Detecting and Interpreting Variable Interactions in Observational Ornithology Data

Daria Sorokina
SCS Carnegie Mellon University
daria@cs.cmu.edu

Rich Caruana
Microsoft Corporation
rcaruana@microsoft.com

Mirek Riedewald
CCIS Northeastern University
mirek@ccs.neu.edu

Wesley M. Hochachka, Steve Kelling
Cornell Lab of Ornithology
{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 this data perform only slightly better than baseline. Previous approaches for interaction detection, including recently proposed algorithm based on Additive Groves, might not work ideally 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. We further demonstrate that with our improvements to the interaction detection algorithm it is possible to detect interactions between important features and the response function, even when the data is this noisy. In the end, we show and interpret the results of our analysis for several bird species.

1. Introduction

Much research in machine learning and data mining focusses 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 they need additional data mining algorithms to answer questions like: What effects do important features have on the response variable? Which features are involved in complex effects — non-additive interactions and therefore must be studied only together with other features? How can we visualize and interpret interactions?

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.

1.1. 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, such as "the number of birds increases linearly with the number of trees". Often natural processes are more complex, e.g., "in the south, there are more birds of species X in winter than in summer; but in the north, there are more of them in summer than in winter (because birds migrate)". Here we cannot describe the seasonal effect without taking into account the values of the latitude variable. This example illustrates an interaction between seasonal effects and latitude. To understand a natural process, it is critical to know which groups of variables are joined in such complex effects and thus must be examined together.

A variable interaction is formally defined as follows [8]. Function $F(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, 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:

$$F(\mathbf{x}) = f_{\setminus j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) + f_{\setminus i}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \quad (1)$$

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.

1.2. Interaction Detection

Although interaction detection is a well-known problem in statistics and a number of techniques have been around for a long time, most established algorithms are not designed to work well with large noisy data sets. Most algorithms have one of the two standard drawbacks: they either assume that interactions take on some very simple mathematical form (e.g., multiplication term), or assume independent distribution of values for interacting variables. Neither assumption holds for large real-life data. In this paper we extend an interaction detection approach that was recently introduced in [16]. It is based on 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 much 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 the model performance - restricted model performs as well as the unrestricted one - then there is no indication of the presence of an interaction between the tested variables in this data.

As a suitable prediction model for this framework, [16] 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 modelling absence of interactions and therefore for building a good restricted model (in restricted Additive Groves every tree is not allowed to use one of the variables from the interaction in question). At the same time, 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 [16].

The basic idea of comparing the performance of restricted and unrestricted models appears deceptively easy. [16] provides results on relatively simple synthetic and standard testbed data sets. In this paper we describe problems that emerged during interaction detection analysis on large and noisy application 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 convenient to analyze log of the response instead of the original response functions. However, logarithm is a non-linear transformation which can add extra interactions not present in the original data. We solve this problem by mimicking log transformation with a different loss function (Section 3).

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 computationaly heavy algorithm and is infeasible for large numbers of features. We therefore split feature selection process on two parts: fast and less accurate first stage (Section 5.1) is followed by backward elimination on a few preselected features (Section 5.2). We also refine the original algorithm by discarding an assumption that removing a feature never improves performance.

3. While it is safe to assume on simple data sets that more complex Additive Groves models perform at least as good as small ones if you bag long enough, this assumption might be heavily broken 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 6).

4. Detecting the presence of interactions alone is only a prerequisite step for studying effects of features on the response. We briefly describe existing methods to visualize 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 7).

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. However, although all techniques discussed here were motivated by a specific domain, they are not application dependent and can be used for many other domains with large and noisy data.

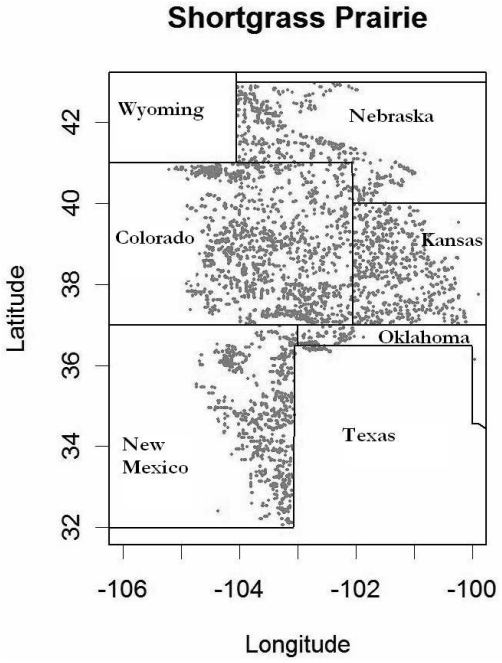

Figure 1. Observation sites.

2. Rocky Mountain Bird Observatory Data

The data used in our analyses come from the data warehouse of the Avian Knowledge Network (AKN) [2], an international collaboration of government and non-government institutions focused on understanding the patterns of distribution and dynamics of bird populations across the Western Hemisphere. This collaboration is creating the framework for gathering and storing existing and new bird-monitoring data from all available sources. It organizes these resources in such a way as to enhance application development, archiving, visualization and exploration, and makes these data generally available. The AKN also creates information products that use its data resources to produce visualizations such as maps, graphs, and tables, as well as scientific and technical analyses.

We selected data from one bird-monitoring program run by the Rocky Mountain Bird Observatory (RMBO) [3] for this analysis. The monitoring program, called the Section Survey, has collected counts of birds of different species observed at over 10,000 sites across a large part of the region known as the shortgrass prairie (Fig. 1). This is an arid zone in the rain shadow of the Rocky Mountains, characterized by short and sparse vegetation. 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 [13]. The Section Survey monitoring scheme is one effort to understand the causes and identify management actions that would reverse these declines. The Section Survey collects data on both abundances of birds (using a distance-sampling protocol [5]), as well as local vegetation at the survey sites. The goal is to identify associations between bird abundance and local vegetation, and the objective of identifying management actions (such as livestock grazing regimes) that would make habitat more suitable for grassland bird species. For our analyses, we used the numbers of detected birds within 100 meters of the observer in a 3-minute period as the response variable, with a different response for each species identified on the survey.

When choosing where to live, birds consider not just local habitat characteristics — such as those measured by the Section Survey protocol — but also habitat configuration over larger regions [14, 17]. We include the larger-scale habitat configuration using interpreted satellite imagery from the 2001 U.S. National Land Cover Data [1], which classify habitat across the United States into 21 classes. Various measures of habitat configuration were calculated from these aggregations using the program FRAGSTATS [11]. These habitat configuration metrics, combined with the observed bird count response variable, are the data we analyse. The resulting data sets contain 700 features and 20000 observations for each bird.

3. Choice of Loss Function

Our technique for finding variable interactions is based on the comparison of the performance of models. To test for an interaction between variables x_i and x_j, we train two models. The *restricted* model is not allowed to model the interaction between the variables. For the unrestricted model, there is no limitation in terms of which interactions can be modeled. If appropriate models are used, then the difference in performance between restricted and unrestricted model indicates the strength of the interaction between x_i and x_j.

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. RMSE penalizes *absolute* deviation from the true response value. For example, predicting 25 birds instead of 20 will be penalized as heavily as predicting 5 birds when there were none. This is not desirable because the estimation error for the smaller response value is much more serious. For this reason 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 the different function $\log(F(\mathbf{x}))$. Since $\log(f_1)+\log(f_2) = \log(f_1 \cdot f_2)$ for any response values f_1, f_2, we would in fact model multiplicative structure, instead of additive, in the original function F. Detecting complex effects in multiplicative structure might be of interest as well, but if we want to find and understand non-additive interactions, working with log counts is not appropriate.

What loss function should be used to penalize errors for low counts more? 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:

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

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 allow 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 similarly weighted standard deviation of response in the data set. The convenient baseline performance for such standardized metric is the performance of the model that predicts average response value for every data point. It is equal to 1 on every data set. Smaller numbers indicate performance better that the 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).

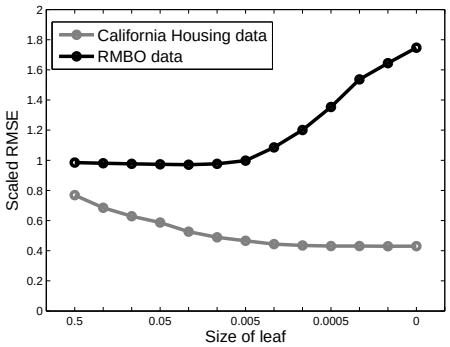

Figure 2. Performance of 100 bagged trees on "standard" California Housing data set vs. noisy RMBO data. Small RMSE means better performance.

4. 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 as we discussed in section 3, we use weighted RMSE instead. Hence we have modified the algorithm for growing trees to use weighted RMSE for selecting splits. We control size of trees using parameter α, the minimum proportion of train set cases that reach an internal node.

4.1. Bagged Trees

Bagging [4] 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. 2 shows the performance of 100 bagged trees of different sizes on the commonly used California Housing [12] data set, and for the Horned Lark, one of the species in the RMBO data set. The difference in performance pattern of bagging for large and small trees on the two data sets is striking. This poor performance of bagging large trees on large noisy data sets can be explained as follows. Bagging can never remove variance completely, because it draws versions of the data again and again from the original training data set. The different training samples inevitably overlap and produce trees that make partially the same predictions. When noise is present, large trees can put noisy data points in separate leaves and therefore make large errors. Due to overlapping train sets they might use the same noisy

Algorithm 1 Additive Groves: layered training of a single grove

function Layered(α,N,$TrainSet\{\mathbf{x}, y\}$)
  $\alpha_0 = 0.5, \alpha_1 = 0.2, \alpha_2 = 0.1, \ldots, \alpha_{\max} = \alpha$
  for $j = 0$ **to** max **do**
    repeat
      for $i = 1$ **to** N **do**
        newTrainSet $= \{\mathbf{x}, y - \Sigma_{k \neq i} \text{Tree}_k(\mathbf{x})\}$
        $\text{Tree}_i = \text{TrainTree}(\alpha_j, \text{newTrainSet})$
    until (change from the last iteration is small)

data points and make the same large errors, and therefore this problem will at least partially stay even after averaging. On cleaner standard data sets this effect is not very visible, but on real noisy data the situation can be as bad as illustrated in Fig. 2.

4.2. Additive Groves

Additive Groves, introduced in [15], 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 (controls size of a single tree) and N, the number of trees in a single grove. As suggested in [16], 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 desired level of complexity. (See Algorithm 1.)

Early experiments in [15] suggest that Additive Groves are robust to overfitting as long as they are bagged sufficiently many iterations. This is the case as long as the bagging process succeeds in removing most variance. Unfortunately, similar to the observation above about bagging individual trees, there are some extremly noisy data sets where this is not achieved. Fig. 3 shows a contour plot of how performance of Additive Groves depends on values of α and N on one of our RMBO data sets. Performance is measured using the weighted RMSE loss described in a previous section, therefore smaller numbers correspond to better performance. We can see that the best performance is reached for comparably small models, and then rapidly decreases when the models become more complex. This property of the data makes the interaction detection process with the RMBO data more complicated.

5. 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.

In general, for detecting an interaction involving a variable x_i, removing x_i from the data set should result in a significant drop in performance. In fact, removing x_i is a stronger limitation for the model than restricting it on an interaction with x_i. If performance does not drop when we completely remove x_i, we cannot expect it to drop when restricting on an interaction with x_i.

For these reasons we have to eliminate all variables (features) from the data until we are left 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.

5.1. Fast Feature Evaluation

For data sets with high or even medium number of 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 [6]. 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. The larger the subset of the train set in the node, the larger the score of the splitting attribute in that node. Experiments in [6] showed that multiple counts, the simplest and fastest of those metrics, produces results of similar quality compared to more expensive methods like sensitivity analysis.

As we mentioned above, using large trees hurts performance for the noisy data. Hence on 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. At the first step we selected 50 useful features for each

Algorithm 2 Backwards elimination

```
repeat
  label A: (μ, Δ) = EstimatePerformance()
  repeat
    for f = 1 to #Features do
      Remove(feature[f])
      newPerf = WRMSE(TrainModel())
      if newPerf − μ > Δ then
        Add(feature[f])
      if newPerf − μ < −Δ then
        goto A (line 2)
  until (No features removed with current μ and Δ)
until (No features removed on last cycle iteration)

function (μ, Δ) = EstimatePerformance()
  for c = 1 to 10 do
    perf[c] = WRMSE(trainModel())
  μ = Mean(perf[1..10])
  Δ = 3 ∗ StdDev(perf[1..10])
```

species using ensembles of 100 of bagged trees each. In most cases the best ensembles consisted of relatively small trees, up to ≈ 10 or 20 nodes.

5.2. Backwards Elimination

To make the first step of feature selection fast enough, we used only bagged trees. On the more fine-grained second step, we want to evaluate the performance of the Additive Groves method, as it will be used in the interaction detection process. At this step we do not know anything about how to set the parameters α and N. We therefore build Additive Groves models for the data set with its remaining preselected features with a variety of parameter combinations. Evaluating this set of models on the validation set, we select values for N and α that resulted in the best performance. 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.

The first version of a suitable backward elimination algorithm that achieves this goal was mentioned in [16]. Here we introduce it in more details.

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 validation set does not degrade by at least Δ, the feature is removed permanently. If it does, 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 coudn't remove any more features with the current estimates of the distribution.

Note that this algorithm implicitly assumes that removing a feature will either degrade the 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 — we have seen cases of significant improvement in performance during the second step of feature selection on RMBO data. To handle this case, we improved the algorithm as follows: if performance is *better* than the original estimate by Δ, the algorithm must recalculate the estimates of the performance distribution. The resulting feature selection procedure is shown in Algorithm 2.

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.

6. 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.

It is easy to meet this requirements when one can increase complexity of a model without harming performance. However, for the noisy RMBO data we often observed that the best performance is achieved by a rather small model with little additive structure, and some compromise is required when choosing parameters for interaction detection. Fig. 3 shows the performance of the model for Horned Lark after feature selection. The best performance is achieved for $N = 2$

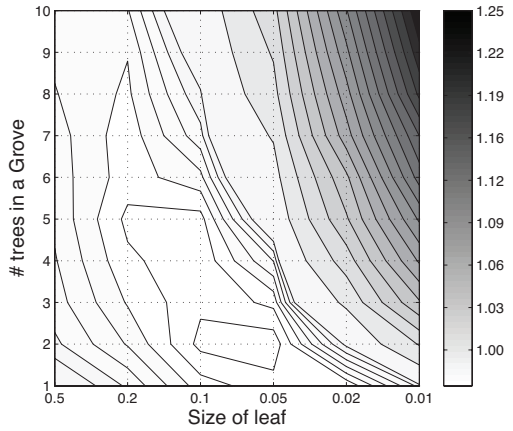

Figure 3. Weighted RMSE of 100 bagged Additive Groves on RMBO data (Horned Lark)

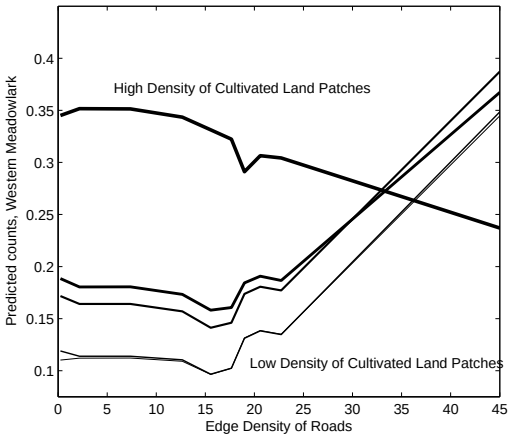

Figure 4. Western Meadowlark. Partial dependence plot, unrestricted model

and trees of moderate size; increasing complexity on any of these parameters degrades performance.

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. Fortunately, one needs to do this only once for each response function and the selected parameters remain the same for the rest of the interaction detection process. Our experience can be summarized as follows:

- In order to make the model additive enough, we need to choose 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. Even with parameter values that result in suboptimal performance, we can still discover interactions. 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 then usual when estimating the performance difference.

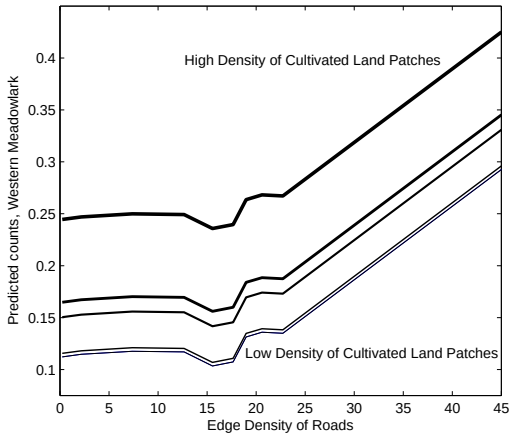

Figure 5. Western Meadowlark. Partial dependence plot, restricted model

For example, we selected $N = 6$ and $\alpha = 0.2$ for Horned Lark based on the countour plot in Fig. 3 and the rules described above.

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.

Similarly to how we define if an attribute is important, interaction is considered significant if the difference between performance of the unrestricted and restricted models is more than Δ. Notice that values of Δ are different for different data sets and/or different model parameters and often indicate the amount of variance in the model.

7. Visualization

After we detected the presence of an interaction between two variables x_i and x_j, we want to see how it influences 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. 6 shows the joint effect of elevation and edge density of shrub patches. Each line correspond to an effect of shrubs at some fixed level of elevation. Non-parallel regions of the lines correspond to interactions and can provide us with 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 [7] 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 [10] 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. Here is a stark example that emerged during our analysis of RMBO data: Fig. 4 pictures a partial dependence plot for joint effect of presence of roads and cultivated crops 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 on Fig. 6. 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] Fig. 5 shows a similar plot produced by a restricted model. We can see that the effect of roads corresponding to the highest

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

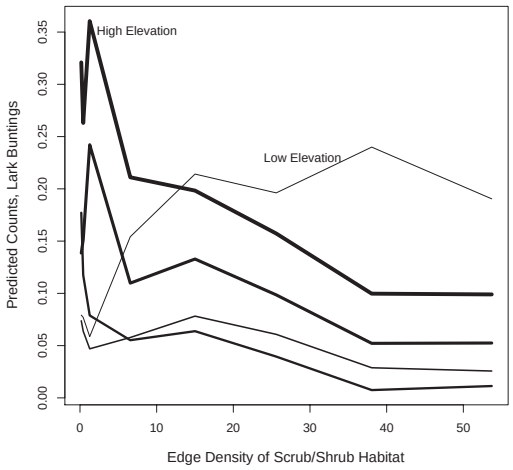

Figure 6. Lark Bunting. Interaction between elevation and density of edges of scrub/shrub vegetation patches

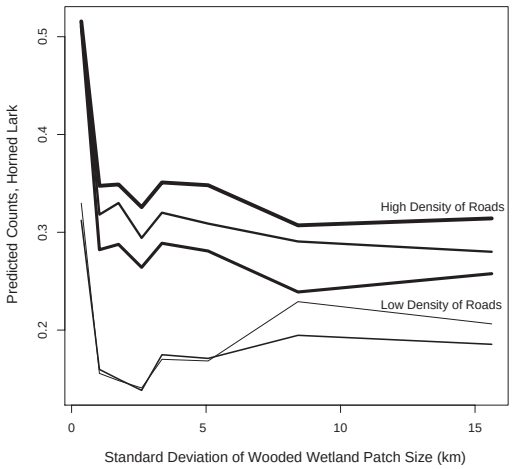

Figure 7. Horned Lark. Interaction between wooded wetlands and density of roads

level of density of cultivated land patches is now very different from the previous picture. However, the performance of the model is the same. The explanation is that there are very few points with high level of cultivated land density in the data, clearly not enough to estimate a real effect. The interaction that we could see on Fig. 4 is a mere random fluctuation. This example illustrates that partial dependence plots should be used for visualization only, when we already have confirmed the presence of interaction in the data by comparing restricted and unrestricted models. Another reason why it is important is because interactions detected in noisy data often are very small and not always visibly distinguishable from spurious irregularities on partial dependence curves.

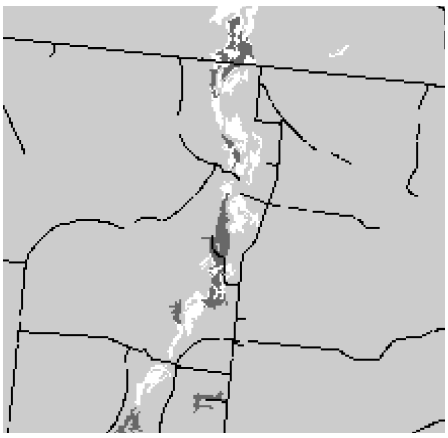

Figure 8. Habitat of Horned Larks. NLCD layers: black - roads; dark - wooded wetlands; light - grassland; white - water.

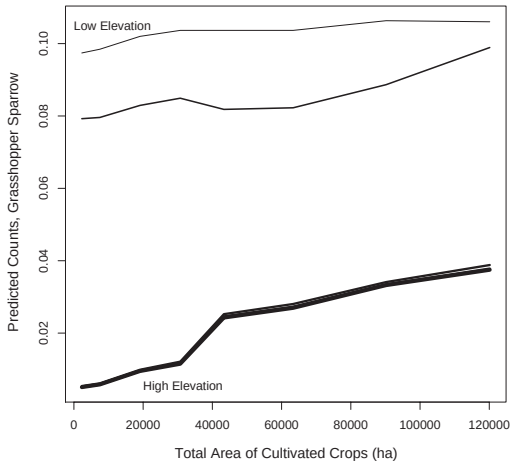

Figure 9. Grasshopper Sparrow. Interaction between elevation and cultivated crops

8. 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, allbeit small, interaction that we identified was for Lark Buntings (*Calamospiza melanocorys*), with elevation and density of scrub/shrub edges simultaneously affecting bunting abundance (Fig. 6). Size of 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. In other words, we believe that this unexpected finding tells us something about our predictors rather than about the birds we are studying, that the scrub/shrub habitat class is more heterogeneous than the classification would at first lead us to suspect.

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 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. Size of interaction 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. Greater variation in size of wooded patches is related to a broader distribution of trees in the overall region and a greater fragmentation of the open habitat that the larks prefer. Fig. 7 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. Fig. 8 shows a representative example of the distribution of habitat types in an area of lark habitat (grassland) in which wooded wetlands and roads are also present in relatively high densities. 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.

The Grasshopper Sparrow (*Ammodramus savannarum*) is a species that lives in moderately lush grass-

land habitat (by the standards of the shortgrass prairie region), an effect that we believe is indicated by the sharp drop in abundance of this sparrow at higher elevations: Drier sites are closer to the rain shadow of the Rocky Mountains. Fig. 9 shows a threshold-like effect; note that three separate partial-dependence prediction lines are essentially overlapping at higher elevations. However, the elevation effect was eased by the presence of cultivated crops at higher-elevation sites within the grasslands. Size of this interaction was estimated as 0.00223, significance threshold as 0.00093. We suspect that this unanticipated finding results from the presence of artificial water sources, irrigating the cropland, creating habitat that was more suitable for Grasshopper Sparrows. Again, interaction detection provided us with evidence that human modification of landscapes affected their suitability to birds, allowing Grasshopper Sparrows to live in areas that would be unsuitable for them under natural conditions.

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 [9]. Such cliques of pairwise interactions never appeared during our analysis.

9. Discussion

All interactions detected in 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 [16] 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. In particular, we have two general observations about the interactions that were detected by our analyses. First, aside from the interaction identified for Lark Buntings (Fig. 6), the effect of variation in one feature was only moderately altered by variation in the second feature in the interactions, as seen by the nearly parallel natures of the lines in the figures. Because of significant difference between restricted and unrestricted models we know that this small difference between line shapes did not happen merely due to a random fluctuation. Our second observation is that most of the habitat types involved in the interactions were relatively uncommon in the shortgrass prairie region. For example, 99% of the areas around individual sites were composed of less than 4% open water and less than 3% wooded wetlands. Two other habitat types were highly patchily distributed, with a median percentage of less than 2.5% (but a maximum in excess of 80%), and a median amount of cultivated crops on less than 18%, although some local areas had roughly 80% of their areas in cultivated crops. Thus, the interactions that we detected are describing biological phenomena that are occurring around only a small proportion of the sites. Both observations highlight the extreme sensitivity of this interaction detection algorithm, which is especially important for the domains with noisy data.

10. Conclusions

We have applied the process of interaction detection to extremely noisy ecological data. We discussed several potential problems that can arise with this kind of real data, proposed possible solutions and presented the real results of applying this analysis to the data. 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 [16] to a real-world noisy dataset.

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

References

[1] 2001 National Land Cover Data (NLCD 2001). www.epa.gov/mrlc/nlcd-2001.html.

[2] Avian Knowledge Network. http://www.avianknowledge.net.

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

[4] L. Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[5] S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. *Introduction to Distance Sampling*. Oxford University Press, 2001.

[6] 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*.

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

[8] J. Friedman and B. Popescu. Predictive learning via rule ensembles. Technical report, Stanford, 2005.
[9] G. Hooker. Discovering ANOVA structure in black box functions. In *Proc. ACM SIGKDD*, 2004.
[10] G. Hooker. Generalized functional ANOVA diagnostics for high dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 2007.
[11] K. McGarigal, S. A. Cushman, M. C. Neel, and E. Ene. FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps.
[12] M. Meyer and P. Vlachos. StatLib. CMU, Dept. of Statistics. http://lib.stat.cmu.edu.
[13] 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.
[14] S. K. Robinson, F. R. Thompson, T. M. Donovan, D. R. Whitehead, and J. Faaborg. Regional forest fragmentation and the nesting success of migratory birds. *Science*, 267:1987–1990, 1995.
[15] D. Sorokina, R. Caruana, and M. Riedewald. Additive groves of regression trees. In *Proc. ECML*, 2007.
[16] D. Sorokina, R. Caruana, M. Riedewald, and D. Fink. Detecting statistical interactions with additive groves. In *Proc. ICML*, 2008.
[17] M. Winter, D. H. Johnson, J. A. Shaffer, T. M. Donovan, and W. D. Svedarsky. Patch size and landscape effects on density and nesting succesws of grassland birds. *Journal of Wildlife Management*, 2006.