number of storks and the birthrate. As long as environmental pollution is not entered as predictor variable in a model of birthrate, the number of storks will act as a spurious variable that could be mistaken to explain birthrate.

Spurious correlations can affect inference from data, and it is important to detect them. Murray and Conner (2009) recognized this problem and offered a solution based on simulations with artificial data. Unfortunately, the simulation design used by Murray and Conner (2009: *Ecological Archives* E090-026-S1), where the response variable was simulated together with the predictor variables, generated a suppressor variable (Velicer 1978, Smith et al. 1992, Maassen and Bakker 2001), rather than a spurious variable. While a suppressor variable also leads to a spurious correlation, the difference between spurious and suppressor variables is that, when considering the parameter estimates in a linear model, the effect of a spurious variable only appears when the truly relevant correlated predictor is absent from the model (Prairie and Bird 1989, Brett 2004). For a suppressor variable, however, the effect appears *only if* another correlated variable is entered into the model (Conger 1974, Tzelgov and Stern 1978, Velicer 1978, Smith et al. 1992).

The distinction between suppressor and spurious variables is not formally recognized by many ecologists (see Juenger and Bergelson (2000) for a notable exception), and may arguably be inconsequential if spurious correlations are considered in general. However, we argue that the distinction is important to facilitate effective communication with statisticians to make use of approaches developed outside the field of ecology. For example, the simulation by Murray and Conner (2009) lead them to erroneously conclude that a spurious variable could be identified by means of simple zero-order correlations. In fact, however, a spurious variable *cannot* be revealed by means of a zero-order correlation (which would indicate a strong association between the number of storks and the birthrate), but only by means of a partial correlation for the number of storks and the birthrate *given* the degree of environmental pollution (which would reveal that, once the truly influential variable is incorporated, the spurious variable “number of storks” proves irrelevant for predicting the birthrate)(Simon 1954). Because Murray and Conner (2009) unknowingly generated suppressor rather than spurious variables in their simulation, they found that the zero-order correlation was zero, while the partial correlation was not zero. This result is correct and unsurprising for describing a suppressor effect, but is misleading and inconsistent with statistical literature due to the incorrect usage of the term “spurious variable”.

Acknowledgments

We appreciate the cooperation of K. Murray in providing source code for their data generation. Three anonymous reviewers provided helpful comments on an earlier draft of the manuscript. SO was funded by a postdoctoral fellowship from the Institute of Arctic Biology.

Literature cited

Anderson, D. R., E. G. Cooch, R. J. Gutierrez, C. J. Krebs, M. S. Lindberg, K. H. Pollock, C. A. Ribic, and T. M. Shenk. 2003. Rigorous science: suggestions on how to raise the bar. Wildlife Society Bulletin **31**:296-305.

Archer, K. J., and R. V. Kimes. 2008. Empirical characterization of random forest variable importance measures. Computational Statistics and Data Analysis **52**:2249-2260.

Bolker, B. 2009. Learning hierarchical models: advice for the rest of us. Ecological Applications **19**:588-592.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution **24**:127-135.

Breiman, L. 2001a. Random forests. Machine Learning **45**:5-32.

Breiman, L. 2001b. Statistical modeling: The two cultures. Statistical Science **16**:199-215.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. 1984. Classification and regression trees. Wadsworth International Group, Belmont, CA.

Brett, M. T. 2004. When is a correlation between non-independent variables "spurious"? Oikos **105**:647-656.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference. A practical information-theoretic approach. 2nd edition edition. Springer, New York.

Conger, A. J. 1974. A revised definition for suppressor variables: a guide to their identification and interpretation. Educational and Psychological Measurement **34**:35-46.

Craig, E., and F. Huettmann. 2008. Using ‘blackbox’ algorithms such as Treenet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: An overview, an example using Golden Eagle satellite data and an outlook for a promising future. Pages 65-84 *in* H. F. Wang, editor. Intelligent data analysis: Developing new methodologies through pattern discovery and recovery. IGI Global, Hershey, PA, USA.

Cressie, N., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications **19**:553-570.

Cutler, A., D. R. Cutler, and J. R. Stevens. 2009. Tree-based methods. Pages 1-19 High-Dimensional Data Analysis in Cancer Research.

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random Forests for classification in ecology. Ecology **88**:2783-2792.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. Ecology **88**:243-251.

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., and C. H. Graham. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography **32**:66-77.

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, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**:129-151.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology **77**:802-813.

Fielding, A. H. 1999. Machine learning methods for ecological applications. Kluwer, Dordrecht, The Netherlands.

Gompert, Z., and C. A. Buerkle. 2009. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. Molecular Ecology **18**:1207-1224.

Grubb, T. G., and R. M. King. 1991. Assessing human disturbance of breeding Bald Eagles with classification tree models. Journal of Wildlife Management **55**:500-511.

Hill, D., P. Coquillard, J. de Vaugelas, and A. Meinesz. 1998. An algorithmic model for invasive species: Application to *Caulerpa taxifolia* (Vahl) C. Agardh development in the North-Western Mediterranean Sea. Ecological Modelling **109**:251-266.

Hochachka, W. M., R. Caruana, D. Fink, A. R. T. Munson, M. Riedewald, D. Sorokina, and S. Kelling. 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management **71**:2427-2437.

Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics **15**:651-674.

Juenger, T., and J. Bergelson. 2000. Does early season browsing influence the effect of self-pollination in Scarlet Gilia? Ecology **81**:41-48.

Koper, N., and M. Manseau. 2009. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. Journal of Applied Ecology **46**:590-599.

Lawler, J. J., S. L. Shafer, D. White, P. Kareiva, E. P. Maurer, A. R. Blaustein, and P. J. Bartlein. 2009. Projected climate-induced faunal change in the Western Hemisphere. Ecology **90**:588-597.

Leathwick, J. R., J. Elith, M. P. Francis, T. Hastie, and P. Taylor. 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series **321**:267-281.

Low, M., M. K. Joy, and T. Makan. 2006. Using regression trees to predict patterns of male provisioning in the stitchbird (hihi). Animal Behaviour **71**:1057-1068.

Maassen, G. H., and A. B. Bakker. 2001. Suppressor variables in path models: Definitions and interpretations. Sociological Methods Research **30**:241-270.

Murray, K., and M. M. Conner. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology **90**:348-355.

O'Connor, R. J., and M. T. Jones. 1997. Using hierarchical models to assess the ecological health of the nation. Transactions of the 62nd North American Wildlife and Natural Resources Conference **62**.

O'Connor, R. J., T. L. Wagner, and N. S. Sodhi. 2004. A test of a regression-tree model of species distribution. Auk **121**:604-609.

Olden, J. D., J. J. Lawler, and N. L. Poff. 2008. Machine learning methods without tears: A primer for ecologists. The Quarterly Review of Biology **83**:171-193.

Oppel, S., A. N. Powell, and D. L. Dickson. 2009. Using an algorithmic model to reveal individually variable movement decisions in a wintering sea duck. Journal of Animal Ecology **78**:524-531.

Perdiguero-Alonso, D., F. E. Montero, A. Kostadinova, J. A. Raga, and J. Barrett. 2008. Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. International Journal for Parasitology **38**:1425-1434.

Peters, J., B. D. Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. D. Becker, and W. Huybrechts. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling **207**:304-318.

Peters, J., R. Samson, and N. E. C. Verhoest. 2005. Predictive ecohydrological modelling using the random forest algorithm. Communications in Agricultural and Applied Biological Sciences **70**:207-211.

Planque, B., and L. Buffaz. 2008. Quantile regression models for fish recruitment-environment relationships: four case studies. Marine Ecology Progress Series **357**:213-223.

Prairie, Y. T., and D. F. Bird. 1989. Some misconceptions about the spurious correlation problem in the ecological literature. Oecologia **81**:285-288.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems **9**:181-199.

Ritter, J. L. 2007. Species distribution models for Denali National Park and Preserve, Alaska. M.Sc. thesis. University of Alaska Fairbanks, Fairbanks, AK.

Simon, H. A. 1954. Spurious Correlation: A Causal Interpretation. Journal of the American Statistical Association **49**:467-479.

Smith, R. L., J. W. Ager, Jr., and D. L. Williams. 1992. Suppressor variables in multiple regression/correlation. Educational and Psychological Measurement **52**:17-29.

Stauffer, H. B. 2008. Contemporary Bayesian and frequentist statistical research methods for natural resource scientists. John Wiley & Sons, Hoboken, NJ.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. BMC Bioinformatics **9**:307.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics **8**:25-45.

Tzelgov, J., and I. Stern. 1978. Relationships between variables in three variable linear regression and the concept of suppressor. Educational and Psychological Measurement **38**:325-335.

Uriarte, M., and C. B. Yackulic. 2009. Preaching to the unconverted. Ecological Applications **19**:592-596.

van der Laan, M. 2006. Statistical inference for variable importance. International Journal of Biostatistics **2**:1008 - 1008.

Velicer, W. F. 1978. Suppressor variables and the semipartial correlation coefficient. Educational and Psychological Measurement **38**:953-958.

Yee, S. H., D. L. Santavya, and M. G. Barron. 2008. Comparing environmental influences on coral bleaching across and within species using clustered binomial regression. Ecological Modelling **218**:162-174.