

Finding Interesting Patterns through Analysis of Complex Prediction Models

Mirek Riedewald, Daniel Fink

1 Introduction

In many scientific disciplines, researchers collect large amounts of observational data. Our running example are ornithologists who have accumulated a large database of tens of millions of bird sightings. These sightings are all described by time, location, and a list of species seen or not seen. The former two are *predictor* attributes (also called *features* or *attributes*)—they describe properties of the observation event. The observed number of individuals of a species is the *response*, i.e., the measure of interest. There are thousands of additional predictor attributes from external sources, describing the habitat, climate, human population information and other features, which can be linked to an observation record based on time and location.

Ornithologists want to find out, which predictors are the most important ones in determining the response. They also would like to learn about major trends involving one or more predictors, e.g., the decline of bird populations in certain regions or habitats. Another goal is to find interactions between predictors, i.e., if the effect of one predictor on the response is dependent on the values of another predictor.

We encountered similar challenges in other scientific fields like high-energy physics and ecology, and it is easy to see that our approach will apply to many other disciplines that deal with observational data.

2 Basic Approach

Observational data is very challenging to analyze. There are missing or incorrect values, observations have inherent biases, and the number of predictors is very large. To address these challenges, we take the following general approach:

1. Clean the data, e.g., by removing noisy and unreliable predictors or by correcting data entries.
2. Add reliable predictors, e.g., from GIS and census data sets.
3. Train a data mining model on a large fraction of the available observation records and evaluate its predictive performance on an independent test set (the remaining data records that were not used for training).
4. Probe the model with a series of carefully selected predictor-value vectors to generate various one- and two-dimensional graphical model summaries.
5. Enable scientists to find the “most interesting” summaries efficiently. Scientists specify *what* they find interesting, they do not need to specify *how* to find it quickly.
6. Based on discovered interesting summaries, scientists can formulate new hypotheses and later test these hypotheses.

In this article, we focus on step 5—finding the most interesting model summaries. The following example illustrates what we mean by ‘model summary’ and ‘interesting’.

Figure 1 shows one-dimensional summaries of a model that was trained on Project FeederWatch observational records in Bird Conservation Region (BCR) 30. (See Figure 2 for a map of BCRs in North America.) We used 92,514 observations in BCR 30 and each record has 197 predictor attributes. About 2/3 of the records were used for training, the remaining 1/3 for testing the model. We trained a bagged decision tree model to predict the probability of seeing House Finches, given a certain location, time, observer effort, habitat features, human population characteristics, and climate features.

The trained model has excellent predictive performance, but it is very complex and by itself does not reveal anything intelligible to a scientist. Essentially the model is a function that for a 197-dimensional vector

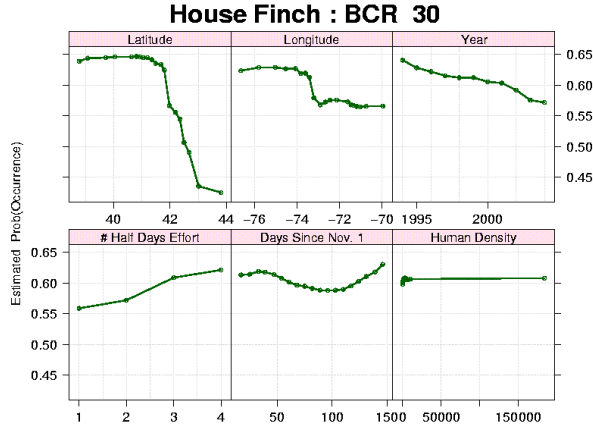


Figure 1: One-dimensional summaries of a complex model learned for House Finch occurrence in BCR 30

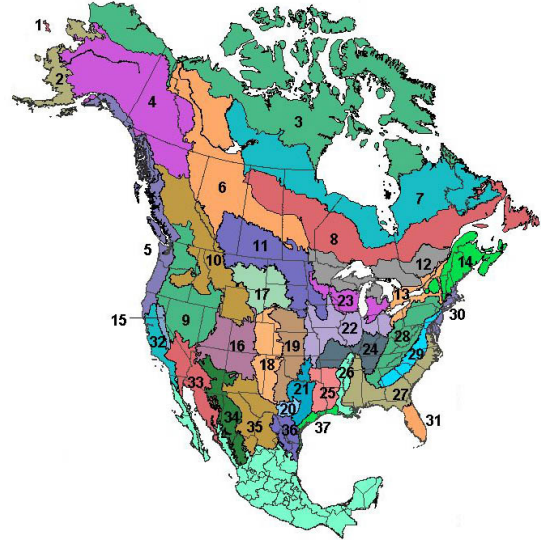


Figure 2: Bird Conservation Regions (BCRs)

of input values describing location, time, observer effort, habitat features, human population characteristics, and climate features, will return a good prediction of the probability of seeing a House Finch for this given input record.

To make the information captured by the 197-dimensional function digestible for a scientist, we can visualize the dependence of House Finch occurrence on a single predictor by appropriately marginalizing over the other 196 predictors. Figure 1 shows such summaries for selected predictors. These summaries do not reveal everything the model has learnt, but a graph with significant differences in estimated probability of occurrence has the potential to signal real biological effects. For example, the 'Year' plot indicates a strong decline of House Finches in BCR 30 since 1995. The '#Half Days Effort' plot brings out the expected relationship between increasing observer effort and higher probability of seeing House Finches. On the other hand, the flat appearance of the 'Human Density' plot does not support the conclusion that human population density has no effect on House Finch occurrence—it might be part of an interaction with other predictors, which together average out human density effects when summarizing the 197-dimensional space in a 1-dimensional plot.

Which of the plots in Figure 1 is most interesting to a scientist depends on individual preferences. Plots showing a large difference between the highest and lowest occurrence probability are often interesting, because they indicate a strong functional relationship between the predictor and the response. From that point of view, the 'Latitude', 'Longitude', and 'Year' plots are much more interesting than the 'Human Density' plot. To another scientist, plots showing a strong trend, e.g., measured as slope of regression line, might be most interesting.

Yet another scientist might be most interested if plots look very different for different geographical regions or for different habitat classes. Figure 3 shows how the observed occurrence of the Purple Finch changes from year to year. The different lines correspond to different BCRs. Two facts could be considered interesting about this graph. First, BCRs 12 and 14 show a clear up-down pattern between even and odd years. Second, there is an apparent difference between the trend in BCRs 12 and 14 versus BCRs 22 and 24. This provides valuable information for a scientist, who can now start looking for reasons why different geographical regions show different patterns.

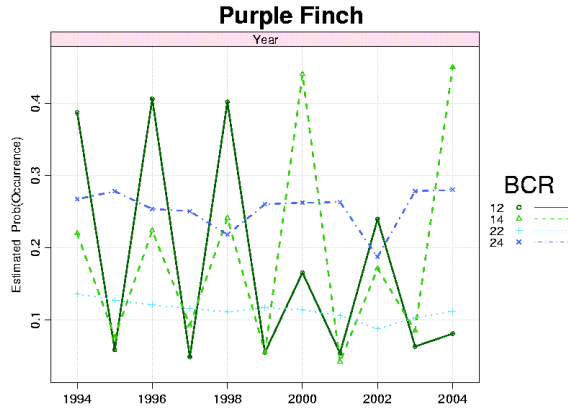


Figure 3: Occurrence of Purple Finch in different BCRs over the years

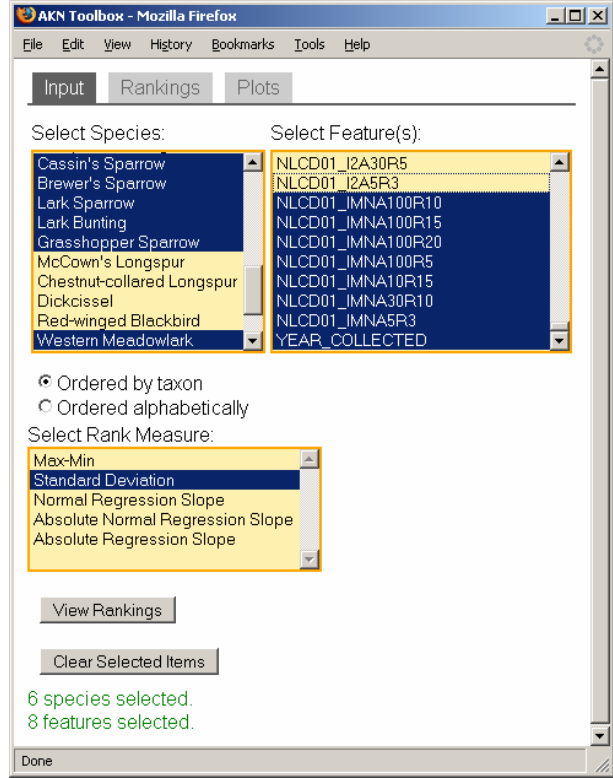


Figure 4: Model summary search engine start page

3 A Search Engine for Model Summaries

Our goal is to build a data mining tool that makes finding interesting model summaries almost as simple and fast as searching the Web. Like a keyword query in a traditional search engine, users will be able to specify their preferences for what they find interesting in an intuitive way. Then the system will respond with those model summaries that match the query best. We have set up a working prototype, which can already be used for analysis. However, it is limited in that it only lets the user choose from a set of pre-defined interestingness ranking measures. In future versions, users will be able to enter personalized ranking measures as well.

Figure 4 shows the interface of the current analysis tool for the RMBO dataset. This is presented to the scientist when she points her Web browser to the appropriate URL. The scientist can then select any set of species and features of interest. In addition, she can select one of the predefined ranking measures. For this example, standard deviation of the summary function was selected, indicating that the scientist is most interested in summaries that are not “flat lines”.

After clicking **View Rankings**, a list of all selected model summaries is shown, ordered based on the selected interestingness ranking measure. Figure 5 shows the result for the example. Notice that there is a model summary for each species-feature combination that was selected in the previous step. The number in each line indicates the interestingness score of the summary for the selected ranking measure. The scientist can now select any set of these summaries and click **View Plots** to see the actual plots. In this case nine summaries were selected and the corresponding plots are shown in Figure 6.

At any point in time, the scientist can go back to the Input tab and modify the selection of species, features, and ranking measure. She can also repeatedly select different sets of summaries from the ranked list under the Rankings tab to generate different groups of plots together.

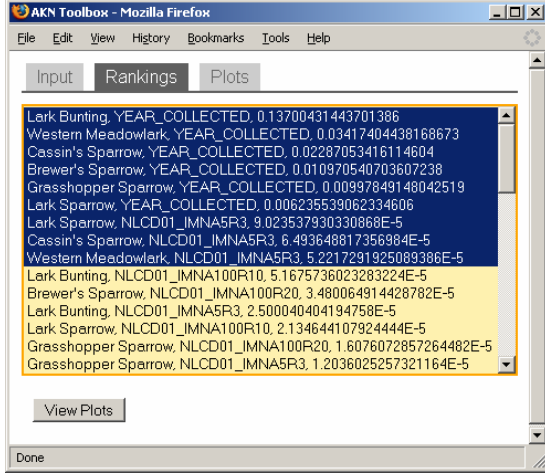


Figure 5: Ranked list of model summaries

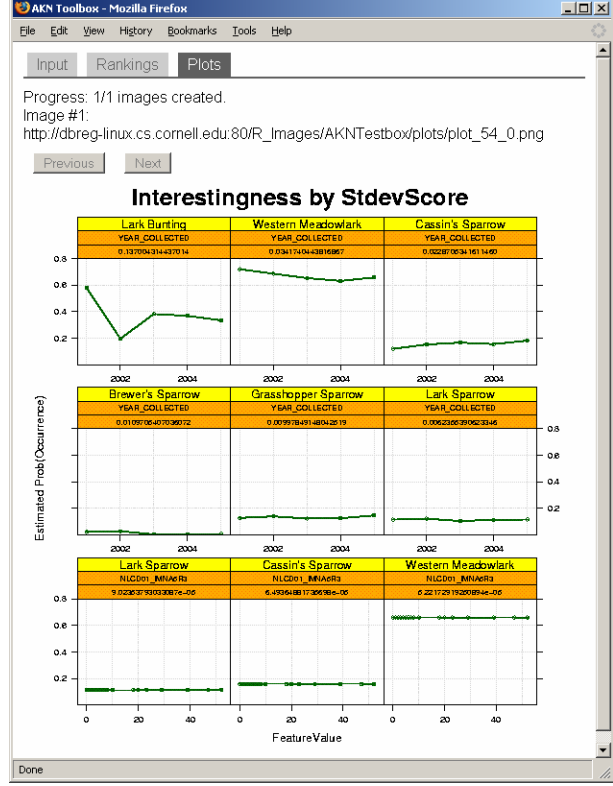


Figure 6: Plots of selected model summaries

4 Measures of Interestingness for Model Summaries

In this section we discuss possible measures of interestingness to rank model summaries. We first define model summaries more formally:

Definition 1. Let $D^{(1)}, D^{(2)}, \dots, D^{(d)}$ be the domains of d predictor variables and D^y be the domain of the response. A model M is a function that maps a d -dimensional predictor-value vector to a response value, i.e., $M : D^{(1)} \times D^{(2)} \times \dots \times D^{(d)} \rightarrow D^y$. A k -dimensional summary of M is a function $S(M, j_1, j_2, \dots, j_k) : D^{(j_1)} \times D^{(j_2)} \times \dots \times D^{(j_k)} \rightarrow D^y$ for $\{j_1, j_2, \dots, j_k\} \subseteq \{1, 2, \dots, d\}$, $|\{j_1, j_2, \dots, j_k\}| = k$, and $k \leq d$.

Stated differently, model M for a given input vector $\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(d)})$, where $x^{(j)} \in D^{(j)}$ for all j , returns the corresponding response value $M(\mathbf{x})$. A model summary similarly maps a k -dimensional input vector to a corresponding response value. In our running example, our model for bird occurrence is a function M from a vector of (location, time, observer effort, habitat features, human population characteristics, climate features) to the corresponding probability of observing a certain bird species. The 'Year' summary in Figure 1 is a function $S(M, \text{Year})$ that maps values of the year predictor to the corresponding average occurrence probability for that year.

In practice a summary will usually be given as a set of pairs of predictor and response values. More formally, the one-dimensional summary for the j -th predictor will be given as $S(M, j) = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} = \{(x_i, y_i)\}_{i=1}^n$, where all x_i are values drawn from domain $D^{(j)}$, the domain of the j -th predictor, and all y_i are from domain D^y . In Figure 6, the large dots indicate these (x, y) pairs that make up the plot. For example, summary $S(M, \text{YEAR_COLLECTED})$ for Lark Bunting consists of the following pairs of year and occurrence probability values: (2001, 0.58), (2002, 0.20), (2003, 0.38),

(2004, 0.37), (2005, 0.33).

Similarly, a two-dimensional summary corresponds to a surface. It is defined by a set of triplets of the form $(x_i^{(j1)}, x_i^{(j2)}, y_i)$.

4.1 Single-Summary Measures

A single-summary measure assigns a *score* to each model summary independently. For instance, the number in each row in Figure 5 shows the score of the corresponding model summary for the standard deviation ranking measure. In the remainder of this subsection, we will enumerate several possible single-summary measures, mostly for one-dimensional summaries. As we discussed above, we assume a summary is given as a set of points $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. We use $\bar{x} = (\sum_{i=1}^n x_i)/n$ and $\bar{y} = (\sum_{i=1}^n y_i)/n$ to denote the mean of the given x and y values, respectively.

Max-Min

- Intuition: difference between highest and lowest response value
- Definition: $\max\{y_1, \dots, y_n\} - \min\{y_1, \dots, y_n\}$
- Motivation: indicates absolute strength of effect of predictor variable on response

Standard-Deviation

- Intuition: common measure of variability of the response values
- Definition: $\sqrt{\frac{1}{N} \sum_{i=1}^n (y_i - \bar{y})^2}$
- Motivation: indicates strength of effect of predictor variable on response

Regression-Slope

- Intuition: absolute value of slope of regression line for linear regression model of given points
- Definition: $\left| \frac{(\sum_{i=1}^n x_i y_i)/n - \bar{x}\bar{y}}{(\sum_{i=1}^n x_i^2)/n - \bar{x}^2} \right|$.
- Motivation: find strongest trends
- Variations: positive slope (to find strongest increasing trends), negative slope (to find strongest decreasing trends), normalized regression slope (multiply slope by $(x_n - x_1)$, i.e., the range of x values, to eliminate effect of range size for different predictors)

Surprisingness

- Intuition: difference between expected summary (e.g., a hypothesis based on previous knowledge) and actual summary
- “Definition”: There are different ways of measuring the difference between two summaries. We can treat each summary $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ as a vector (y_1, y_2, \dots, y_n) in an n -dimensional space. Then we can use Euclidean distance, or any other appropriate distance measure, between the hypothesis vector and the model summary vector.
- Motivation: When there exists prior knowledge about how a summary like the annual abundance trend for a certain species should look like, then it is most interesting to find summaries that do not conform to this hypothesis.

Many more such single-summary measures could be used. Many of them will naturally generalize to higher-dimensional summaries.

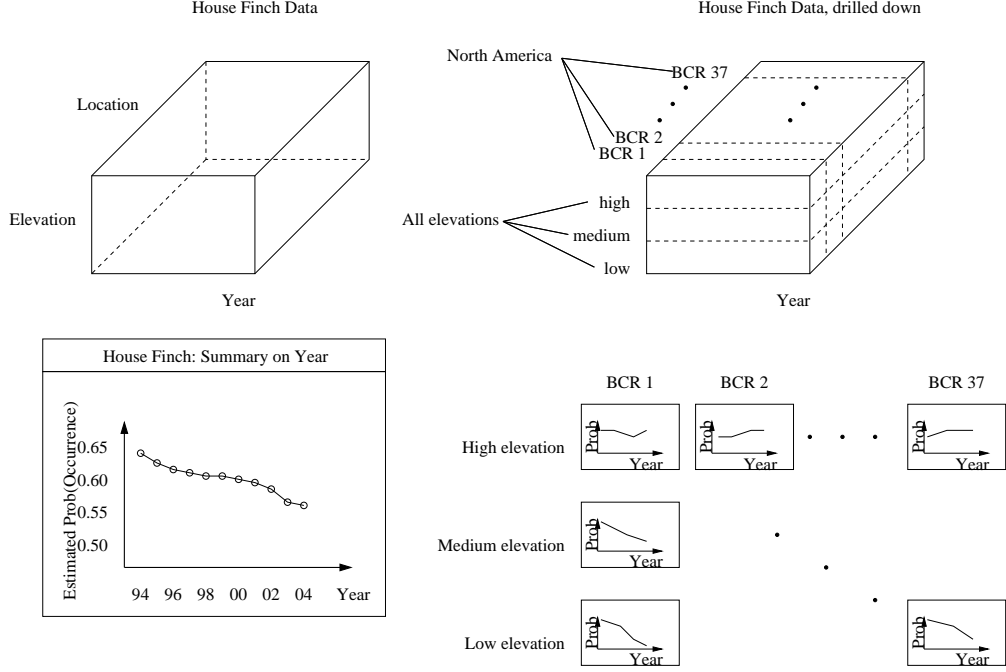


Figure 7: Creating a group of related summaries by drilling down on elevation and location

4.2 Multiple-Summary Measures

The measures discussed so far are applied to single function summaries individually. Another way of finding interesting patterns is to compare a function summary to other *related* function summaries or to additional data that encode the user’s prior knowledge or expectations. Figure 3 shows a typical example. The summary plots for BCRs 22 and 24 taken in isolation might not appear very interesting. However, together the four summaries are interesting because they indicate that there are two groups of BCRs with very distinct annual patterns.

There are two parts to computing multiple-summary scores for ranking. First, the user has to specify which summaries together should form a group. Second, a score has to be computed for each group. In the example shown in Figure 3, each group is identified by a species (Purple Finch) and a predictor variable (year). The members of the group are the model summaries for this species-feature combination, but now at the more fine-grained level of individual summaries for each BCR.

4.2.1 Specifying Groups of Summaries

A common way how groups are obtained is by *drilling down*, i.e., increasing the resolution of one or more predictors. Consider a model summary like the ‘Year’ plot in Figure 1. This summary was obtained by marginalizing over all other predictors, including latitude, longitude, elevation, and so on. By drilling down on location from “all locations” to “BCR” level, we now have a more fine-grained representation of the year summary: one summary for each BCR. One could also drill down on elevation to see how the annual pattern compares for observations at different altitudes. Or one can drill down on multiple predictors simultaneously. Figure 7 illustrates how drilling down on location and elevation creates a group of related summaries—one for each (elevation level, BCR) combination. These summaries represent a fine-grained version of the summary on the left half of the figure, which summarizes observed occurrence across all elevations and BCRs.

The main steps for creating groups of summaries are the following:

1. Partition the data space by drilling down on some predictors along their *granularity hierarchies*.
2. For each partition, create sub-partitions of the the data space by drilling down further, usually on predictors that were not drilled down on in the first step.
3. Create a summary for the predictors of interest for each sub-partition.

The following example illustrates this process. As before, we have models for predicting observed bird abundance given a vector of predictor values describing location, time, observer effort, habitat features, human population characteristics, and climate features. We first partition the data space by drilling down on the species dimension from level “All” to level “Individual Species”. This partitions the original data space in a way that there is now one partition per species. For each of these partitions we create a group of summaries by drilling down on the location dimension from level “North America” to level “BCR” and on the elevation dimension from level “All” to level “high/medium/low”. As a result, for each bird species we now have groups of $37 \cdot 3 = 111$ sub-partitions—one sub-partition per BCR-elevation range combination. In the last step the user specifies predictors for which to generate visualizations. If she picks ‘Year’, then the system will compute a score for each group of 111 ‘Year’ summaries for the combinations of different BCRs and elevation ranges for each bird species separately.

More concretely, for the House Finch, the system would consider all the one-dimensional ‘Year’ summaries shown on the lower right in Figure 7. And it would compute some multiple-summary score for this set of 111 summaries. It would then consider the corresponding set of 111 summaries for the Purple Finch, the Downy Woodpecker, and so on. Depending on the selected multiple-summary measure of interestingness, some of these sets of 111 related summaries will be ranked higher than others. This helps researchers to identify bird species whose populations are affected differently depending on elevation and geographical region, potentially leading to new hypotheses about environmental effects on these species.

Notice that this ranking does not need to be limited to a single predictor like ‘Year’. The user could also ask for one-dimensional summaries on ‘Effort’, ‘Human Density’, etc. Each time there would be 111 summaries for the different BCR-elevation level combinations, but now for the other predictors. In Figure 7, the plots then would have ‘Effort’ on their x-axis rather than ‘Year’.

4.2.2 Scoring Groups of Multiple Summaries

Instead of computing a score for each individual summary, we now need to compute a score for each group of summaries. For the example in Figure 7, the entire group of 111 ‘Year’ summaries on the bottom right would receive a single score, which would be compared to the scores of other summary groups.

Diversity

- Intuition: How dissimilar are the different summaries in the group from each other?
- Definition: There are many ways to measure dissimilarity of summaries, e.g.,
 - Treat each summary $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ as a vector (y_1, y_2, \dots, y_n) in an n -dimensional space. Then define a diversity score on the set of all pair-wise vector distances between the summaries. An example would be the variance of the distances.
 - Treat each summary as an n -dimensional vector as above and run k-means clustering (or some other clustering algorithm). Then define a diversity score for a clustering, taking into account cluster diameters and distances between cluster centers. An example would be a weighted sum of average cluster diameter and average cluster center distance.

- Use any of the single-summary measures (see Section 4.1) to compute a score for each individual summary. Then measure diversity as the standard deviation of these individual scores.
- Motivation: In many cases, the baseline assumption is that the related summaries should be similar. E.g., the annual occurrence pattern is similar in the different BCRs. It is therefore interesting, and possibly very surprising, if related summaries show a great diversity.
- Note: We can choose a diversity measure that is invariant to shifting along the y-axis, i.e., summaries $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ and $\{(x_1, y_1 + c), (x_2, y_2 + c), \dots, (x_n, y_n + c)\}$ would be considered (almost) identical for any constant c . Then the diversity measure can be used to find statistical interaction patterns.

Compact-Clustering

- Intuition: Compute a clustering of the summaries in the group and compute a score based on how distinct the different clusters are.
- Definition: Treat each summary $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ as a vector (y_1, y_2, \dots, y_n) in an n -dimensional space. Compute a clustering, e.g., by using the k-means algorithm. Then use a scoring function on the clustering. This function should give higher scores for fewer clusters, smaller cluster diameter, larger distances between cluster centers, and lower diversity of summaries in the same cluster.
- Motivation: Figure 3 shows a clustering that would receive a high score. The summaries for BCRs 12 and 14 are very similar to each other, but very different from the summaries for BCRs 22 and 24. The existence of two very distinct pattern types makes the summaries in Figure 3 very interesting.

In the next section we discuss a case study how the Compact-Clustering measure can help in the discovery of interesting migration patterns.

5 Case Study: Finding Migration Patterns through Multiple-Summary Measures

Ornithologists want to understand bird migrations, a domain example characterized by complex, dynamic patterns in space and time. The ecological goals are to explore migration patterns and to delineate wintering and breeding ranges. Developing the tools to generate empirical estimates of breeding, wintering, and migration ranges represents a significant advance for ornithology and conservation managers who rely on hand-drawn maps based on expert opinion for many species. Our analytical framework can be used to detect and describe a broad class of predictor patterns.

In the remainder of this section we first present an example of an interesting bird migration pattern that we discovered through *expert-centered analysis* of a complex data mining model. Unfortunately, this pattern would not have been discovered without experts suspecting its existence. We then discuss how such patterns could be discovered *automatically*, even without having any experts hypothesizing their existence.

5.1 Manually Discovering a Migration Pattern

We explore the annual migration patterns of the North American Tree Swallow across the Eastern United States. Domain experts are interested in this species because of its known complex migration. The ecological challenge is to identify and describe the regions with the highest seasonal population densities. More specifically, biologists need to know

1. Are there regions with seasonally high or low concentrations of birds? (Existence)
2. How many distinct regions are there? (Ordering)
3. What is the duration of the seasonal concentrations? (Temporal description)
4. Where are these regions of seasonal concentration? (Spatial description)

Analytically, the goal is to automatically detect and describe potentially complex regions of high density in space and time. Technically, this requires an analysis flexible enough to detect a wide range of patterns. Temporally, trajectories can vary dramatically with both smooth and sharp transitions. Spatially, there are an unknown number of regions with potentially complex boundaries that may be either contiguous or non-contiguous. Spatial structure, both number and region configuration, often change through time.

5.1.1 The Predictive Model

The predictive model studied here is a Bagged Decision Tree analysis of Tree Swallow data collected with eBird, the online checklist program operated by the Cornell Lab of Ornithology (<http://www.ebird.org>). This program collects data from tens of thousands of citizen scientists across the country each year. There were approximately 60,000 observations from 2004-2007. This predictive model accounted for habitat-selectivity of birds using remotely-sensed habitat information compiled at a 15km square resolution. Variation in detection rates is modeled as a function of effort spent watching birds, and variation in availability for detection is modeled as a function of the observation time of day and date.

The predictions used for the “drill down” were made over a dense set of 150,000 random locations uniformly distributed over the Eastern US. Predictions of abundance were standardized for observation effort and time of day to better infer the underlying biological processes. At each location, 100 evenly spaced predictions were made between January 1, 2006 to December 31, 2006. The resulting abundance estimates are considered as a “trajectory” at each location, a 100-dimensional multivariate “response” for each location.

5.1.2 Exploratory Clustering of Spatio-Temporal Patterns

In the first step of this analysis, intra-seasonal abundance trajectories are clustered to reveal relationships between locations. We want to identify and group similar temporal characteristics into clusters such that the resulting clusters provide a compact, sufficient description of the temporal patterns in the data. The second step is to visualize the resulting temporal and spatial patterns, discussed in the following sections.

We use a robust version of a k-means clustering algorithm called *k-medoids* [1] to cluster the trajectories. This algorithm searches for k representative objects, or medoids, among the observations of the dataset. After finding a set of k medoids, k clusters are constructed by assigning each observation to the nearest medoid. The goal is to find k medoids that minimize the sum of the dissimilarities of the observations to their closest representative object. For simplicity, we use Euclidian distance to measure dissimilarity between trajectories. This approach is very flexible, allowing a wide variety of temporal patterns and complex spatial patterns.

To automate choosing k , the number of clusters, we adopt a scoring function called the *silhouette* [2], which measures the distance between an observation and its cluster and the distance between the observation and its closest neighboring cluster. Observations with a large silhouette (close to 1) are very well clustered, a small silhouette (around 0) means that the observation lies between two clusters, and observations with

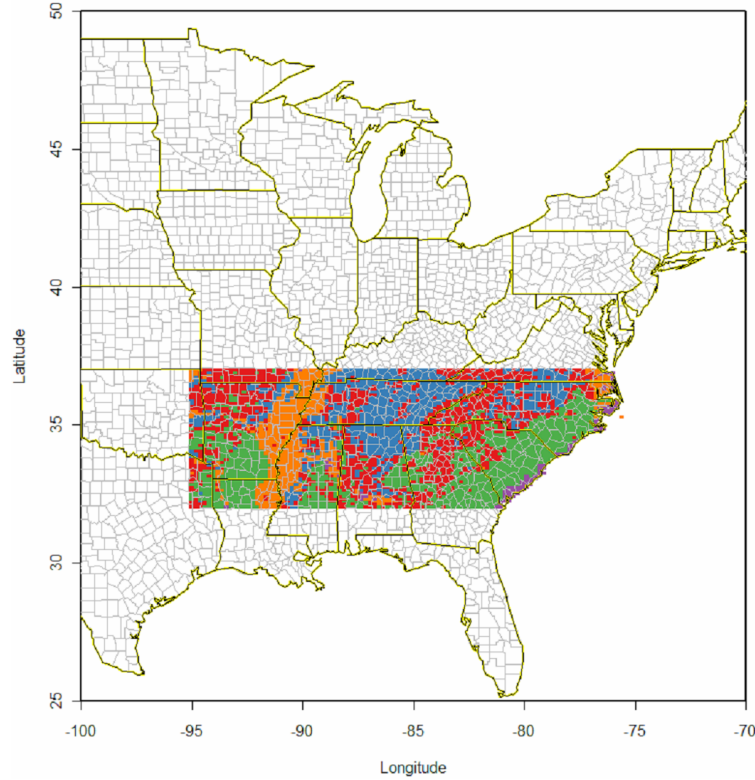


Figure 8: Spatial extent of migration analysis and abundance clusters

a negative silhouette are probably placed in the wrong cluster. The average silhouette is used as a scoring function on the clustering.

The ecological scale of inference is at the regional-state, or perhaps a multi-county level. Thus, this data represents a very fine-grained spatial resolution of the response. In order to facilitate computation we “pixelate” or grid the spatial locations. While pixelation will impart some additional smoothness of the abundance surface, we balance this smoothing so that there are very many pixels to support any regional inferences. We have chosen not to scale the abundance response because it is biologically interesting to identify and establish abundance orderings, e.g. core breeding grounds have the absolute highest concentration of birds during the summer.

5.1.3 Temporal Analysis of the Spring and Fall Migrations

We begin with a focused exploration of the spring and fall migration flyways. To facilitate this we have restricted the spatial domain to lie between 37 and 32 degrees latitude, the southern limit of the breeding range and the northern limit of the winter range for Tree Swallow, respectively (Figure 8). This region is pixelated into a 50-by-100 grid resulting in 4934 non-zero trajectories. The spring and fall migration trajectories are analyzed separately. We use 31 abundance estimates from February 3 until May 24 to study the spring migration and 31 abundance estimates from August 6 until November 24 to study fall migration patterns.

Each analysis begins by investigating the number of clusters. For both spring and fall data we tried clusterings with 2 to 15 clusters. The corresponding average silhouette values are plotted as a function of the cluster number for spring and fall data in Figures 9 and 10, respectively. Both of these plots suggest that



Figure 9: Average silhouette plot for spring data. The plot suggests that two clusters are best supported by the data.

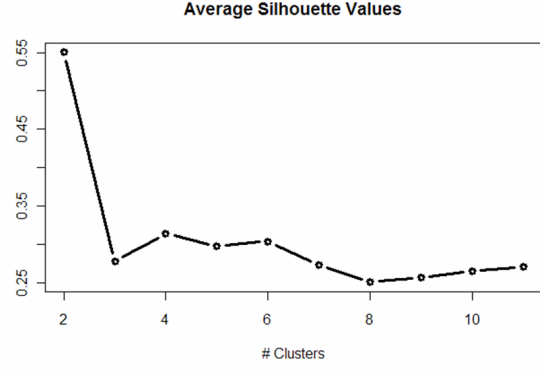


Figure 10: Average silhouette plot for fall data. The plot suggests that two clusters are best supported by the data.

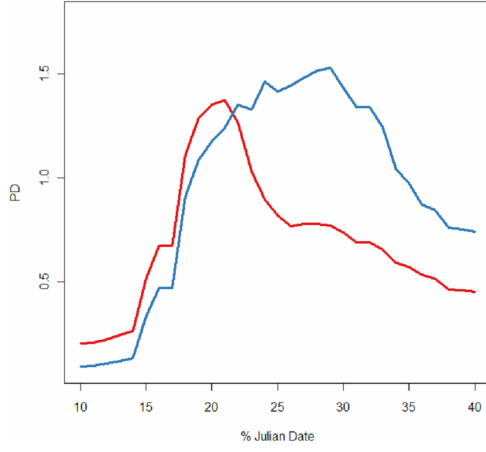


Figure 11: Average cluster trajectories for cluster 1 (blue) and cluster 2 (red) for spring.

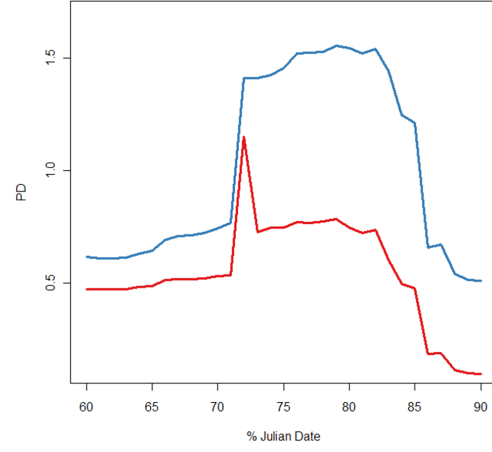


Figure 12: Average cluster trajectories for cluster 1 (blue) and cluster 2 (red) for fall.

the data best support a two cluster resolution. The average cluster-wide trajectories represent the temporal patterns identified by the clustering. Figures 11 and 12 show the average spring and fall cluster trajectories, respectively. The highest concentrations shown in the spring trajectories describe two similar migration patterns: a smaller early onset migration in the red region versus a slower, longer duration migration wave in the blue region. The fall trajectories describe regions with similar migration onset and duration (ignoring the spike), though with the highest density of Tree Swallows in the blue region.

5.1.4 Spatial Analysis of the Spring and Fall Migrations

The regions corresponding to the temporal clusters above are mapped back to latitude and longitude with corresponding colors (see Figures 13 and 14). Note that clustering was performed on the temporal trajectories. Spatial information, e.g., geographic proximity, was not used as a clustering criterion in any way. Hence there are no constraints on the types of spatial patterns that can be revealed under this visualization. The spring map (Figure 13) shows an early initial wave of migrants flying north through the red

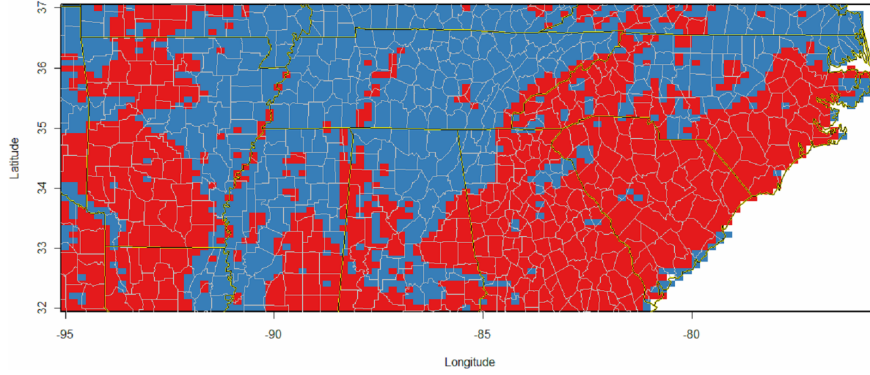


Figure 13: Spring migration cluster map

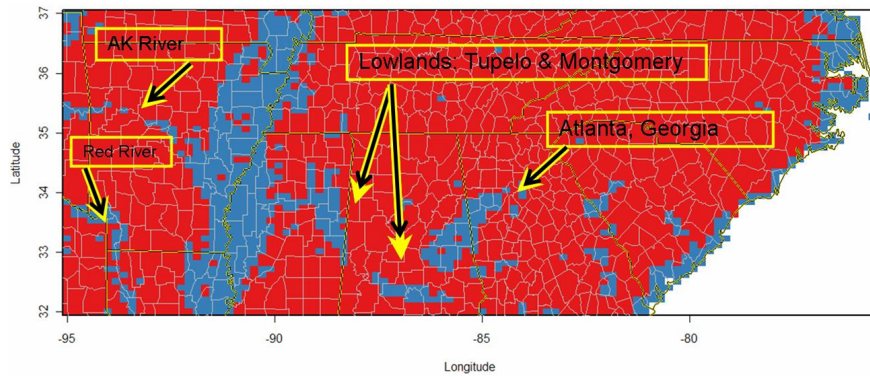


Figure 14: Fall migration cluster map with geographical annotation

Piedmont/Southern Appalachian region and the larger wave of migrants moves north a little later through the blue region.

The fall surface (Figure 14) clearly identifies two southerly flyways, both extending the full 37–32 degrees along the Atlantic coast and the Mississippi River Valley. This confirms what experts are beginning to find based on radar data [3]. The Fall migration map is especially interesting because it also identifies some additional “flyways” attached to the Mississippi River valley. By comparing these regions to a map, it can be seen that most of these areas are major rivers or lowlands attached to the Mississippi. This makes sense since the preferred Tree Swallow food source, insects, persist longer in riparian areas. We do not have an explanation for why the Atlanta, Georgia region was included in the blue region.

5.1.5 Empirically Generated Core Breeding and Non-Breeding Maps

Finally, we use our exploratory clustering pattern recognition to generate core breeding and non-breeding range maps based on observational data. The idea here is to cluster *annual* trajectories from across the Eastern United States and visualize the resulting clusters. This region is pixelated into a 75-by-105 grid resulting in 3834 non-zero trajectories. The annual trajectories consist of 81 abundance estimates from February 3 until November 24. The four-cluster map and corresponding trajectories are shown in Figures 15 and 16, respectively. The core breeding region is identified in green and the northernmost wintering region is shown in purple.

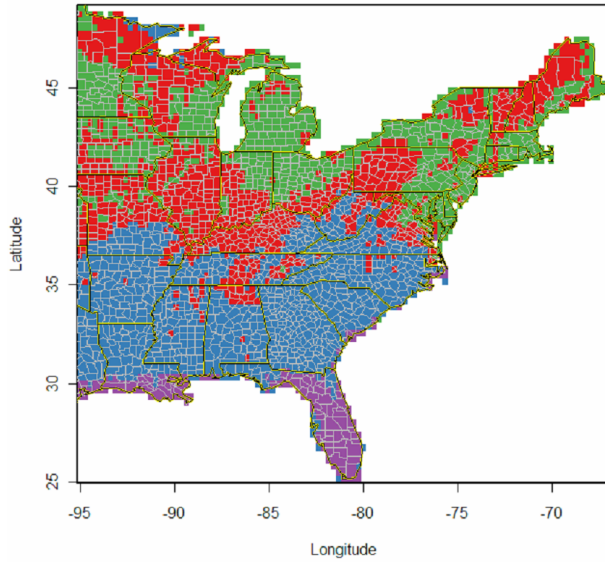


Figure 15: Annual cluster map

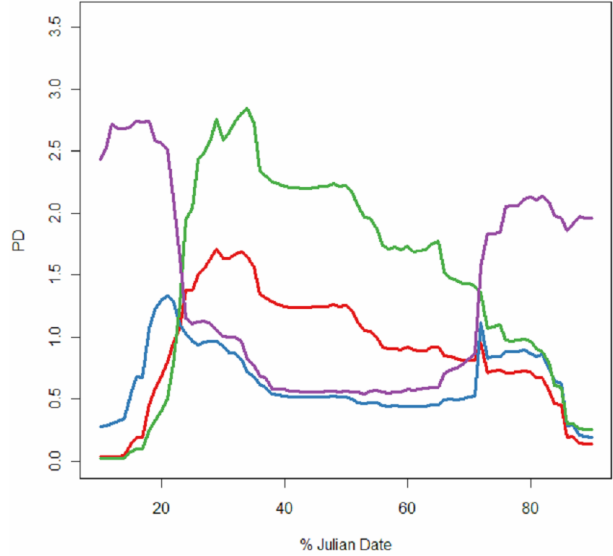


Figure 16: Average cluster trajectories for annual cluster map

5.2 Discovering Interesting Migration Patterns Automatically

The migration patterns discussed so far were discovered using semi-automatic data mining techniques, but a domain expert still had to direct exploration to the right species (Tree Swallow), the right predictors (space and time), and the right resolution (spring/fall or entire year). We now discuss how one could find such patterns automatically by searching various species, predictors, and predictor granularities and specifying an appropriate interestingness measure.

5.2.1 Benefits of Automatic Discovery

Having seen the above migration patterns for Tree Swallows, several questions might arise naturally:

1. Are there other species that show interesting and interpretable migration patterns?
2. Could we discover such patterns even without a priori knowledge on which time periods or geographical regions to concentrate?
3. Are there more complex patterns involving other predictors? For example, are there interesting patterns if we limit our attention to certain habitat types, elevation ranges, or human population characteristics of the environment?

Exploring these questions by manually trying different species, regions, time periods, and predictor combinations is infeasible. Here is where our pattern search engine can make a huge difference. Users only need to specify the pattern search space, what types of patterns they are interested in, and then they will receive the most promising patterns found automatically based on their preferences.

5.2.2 Defining the Pattern Search Space

Selecting species. Instead of limiting the analysis to a single species, the researcher would like to see if other species also show an interesting migration pattern. She could select any set of species of interest.

Selecting summary predictor(s) and response. For migration patterns, researchers are usually interested in seeing absolute abundance for different days of the year. In this case the summary predictor is 'Day of Year' (also called '% Julian Date') and the response is 'Abundance'. (For other analysis tasks, users might select other summary predictors and responses.)

Selecting the data space region of interest. Migration patterns, and other bird ecology patterns, might be visible only at certain scales or in certain habitat types, geographical regions, and so on. Hence the user has to specify all data regions of interest where the system should search for patterns.

In the migration example in Section 5.1.3, the domain experts selected latitude to be between 32 and 37 degrees and longitude to be between -75 and -95 degrees. For the larger-scale breeding range analysis (Section 5.1.5), the data space region of interest were latitudes between 25 and 50 degrees and longitudes between -65 and -95 degrees. For all other predictors the range was not limited. If, for example, the researcher wants to focus on migration patterns in grassland habitat only, she could add a corresponding selection for the habitat type predictor.

Specifying groups of summaries. Migration patterns like the ones presented in Section 5.1.3 can be discovered with multiple-summary measures of interestingness. As discussed in Section 4.2.1, these measures require the user to specify groups of summaries based on drilling down in the data space. In the example, we drilled down on both latitude and longitude to create a regular 50-by-100 grid of the region of interest.

Instead of limiting drill-down to a fixed grid size, the researcher could ask the system to try region partitionings of varying scales. Also, instead of drilling down on latitude and longitude, she might want the system to explore other predictors as well. For example, one could drill down based on elevation ranges and land cover characteristics to find clusters of elevation and land cover combinations with similar migration trajectories.

5.2.3 Choosing a Measure of Interestingness

For each combination of species, summary predictor(s), response, data space region of interest, and group definition that the user specified, there is a corresponding set of model summaries like those on the bottom right of Figure 7. As we discussed in Section 4.2.2, a score can be computed for such a set of summaries based on a given multiple-summary measure of interestingness. The system can then rank the summary sets by their interestingness and direct the researcher's attention to the most relevant ones first.

The main challenge is to select the right measure of interestingness that matches the intuition of what the researcher is looking for.

To make this concrete, assume the researcher did not suspect that Tree Swallows might show interesting migration patterns in the Eastern US. How could we ensure that for example the fall migration pattern for the Tree Swallow in the Eastern US between 32 and 37 degrees latitude receives a high score, while less interesting patterns receive a lower score? Asked differently, the question is what makes the fall migration pattern in that region interesting for a researcher.

The two most important criteria for interestingness of a migration pattern are that it (1) is concise and accurate, and (2) matches established domain knowledge. The Tree Swallow fall migration pattern as summarized by Figures 12 and 14 is concise, because it represents the 4934 trajectories with only two clusters. It is accurate as evidenced by the high silhouette score for 2 clusters (see Figure 10). A high silhouette score implies that the cluster medeoid trajectories are good representatives of the cluster members, i.e., they accurately capture individual trajectories in the clusters.

The migration pattern matches established domain knowledge in the following sense. First, even though the clustering did not take spatial information into consideration at all, the clusters create a remarkably clean partitioning into large consecutive red and blue regions (see Figure 14). This matches the expectation that

neighboring regions should generally show similar migration patterns. Hence a pattern with large consecutive regions belonging to the same cluster usually is more interesting than a “patchwork” of small red and blue regions all over the map.

Notice that matching established domain knowledge is a rather soft criterion. And in some cases patterns that run contrary to existing belief are more interesting. Examining such patterns could lead to surprising new discoveries.

In the end, our system is value neutral and we want to give the researcher a query language that is expressive enough to specify her preferences. For the migration analysis example, the score of a pattern should depend on

1. Number of clusters found (the fewer, the better)
2. Silhouette score or similar measure of clustering quality (the greater, the better)
3. Patchiness of geographical region belonging to each cluster (few larger regions better than many tiny patches).

5.3 Further Work

The general strategy employed here can be used to explore patterns among any set of predictors. It should be noted that the “time” predictor along which the trajectories are clustered plays an important role in determining the kinds of patterns to be explored. The spatial surfaces plotted here are just two-dimensional surface plots conditioned on some clustering based on the third predictor, and can be used to visualize interaction surfaces across any two quantitative predictors.

Though these tools are adequate for generating interesting regions, we believe that the value of the clustering may be enhanced by using other cluster scores, especially scores tied back to predictive performance on a test or validation set. A similar concern is to account for potential uncertainty from making extrapolations over regions with sparse data. The implementation used here may also be generalized by clustering dissimilarity matrices, giving the analyst more control over the clustering algorithm.

6 Advanced Topics

6.1 Measuring Distance Between Summaries

6.1.1 Distance of Response Value Vectors

There are many ways to measure the distance between two one-dimensional summaries. A common approach is to define a distance metric on the vector of response values: for summary $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, this vector is (y_1, y_2, \dots, y_n) .

Given the vectors for two function summaries, say (a_1, a_2, \dots, a_n) and (b_1, b_2, \dots, b_n) , we can use the Euclidean distance metric. Alternatively, we can use another L_i metric.

In some cases the researcher would like to make the distance metric invariant to shifting and/or scaling. Shift invariance can be achieved by normalizing all vector components by subtracting the vector mean.

6.1.2 Spline-Based Distance

As an alternative to the L_i metrics on the response value vector, one could also fit a spline to each summary $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Then distance can be measured based on differences between spline coefficients.

6.1.3 Regular-Expression Based Distance

Regular expressions, like splines, can be “fit” to a *sequence* pattern. Regular expressions are typically used for string matching, but we believe they would also be useful for our problem. In general, regular expressions (regex) describe sets of strings over a finite alphabet Σ :

- The empty string e is a regex, denoting the set containing the empty string. Any alphabet character $c \in \Sigma$ is a regex, denoting the set that consists only of that character.
- If R and S are regex', then so are RS (set of all strings obtained by concatenating each string in R with each string in S), $R|S$ (union of R and S), and R^* (smallest set that contains e and R and is closed under concatenation).

For example, sequence $ABABABAB$ can be described as $(AB)^*$. Intuitively, the graphs for BCRs 12 and 14 in Figure 3 have such a regular sequence pattern for even and odd years, while others do not. However, a straightforward search for regular expressions will not suffice, because the same value of the response attribute (probability of occurrence) will almost never occur more than once for different predictor values. For example, the line for BCR 12 has the following sequence of y-values: 0.39, 0.058, 0.41, 0.048, 0.4, 0.054, 0.17, 0.053, 0.24, 0.063, 0.08. Without value repetition, the regular expression will be as complex as the original sequence.

We can address this issue by quantizing the response values, so that similar values map to the same 'bucket'. Another option is to work with the *relative change* between consecutive response attribute values in the sequence. By simply considering their sign, we can describe a sequence like the graph for BCR 12 in Figure 3 as **(down up)*** over alphabet **{up, down}**. These are initial ideas, and much more work is needed to identify the best way of fitting regular expression patterns to function summaries. We will also examine how to measure the interestingness of a regular expression pattern. For example, is a shorter pattern more interesting than a longer one? Or is a pattern over a smaller alphabet more interesting?

6.2 Interestingness for Unordered and Partially Ordered Domains

For predictors like 'Land Cover Type' there is no natural notion of an order between domain values. In this case some of the previously discussed ranking measures like slope of regression line are not meaningful. Clustering-based measures will usually work, but have to be carefully re-evaluated.

Partially ordered sets (posets) are somewhere between sequences and unordered sets. They arise naturally when we consider multi-dimensional spaces, e.g., two-dimensional partial dependence plots. As a default one can use interestingness measures developed for unordered sets also for posets. However, this approach might lose some information and we are therefore planning to investigate this problem in more depth.

References

- [1] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.
- [2] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [3] D. W. Winkler. Personal communication, September 2008.