Detecting and Interpreting Variable Interactions in Observational Ornithology Data

Daria Sorokina*, Rich Caruana†, Mirek Riedewald‡, Wesley M. Hochachka§, Steve Kelling§
*SCS Carnegie Mellon University, Pittsburgh, PA. daria@cs.cmu.edu
†Microsoft Corporation, Redmond, WA. rcaruana@microsoft.com
‡CCIS Northeastern University, Boston, MA. mirek@ccs.neu.edu
§Cornell Lab of Ornithology, Ithaca, NY. {wmh6, stk2}@cornell.edu

***Abstract*—In this paper we demonstrate a practical approach to interaction detection on real data describing the abundance of different species of birds in the prairies east of the southern Rocky Mountains. This data is very noisy—predictive models built from it perform only slightly better than baseline. Previous approaches for interaction detection, including a recently proposed algorithm based on Additive Groves, often do not work well on such noisy data for a number of reasons. We describe the issues that appear when working with such data sets and suggest solutions to them. In the end, we discuss results of our analysis for several bird species.**

This is a short version of the paper. The full version is located at www.cs.cmu.edu/˜daria/papers.htm

I. Introduction

Much research in machine learning and data mining focuses on building prediction models with the best possible performance. In most cases such models act as *black boxes*: they make good predictions, but do not provide much insight into the decision making process. However, domain scientists often are more interested in performing descriptive analysis and therefore need additional data mining tools.

In this paper we study the process of applying an *interaction detection* algorithm, using a very challenging ecological data set describing the abundance of a variety of bird species. We could not train a high-performing predictive model for this data, but we still were able to detect important biological dependencies. Apart from presenting a detailed application of a general technique to real life data, we also introduce a number of necessary important additions to the earlier procedure to make it useful for noisy data sets.

A. Interactions

Interactions are complex non-additive effects that groups of variables exert on the response of the function. If a variable is not involved in any interactions, its effect can be studied alone and often described by a simple rule. To understand a natural process, it is critical to know which groups of variables are joined in complex effects and thus must be examined together.

A variable interaction is formally defined as follows [1]. Function $F(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \dots, x_n)$, shows no interaction between variables x_i and x_j if it can be expressed as the sum of two functions, $f_{\setminus j}$ and $f_{\setminus i}$, where $f_{\setminus j}$ does not depend on x_j and $f_{\setminus i}$ does not depend on x_i:

$$\begin{aligned} F(\mathbf{x}) &= f_{\setminus j}(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n) \\ &\quad + f_{\setminus i}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \end{aligned} \tag{1}$$

Note that the term statistical interaction describes only the effect of variable values on the response function and should not be confused with any dependencies between the variables themselves, e.g., correlation.

In this paper we extend an interaction detection approach that was recently introduced in [2]. It is based on the comparison of the performance of two models: an *unrestricted* one that is allowed to model a given interaction, and a *restricted* one that is not allowed to model this interaction. If the unrestricted model performs significantly better, then we conclude that modeling the interaction was crucial for good performance and hence there is an interaction between the variables. But if eliminating a specific interaction does not impact model performance then there is no evidence for the presence of an interaction between the tested variables.

As a suitable prediction model for this framework, [2] suggest Additive Groves — an additive-model based ensemble of trees that is good at capturing the additive structure of the function. Additive structure is crucial for modeling absence of interactions and therefore for building a good restricted model. At the same time, the ability to use large trees allows Additive Groves to capture very complex interactions and interactions of small magnitude. For detailed discussions on why Additive Groves fit this framework better than many other models, as well as why this interaction detection approach is more efficient than earlier methods, we refer the reader to the original paper, where this algorithm was introduced [2].

The basic idea of comparing the performance of restricted and unrestricted models appears deceptively simple. [2] provides results on commonly used real and relatively simple synthetic data sets. In this paper we describe problems that emerged during interaction detection analysis on large and noisy real data and suggest how to approach them. In particular our contributions concern the following issues:

1) For a large class of regression data sets, including our ecological data, it is more appropriate to analyze the logarithm-transformed response. However, logarithm

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.84

is a non-linear transformation which can introduce additional interactions not present in the original data. We solve this problem by mimicking the log transformation with a different loss function (Section III).

2) Interaction detection requires feature selection as a preprocessing step, and backward elimination is the most suitable type of feature selection for this purpose. Unfortunately, it is also a computationally expensive algorithm and hence infeasible for large numbers of features. We therefore split the feature selection process into two parts: a fast and less accurate first stage (Section V-A) is followed by backward elimination on the fewer remaining features (Section V-B). We also refine the original algorithm by discarding the assumption that removing a feature never improves performance.
3) It is fairly safe to assume on simple data sets that more complex Additive Grove models perform at least as well as smaller ones, provided the algorithm performs a sufficient number of bagging iterations. But this assumption might not hold on noisy data. Because of this, parameters resulting in the best predictive performance will not necessarily result in the best model for interaction detection. We provide several heuristics that can aid in choosing a model of the right size. (Section VI).
4) Detecting the presence of interactions is only a prerequisite step for studying effects of variables on the response. We briefly describe existing methods for visualizing joint effects of pairs of variables and demonstrate on real examples from our data why they should be used as a visualization aid only, not a tool for detecting interactions by themselves. (Section VII).

In this paper we demonstrate the interaction detection analysis on a specific application: extracting domain knowledge from an ornithological data set and show that this type of analysis can provide useful findings for the field of ecology.

II. Rocky Mountain Bird Observatory

We selected data from one bird-monitoring program run by the Rocky Mountain Bird Observatory (RMBO) [3] for our analysis. Bird species specialized to grassland habitats, including those living in the shortgrass prairies, are some of the fastest and most consistently declining bird species in North America [4]. The monitoring program, called the Section Survey, is one effort to understand the causes and identify management actions that would reverse these declines. The goal is to identify associations between bird abundance and local vegetation, and the objective of identifying management actions that would make habitat more suitable for grassland bird species.

When choosing where to live, birds consider not just local habitat characteristics, but also habitat configuration over larger regions. We therefore include the larger-scale habitat configuration using interpreted satellite imagery from the 2001 U.S. National Land Cover Data [5], which classify habitat across the United States into 21 classes. The resulting data sets contain 700 features and 20000 observations for each bird.

III. Choice of Loss Function

The first fundamental challenge is to select the appropriate performance measure, or *loss function*. A common choice for general regression problems is root mean squared error (RMSE). However, this metric is less appropriate for bird observation data, which are counts. Analysis of point counts is often conducted using the logarithm of the original response function. This is a standard way to treat such data sets in ecology and similar areas.

Unfortunately, working with log-transformed response values has an undesirable side-effect on the interaction detection task. Instead of discovering additive structure in the original function $F(\mathbf{x})$, we would now search for additive structure in function $\log(F(\mathbf{x}))$.

To overcome this problem, instead of changing the response function, we change the loss that our models are trying to minimize. In order to still obtain a simple additive loss and at the same time achieve approximately the same effect as log-transforming the counts, we use the first 3 terms of the Taylor expansion of the squared error of log counts. Since the first 2 terms of this particular expansion are equal to 0, this is equivalent to only using the third term:

$$(log(y+1) - log(F+1))^2 \approx (\frac{1}{y+1}(y-F))^2 \qquad (2)$$

Here y corresponds to the original response, F corresponds to the predicted value. A constant value (usually 1) is added to the counts before taking the logarithm in order to be able to handle zero counts. To derive this approximation, we view the loss function as a function of F with y fixed and take the Taylor expansion at the point $F = y$.

We substitute squared error in RMSE with the obtained weighted squared error $(\frac{y-F}{y+1})^2$ and refer to the new loss as *weighted RMSE*. To make the results comparable across data sets, we use a standardized version of this metric: we divide it by the similarly weighted standard deviation of the response in the data set. The baseline performance for such standardized metric is the performance of the model that predicts the average response value for every data point. This model has a loss of 1 on every data set. Smaller numbers indicate performance better than baseline.

Predictive modeling of RMBO data is very challenging. The improvement over baseline typically is only 2%-5%. For example, for Horned Lark, the bird species about which we could extract the most information, the best performance we could achieve is 0.974 (measured by the loss discussed above with baseline 1.0).

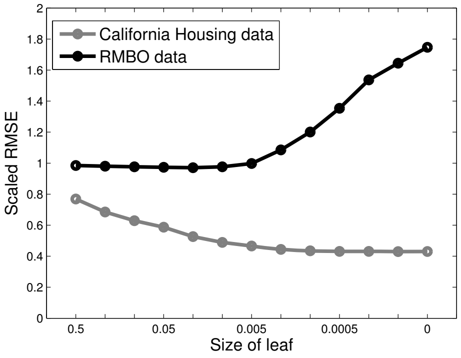

Figure 1. Performance of 100 bagged trees on the "standard" California Housing data set vs. noisy RMBO data.

IV. Tree-Based Models

Our models used for the interaction detection task are ensembles of binary regression trees. Usually these regression trees optimize for RMSE, but we have modified the algorithm for growing trees to use weighted RMSE for selecting splits. We control the size of trees using parameter α, the minimum proportion of train set cases that reach an internal node.

A. Bagged Trees

Bagging [6] is a well-known ensemble method that creates a set of diverse models by sampling from the training set, and then decreases variance by averaging the predictions of these models. Large decision trees are low-bias, high variance models that benefit significantly from bagging. Because of this, often bagging works best with larger trees. However, on noisy data, large trees perform much worse than small trees, even after a large number of bagging iterations. Fig. 1 shows the performance of 100 bagged trees of different sizes on the commonly used California Housing [7] data set, and for the Horned Lark, one of the species in the RMBO data set. The difference in performance patterns of bagging for large and small trees on the two data sets is striking.

B. Additive Groves

Additive Groves, introduced in [8], is a regression ensemble consisting of bagged additive models, where each additive component is a tree. Its size is defined by 2 parameters: α—the minimum proportion of train set cases in a leaf (controls size of a single tree) and N, the number of trees in a single grove. As suggested in [2], for interaction detection we use the "layered" style of training: the second parameter, number of trees, is fixed during training, while the size of trees is gradually changed from very small up to the desired level of complexity.

Early experiments in [8] suggest that Additive Groves are robust to overfitting as long as they are bagged sufficiently many iterations. Unfortunately, similar to the observation above about bagging individual trees, there are some extremely noisy data sets where this is not achieved. This property of the data makes the interaction detection process with the RMBO data more complicated.

V. Feature Selection

Correlations between variables pose a problem for any interaction detection algorithm. For our approach based on model comparison, they can "hide" existing interactions. Suppose we want to test for an interaction between x_i and x_j, and there is another variable x_k that is almost identical to x_j. When we restrict a model on interactions between x_i and x_j, it can use x_k instead of x_j and thus bypass the restriction. Hence even if x_i and x_j interact, we can not discover this unless we remove x_k from the data.

For these reasons we have to eliminate all variables (features) from the data until we are left only with a set of variables such that removing any of them would significantly decrease model performance. We discuss how to do this in the remainder of this section.

A. Fast Feature Evaluation

For data sets with many features a thorough feature selection based on generating different models for different combinations of features is infeasible due to the large number of models that need to be trained. We therefore adopt a two-step approach. In the first step we perform fast but rather crude elimination of the least important features. In the second step we perform a more careful selection from the remaining features.

To preselect a reasonable number of useful features, we use one of the "white-box" feature evaluation techniques that were recently proposed for bagged trees [9]. In particular, we used the "multiple counts" method. This technique ranks attributes based on how often trees in the ensemble use them in their nodes. During the preselection stage we generate several ensembles using trees of different sizes, test their performance on the test set and then chose the best performing one to use for determining feature importance.

Our version of the RMBO data with the NLCD land cover information at different scales has 763 features. In the first step we selected 50 useful features for each species using ensembles of 100 bagged trees. In most cases the best ensembles consisted of relatively small trees, up to ≈ 10 or 20 nodes.

B. Backwards Elimination

To make the first step of feature selection fast enough, we used only bagged trees. During the more finegrained second step, we want to evaluate the performance of the Additive Groves method, as it will be used in the interaction detection process. Hence in this step we build Additive Groves models for the data set with its remaining preselected features for

Algorithm 1 Backwards elimination

```
repeat
  label A: $(\mu, \Delta) = \text{EstimatePerformance}()$
  repeat
    for $f = 1$ to $\#Features$ do
      $\text{Remove}(feature[f])$
      $newPerf = \text{WRMSE}(\text{TrainModel}())$
      if $newPerf - \mu > \Delta$ then
        $\text{Add}(feature[f])$
      if $newPerf - \mu < -\Delta$ then
        goto A (line 2)
  until (No features removed with current $\mu$ and $\Delta$)
until (No features removed on last cycle iteration)

function $(\mu, \Delta) = \text{EstimatePerformance}()$
  for $c = 1$ to $10$ do
    $perf[c] = \text{WRMSE}(\text{trainModel}())$
  $\mu = \text{Mean}(perf[1..10])$
  $\Delta = 3 * \text{StdDev}(perf[1..10])$
```

a variety of parameter combinations. Then we select values for N and α that resulted in the best performance on a validation set. These values are used for all models that are built during the second stage of feature selection.

Recall that in order to be able to run effective interaction detection, we need to be left with a small set of important features. Important here means the following property: if we remove this feature, the performance degrades by more than Δ, where Δ is defined to indicate a significant difference.

To calculate Δ, we estimate the distribution of Additive Groves performances on the data by training several models with different random seeds and evaluating their performances on the validation set. After that the threshold of statistical significance is defined following the common practice in statistics as $\Delta = 3 * \sigma$, where σ is the standard deviation of the estimated distribution. This estimates are used in the backward elimination algorithm. In the beginning all features are present. Then the algorithm tries to remove features one-by-one. If the performance on the validation set does not degrade by at least Δ, the feature is removed permanently. Otherwise the feature is considered important and left in the data. Several passes through the set of remaining features are done until no features can be removed.

As removing features can change the distribution of performances, this distribution needs to be recalculated occasionally. In the first version of the algorithm it happened only when selection could not remove any more features with the current estimates of the distribution.

Note that this algorithm implicitly assumes that removing a feature will either degrade performance or leave it approximately the same. However, this is not always the case for noisy data sets. Trees can mistakenly use "bad" features and benefit when those features are removed. To handle this case, we improved the algorithm as follows: if performance is *better* than the original estimate by Δ, the algorithm recalculates the estimates of the performance distribution. The resulting feature selection procedure is shown in Algorithm 1.

Given the weak predictive performance of models trained on the RMBO data, we were not surprised that feature selection left few important features for most bird species. In the best case (Horned Lark) we had 8 features left, in the worst cases, only 1 or 2.

VI. Interaction Detection

After we are left with only a few important features, we need to choose the right type of Additive Groves model to be used for interaction detection. Our model should represent the function well and at the same time should have sufficient additive structure to allow for restrictions.

In RMBO data the final parameters suitable for interaction detection were very different for different species. Occasionally the search for good parameters required multiple trials with a human in the loop. Our experience can be summarized as follows:

- In order to make the model "additive enough", we need to a large N. From our experience, $N = 8$ usually is a safe value; $N = 6$ will work for most data sets, but smaller values usually hurt the performance of the restricted models.
- Since interaction detection uses the same basic model for the restricted and unrestricted case, the process is fairly robust with respect to choosing Additive Groves parameters. In most cases we can lose $\approx 8\Delta$ of predictive performance without hurting final interaction results.
- It is safer to choose a parameter combination for which Additive Groves slightly underfit (simpler than the best model), rather than overfit, because variance will be higher with the overfit models making the results less reliable.
- Even if there is no clearly optimal point with large N on the grid, we can try points with small N and set the threshold for interaction presence higher than usual when estimating the performance difference.

If different parameters are selected than those used during backward elimination, it is necessary to run another round of backward elimination to make sure that each feature is still important for the new Additive Groves configuration.

Similar to how we define if an attribute is important, an interaction is considered significant if the difference between performance of the unrestricted and restricted models is more than Δ.

VII. Visualization

After we detect the presence of an interaction between two variables x_i and x_j, we want to see how it influences

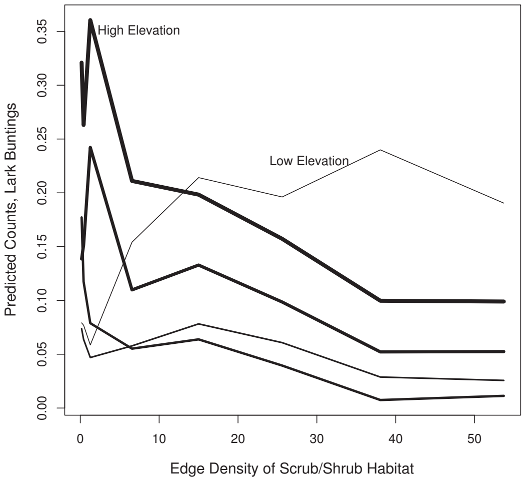

Figure 2. Lark Bunting. Interaction between elevation and density of edges of scrub/shrub vegetation patches

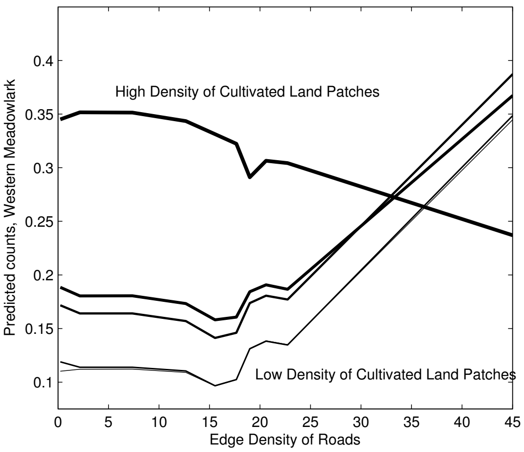

Figure 3. Western Meadowlark. Partial dependence plot, unrestricted model

the response function. In other words, we need to represent the response as a function of x_i and x_j only. After that we can plot the joint effect of two variables as several one-dimensional plots, each of which shows the dependence of the response value on x_i for a fixed value of x_j. Different lines on the plot correspond to different values of x_j. For example, Fig. 2 shows the joint effect of elevation and edge density of shrub patches. Each line corresponds to an effect of shrubs at some fixed level of elevation. Non-parallel regions of the lines correspond to interactions and can provide insight into its nature. In this example we can see that the presence of shrubs shows a positive effect on abundance of Lark Buntings at the lowest elevation, but at higher elevations larger amounts of shrubs patches have the opposite effect and discourage this species.

An efficient method for creating such two-dimensional models, partial dependence plots, was introduced by Friedman [10] as a tool to visualize the effects of a fixed number of variables averaged over the values of all other variables.

It is very important to notice that partial dependence plots by themselves are unreliable for interaction detection, because they depict interactions in the model instead of the data. Hooker [11] demonstrated that potential spurious interactions of arbitrary strength can appear in a partial dependence plot. This happens when some parts of prediction model are unsupported by the data and only emerge because of a presence of a few outliers. A stark example of this emerged during our analysis of RMBO data: Fig. 3 pictures a partial dependence plot for the joint effect of presence of roads and cultivated crop areas on Western Meadowlark abundance generated by an unrestricted model. The plot clearly shows a strong interaction similar to the one we have just seen in Fig. 2. However, there is no such interaction in the data! The restricted model that does not have this interaction has the same predictive performance: our performance comparison method estimated the size of interaction as -0.00009 and the significance threshold as 0.0005, which clearly indicates absence of interaction. [1]

VIII. Results

In this section we present and explain selected results of the application of this interaction detection procedure to the RMBO data. This analysis provided findings about collected data and biological relationships that were previously unknown, and yet are consistent with the general body of ecological knowledge.

The most complex, albeit small, interaction that we identified was for Lark Buntings (*Calamospiza melanocorys*), with elevation and density of scrub/shrub edges simultaneously affecting bunting abundance (Fig. 2). The strength of the interaction is estimated as 0.00037, significance threshold as 0.00032. At the lowest elevation sites, farthest from the base of the Rocky Mountains, Lark Buntings were more abundant in areas with a higher amount of patchily-distributed scrub/shrub vegetation. However, closer to the Rocky Mountains, the presence of scrub/shrub habitat inhibited Lark Buntings from settling. We believe that this result indicates that the habitat classified as "scrub/shrub" represents very different things in different parts of the study region, and that at higher elevations "scrub/shrub" contains plant species or habitat configurations that are unsuitable for Lark Buntings.

The Horned Lark (*Eremophila alpestris*; known as the Shore Lark in Europe) is a species widely distributed across the Northern Hemisphere. It preferentially lives in barren habitat with short and patchy vegetation. The most unexpected interaction we found was related to this preference

[1] When estimating a size of a non-existing interaction, negative numbers insignificantly different from 0 can happen as often as positive numbers. Negative numbers *significantly* different from 0 would indicate some problem, most probably a poor choice of Additive Groves parameters.

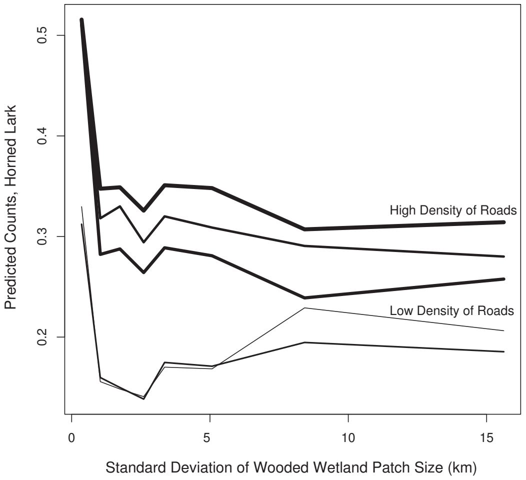

Figure 4. Horned Lark. Interaction between wooded wetlands and density of roads

for barren habitat: abundance of Horned Larks differed across our study area as a function of both the density of roads and the variation in sizes of patches of wooded wetland. Interaction strength is estimated as 0.00163, significance threshold as 0.00085. In the shortgrass prairie region, "wooded wetland" effectively means wooded areas along rivers and these are essentially the only large areas of taller vegetation in the entire region. Fig. 4 shows that there is a sharp drop in abundance of Horned Larks as soon as there is any substantial amount of wooded wetland habitat. Horned Larks do not like wooded habitat. However, the effect of woodland was ameliorated by the presence of roads, with more Horned Larks present, even in areas with higher amounts of forest, when these regions had a higher density of roads: not only the curves corresponding to higher level of road density are above the curves of lower levels, they are also showing slower decrease in birds abundance in dense wetlands. Effectively, the roads create open areas of habitat preferred by Horned Larks. Detecting this interaction has helped us to identify an unexpected impact of human modification of landscape which can be important when assessing implications for Horned Lark from human activity in the future.

Although the original interaction detection technique allows detection of higher-order interactions, we did not have an opportunity to conduct these tests for RMBO data sets. K-way interactions are possible only between those groups of variables that are involved in all possible $K(K-1)/2$ 2-way interactions between each other [12]. Such cliques of pairwise interactions never appeared during our analysis.

IX. Discussion

All interactions detected in the RMBO data were relatively small and could not be reliably detected from partial dependence plots alone. For comparison, most interactions in data sets described in [2] are larger by an order of magnitude or more. This is expected when the data is noisy and difficult to model. The noise obscures interactions that might have been more striking otherwise, because it is impossible to improve performance much over the restricted models. However, as long as these small improvements are significant, they clearly indicate a presence of a real interaction in the data and in the domain.

Techniques introduced here can be easily adapted to other domains and we believe that the experience described in this paper is of direct practical use to any researcher who is interested in applying the general interaction detection method from [2] to a real-world noisy data set.

Acknowledgements. This work was funded by the Leon Levy Foundation and the National Science Foundation (Grant Numbers DBI-0542868, DUE-0734857, IIS-0748626, IIS-0612031, EF-0427914).

References

[1] J. Friedman and B. Popescu, "Predictive learning via rule ensembles," Stanford, Tech. Rep., 2005.

[2] D. Sorokina, R. Caruana, M. Riedewald, and D. Fink, "Detecting Statistical Interactions with Additive Groves," in *Proc. ICML*, 2008.

[3] "Rocky Mountains Bird Observatory," www.rmbo.org.

[4] B. G. Peterjohn and J. R. Sauer, *Ecology and Conservation of Grassland Birds of the Western Hemisphere, 1966-1996.* Cooper Ornithological Society Studies in Avian Biology, 1999.

[5] "2001 National Land Cover Data (NLCD 2001)," www.epa.gov/mrlc/nlcd-2001.html.

[6] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[7] M. Meyer and P. Vlachos, "StatLib. CMU, Dept. of Statistics," http://lib.stat.cmu.edu.

[8] D. Sorokina, R. Caruana, and M. Riedewald, "Additive Groves of Regression Trees," in *Proc. ECML*, 2007.

[9] R. Caruana, M. Elhawary, A. Munson, M. Riedewald, D. Sorokina, D. Fink, W. M. Hochachka, and S. Kelling, "Mining citizen science data to predict prevalence of wild bird species," in *Proc. of KDD'06*.

[10] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 2001.

[11] G. Hooker, "Generalized functional ANOVA diagnostics for high dimensional functions of dependent variables," *Journal of Computational and Graphical Statistics*, 2007.

[12] G. Hooker, "Discovering ANOVA structure in black box functions," *Proc. ACM SIGKDD*, 2004.