MULTI-VARIATE ATTRIBUTE SELECTION

FOR AGRICULTURAL DATA

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Xing John Xu

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

October 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

MULTI-VARIATE ATTRIBUTE SELECTION FOR AGRICULTURAL DATA

By

Xing John Xu

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Anne Denton
Chair

Dean Knudson

John Nowatzki

Simone Ludwig

Approved:

11/03/2015
Date

Brian Slator
Department Chair

ABSTRACT

Farmers always have been concerned about the quantity of crops (yield) as well as the quality of crops (sugar content of the sugar beets). The quality and quantity of crops are affected by various attributes, some are natural elements (rain, sunshine etc) and some are not (the amount of fertilizer, seed type etc). Some techniques have been developed to discover attributes that are important to different crops' yield. But within those selected attributes, how can we tell one attribute is more important than the other? The proposed algorithm is aimed to utilize the advantages of multiple response attributes to select the important attributes and then put the selected attributes in a hierarchical order. Although at the end this paper only focuses on yield prediction, any other target attribute can be a candidate for the prediction model.

ACKNOWLEDGMENTS

I would like to give special thanks for Anne Denton for her vision and guidance, many good discussions with my colleague, Eric Momsen; last but not the least, many thanks for all the encouragement from my parents, family members and friends.

This study is made possible by Grant No. 1114363 from National Science Foundation. This work also got a lot of support from the NDSU-Industry Consortium, and special thanks for the American Crystal Sugar Company for providing valuable data.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Agricultural applications provide a variety of data sources and with that complexity the existing data mining techniques are not sufficient enough to answer all the questions. More and more farmers are becoming interested in more than just the quantity of the crops, which is measured by the pure weight of crops. For example, many wheat farmers also care a lot about the protein content of the wheat and sugar beet farmers care about sugar content and sugar lost to molasses. This means that the sugar beet production can be described as three-dimensional vector data: yield, sugar content and sugar lost to molasses. Weather affects the growth of sugar beets greatly, therefore rainfall and temperature data are used in the data analysis; satellite images have also been used to measure the health of the crops by measuring how green they are. Most of the existing statistical or data mining algorithms use multiple explanatory attributes to predict single target attribute, but they lack of the ability to analyze multi-dimensional target attributes. There are other data mining techniques that can do attribute selection for multiple response attributes, but some of them don't rank the importance of the selected attributes. Therefore to find interesting patterns and to derive the top most important attributes, it's important to use all the available information including all the target and explanatory attributes.

1.1. Problem Statement

There are many existing data mining techniques that can capture the relationship between multiple explanatory attributes and a single categorical or continuous target attribute, but the problem is that many times we can not really find meaningful patterns using a single target attribute and there can be some meaningful patterns found when multiple target

attributes are considered. As shown in Figure 1, the X's and O's represent two data types and they are normalized. Both of the data types are evenly distributed when they are projected on either axis a or axis b, a and b here represent single target attribute; therefore if we only have one target attribute we can not see any difference between X's and O's and there is no meaningful information. However, when we evaluate the data on both target attribute a and b the same time, we can see that X's have an obvious pattern, they gather together in both corners of the plot; but O's are just some scattered points. This problem is well known in the world of classification.

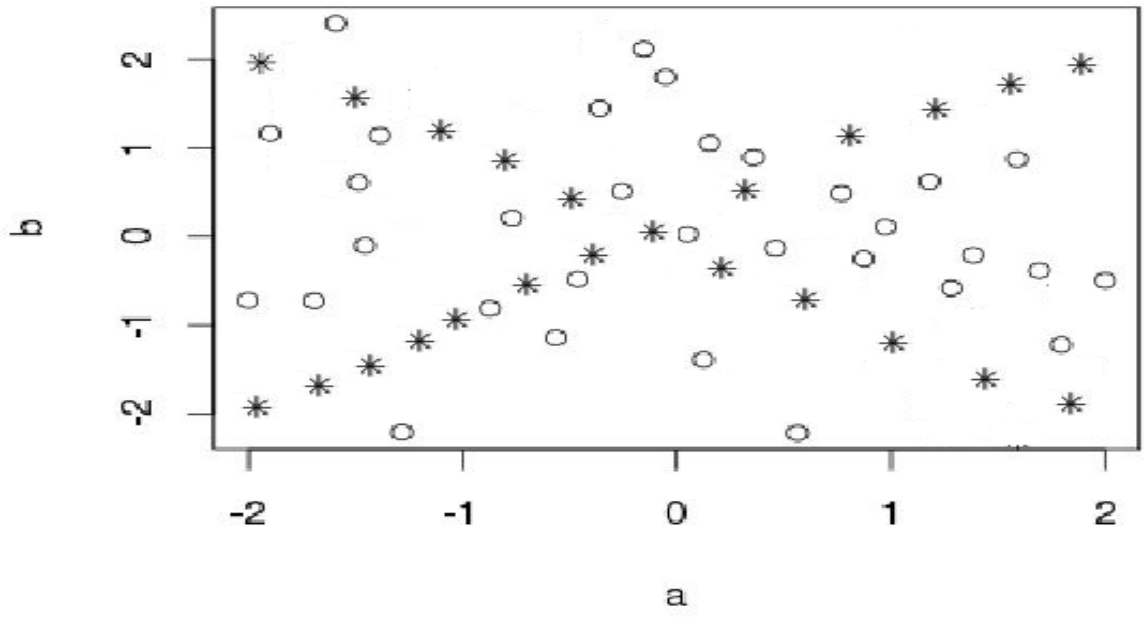

Figure 1. Example to show common problem of only considering one demension data. But meaningful patterns can be found when multiple target attributes are considered.

One solution to the problem presented in Figure 1, is to find more target attributes that will give us more information about the relationship between the independent attributes and target attributes. After more target attributes are selected, one method is to develop a model for each target attribute individually and combine them afterwords. However, this method runs into the risk of the interrelationship between the target attributes; it assumes that the response attributes are independent to each other. Another way is to combine the target attributes into a vector format and evaluate them at the same time, which is what we propose here.

Figure 2 illustrates our solution. On the left hand side of the plot, only one response attribute is under consideration and the highlighted data points in both plots seem to have some patterns; the gray lines seem to be randomly distributed data points. However, when we evaluate them in multiple response attributes and plot them out on the right hand side of the Figure 2, we find they are very much different. Data set from the top plots have significant pattern in multiple attributes A, B, C; but data set from the bottom plots don't have any patterns on the multiple attributes.

When approaching the problem described in Figure 1, the key is to notice that both the explanatory attributes *and* response attributes are mutli dimensional. With the existing data mining techniques, they work well with multi-dimensional explanatory attributes and single response attribute; but it lacks the ability to work with multi-dimensional response attribute. There are some alternative ways. One is to try to use explanatory attributes to predict each response attribute individually; the other way is to project all response attributes into one single attribute. However, both techniques fail to capture all information and fail to discover some significant patterns like shown in Figure 2. This proposed algorithm

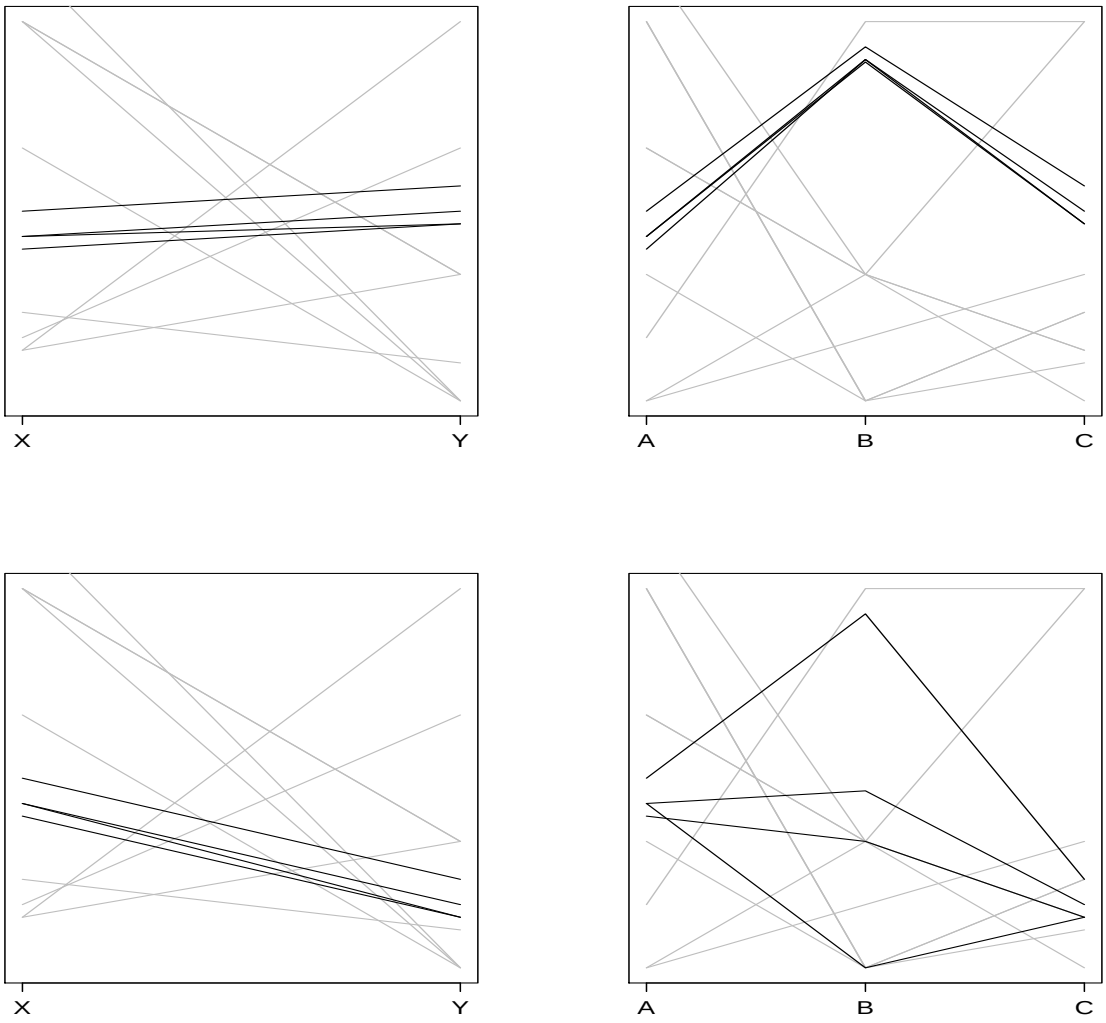

Figure 2. Concept question: if some patterns exist in attributes X and Y, can the similar patterns be found on multiple target attributes: A, B, C? In this case, on the left side plots, certain patterns (highlighted with black lines) exist in attributes X and Y. But when the same data is plotted over multiple attributes A, B, C, the top data set shows significant patterns but not with the bottom data set.

utilizes all the response attributes to find the significant patterns and find the significant explanatory attributes, ranking them in the order of importance.

Comparing with conventional data mining project data sets, which usually have just one categorical or continuous target attribute, the agricultural data set used in this study has multiple characteristics described by multiple attributes. In the agricultural domain, farmers care about quality of crops as much as about quantity. In this specific case, we place high importance on sugar beet total yield (quantity), its sugar content and sugar lost to molasses (quality). Here sugar content is measured by sugar per weight of beet. For instance, a sugar beet field with high yield but with low sugar content is not desired, because processing sugar beets takes a lot of energy to get rid of the extra water out of the sugar beets. Therefore, the lower the sugar content, the more energy is needed to produce the same amount of sugar from a high sugar content field with the same yield. On the other hand, a field with a low yield and high sugar content is also not desirable because the total sugar content is very low.

CHAPTER 2. RELATED WORKS

In this section we are going to look at some existing models for predictions and attribute selections. For example: linear model, Kullback-Leiber Voronoi and MANOVA etc. We also examined some information gain techniques, for example Kullback-Leiber divergence. Since this research is focused on agriculture, dynamic models from a biology perspective are mentioned in the section too. In the end, we list some data types used in the research, Normalized Difference Vegetation Index (NDVI) from sattelite images and other weather related data types.

Traditionally data mining techniques implement one of several techniques to recognize patterns in data sets or to predict outcomes. Often those techniques only work with a single target data. For example in [1], in order to identify which attributes effect a person's decision to purchase a certain item, association rule mining is applied to recognize patterns within the independent attributes and Chi-square test is performed to evaluate the significance level among the attributes. Those selected attributes are then used to predict target attributes. There are several existing models for prediction, Artificial Neural network and Classification And Regression tree (CART) models are used to predict the cost for cancer treatment [3].

$$Y = a_0 + a_1 * X_1 + a_2 * X_2 + ... + a_n * X_n \tag{1}$$

Linear model is often used for prediction, as it is pointed out in [8], Equation 1 provides a generalized equation of linear model; each X_i represents individual explanatory attribute and Y is the target attribute, and the key of finding a good model is to find the best coefficients a_i. [8] also uses cross-validation in its linear model. Cross-validation involves using parts of the data for training and the rest of the data for testing. For example, if a

data set consists of n data points, n_c data points are used to fit a linear model, and n-n_c data points are reserved to assess the predictive ability of the fitted model. One classic cross-validation procedure is called leave-one-out cross validation, it corresponds to the choice $n_c = n$ - 1, which is used in this research too. The purpose of cross-validation is to prevent overfitting the prediction model and to make sure it is more generalized for independent data set. Linear model makes certain assumptions that the explanatory attributes are independent of each other and the relationship between explanatory attributes and target attribute is linear. Although linear model does prediction, it does not rank the explanatory attributes and it lacks the ability of multivariate selection.

More complex data mining algorithms have been developed to find patterns on two or more continuous target attributes jointly. From a mathematical point of view, those target attributes can be represented as a vector. One way to find a relationship between target attributes and the explanatory attributes is finding distinguished patterns. This technique is known as vector-item pattern mining. In order to classify the gene sequence signatures in a gene data set [14], vector-item pattern mining is used to uncover patterns of different genes' presence in different gene sequences. Furthermore, [14] shows that the relationship discovered by pattern mining is not found using a classification algorithm. Even though vector-item pattern mining can give us more insight about how explanatory attributes effect response attributes, the technique does not rank the importance of the selected attributes.

Decision Tree is one model to select attributes, other models used for attribute selection include: Neural Networks and Support Vector Machine [18]. In [17], a Neural Network model is used to predict wheat yield. A Neural Network is a network of neurons, or nodes. A node receives an input then produces an output passed onto the another node. Neural Network

is a complex and adaptive system. A Neural Network can change its internal structure depending on the information passing between different nodes. Typically, the information passed between two nodes is controlled by the weights, and each connection between two nodes has a weight. For more detailed information about Neural Networks please refer to [9] and [10]. However, those models mentioned above traditionally only consider one target attribute not multiple target attributes.

The models mentioned above are some common known data mining techniques, extensive work has been done on developing dynamic models from a biology perspective too [24]. For example a dynamic model is specifically built for predicting the growth of tomatoes [25]. The limitation of a dynamic model is that it only works well for one crop, the same model does not work for other crops; also developing one dynamic model can be very time consuming. A limited number of crops have been chosen for dynamic model researching, for example tomato [25] and wheat [11]. Therefore, there isn't one dynamic model that works for a variety of crops. With our study, our aim is for our algorithm to select and rank important attributes for multiple target attributes yet not be limited to a specific crop.

Multivariate Analysis of Variance (MANOVA) incorporates multiple target attributes. MANOVA expands univariate ANOVA by involving several dependent attributes. Univariate ANOVA compares the means between two or more groups, in contrast MANOVA discovers the difference between two or more vectors of means. It discovers how the independent attributes affect the dependent attributes (target attributes) based on comparison of the error variance matrix and the effect variance matrix. In order to test the multiple dependent attributes, some artificial dependent attributes are created; those artificial dependent attributes are linear combinations of the measured dependent attributes. For example, MANOVA can

be used to compare several test scores for two populations, students in a Catholic school and students in a public school. In [21], MANOVA is used to study how people from different culture backgrounds differ in the Big Five personality traits: Neuroticism, Extraversion, Openness-to-Experience, Agreeableness, and Conscientiousness. However, MANOVA works under the assumption that dependable variables are normally distributed and have a linear relationship. Also it is very sensitive to outliers in the data set [22]. Attribute ranking is not necessary in all the multi-variate attribute selection methods, but often they are included in the algorithm [19]. Different algorithms use various informational gain techniques to rank the attributes, MANOVA isn't really known for ranking the selected attributes, but we can rank the attributes from comparing variance-covariance values [21], others use chi-squared value or p-value etc [1, 20]. One obstacle in agriculture is to obtain those attributes, a variety of technologies have been used to achieve that goal.

Researchers have taken advantage of remote sensing technology. There are various types, one of them is using satellites to capture an image of a massive area. Then the image can be processed to give us more insights of the crop's health[27]. One of the remote sensed attributes is Normalized Difference Vegetation Index (NDVI)[16], NDVI can be calculated from images captured by satellite or unmanned aircraft, this study uses satellite images. NDVI can be used to reflect how green crops are and it also can be used to predict next year's yield for the crop, in this case sugar beets. NDVI can be calculated throughout different stages of sugar beets, and NDVI during grainfill period improves the yield prediction for spring wheat [27].

A challenge with data pertaining to agriculture is which attributes to include, NDVI by itself is clearly not enough. Weather plays a big role in agriculture in general and previous

research has been done to show that by incorporating weather data in the model, it helps with yield prediction. Temperature data has been proved useful in predicting yield for spring wheat [27], and it is shown that precipitation has strong correlation with yield in [26]. But the problem is that there are so many other attributes that can be used in the prediction model, which attributes should be included in the model and which ones shouldn't be needs to be resolved. The proposed technique will allow us to find more information on the relationship between the explanatory attributes and multiple target attributes. The technique selects the important explanatory attributes then ranks them in the order of importance.

CHAPTER 3. CONCEPTS AND ALGORITHM

3.1. Concepts

The goal of our vector-vector pattern mining algorithm is to select significant explanatory attributes in a data set with multi-dimensional response attributes and also to rank the selected explanatory attributes with the more important ones on top of a tree structure, since the more important attributes give us significant patterns than with the less important attributes. As shown in Figure 3, the independent attributes are in a hierarchical structure, and in order to select each node attribute, their Kullback-Leibler divergence values are calculated. It measures the distribution difference between the left side of the branch (denoted as P) and all the data on that node. Whichever attribute gives the biggest K-L divergence value is the most effective one, so it is selected for that node. The same algorithm is applied for each node until reaching the end of the tree.

A technique often used to rank the importance of the selected attributes is Decision Tree, it often works well with a single target attribute. In [3], a Decision Tree is used to select variables that effect the cost of cancer treatment and then rank them. Decision Tree utilizes information gain to evaluate which attributes are important to the target attribute. There are several reasons why Decision Tree was chosen in [3], there are many different categories that effect the cost but with a Decision Tree it's easier to visualize and interpret, it can represent different scenarios etc. There are different methods to calculate information gain such as entropy and Kullback-Leibler divergence. Traditionally entropy is used in Decision Trees, however the limitation of entropy is that it only works with a single target attribute. In contrast to entropy, Kullback-Leibler divergence can target multiple attributes.

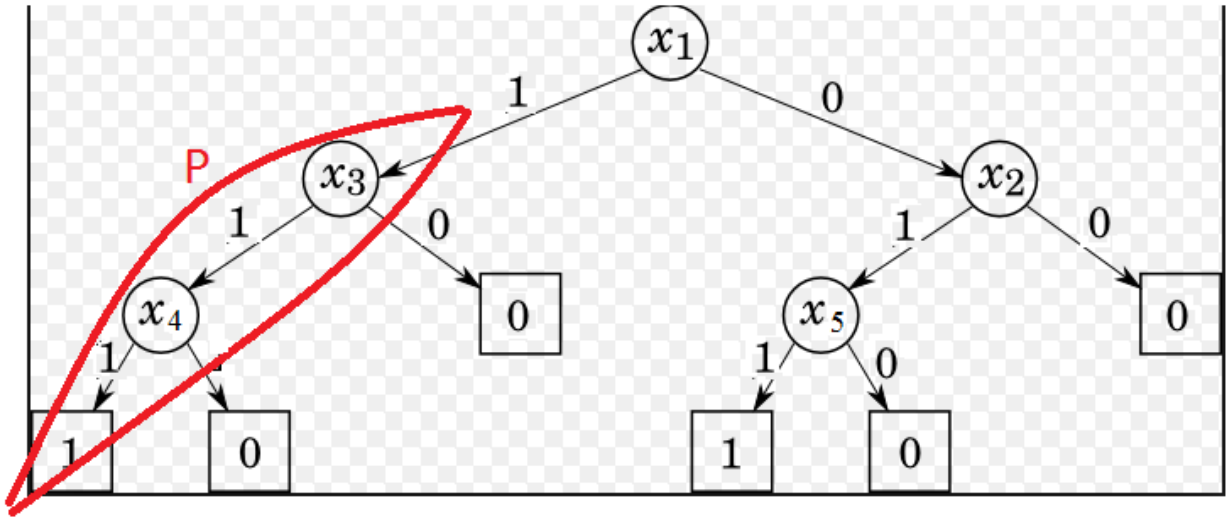

Figure 3. Combining Kullback-Leibler divergence and Decsion Tree

Since multiple target attributes are evaluated in this research, we choose Kullback-Leibler divergence for this paper.

Kullback-Leibler divergence, or K-L divergence, is a non-symmetric measurement of difference between two distributions P and Q. The KL divergence of Q from P is denoted as

$$D_{KL}(P||Q) \tag{2}$$

it is the information lost when Q is used to approximate P. The equation below shows the formula, p and q denote the densities of P and Q; where q is the reference distribution, which in this case is the full distribution of all the data points, and p is the distribution of data points that have the item of interest, and KL divergence is a non-negative value. It's worth noting that because KL divergence is non-symmetric, $D_{KL}(P||Q)$ is different from $D_{KL}(Q||P)$.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) ln(\frac{p(x)}{q(x)}) \, \mathrm{d}x \tag{3}$$

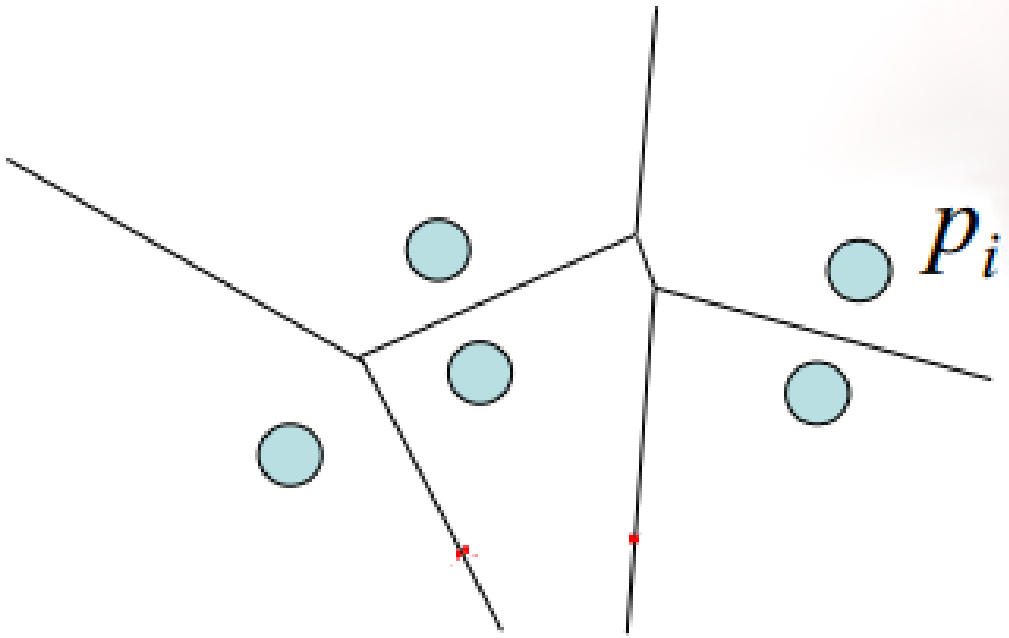

Figure 4. A Voronoi construction with all the P values

K-L divergence is frequently used in data mining algorithms[28, 4, 5]. As in the example of text categorization [6], each category is considered as a discrete target attribute and their distributions are compared with each other using K-L divergence. In this project the target attributes are continuous instead of discrete and all our explanatory attributes are in discrete format, to find the nearest neighbor a Voronoi cell based on K-L divergence is calculated, also noted as K-L Voronoi.

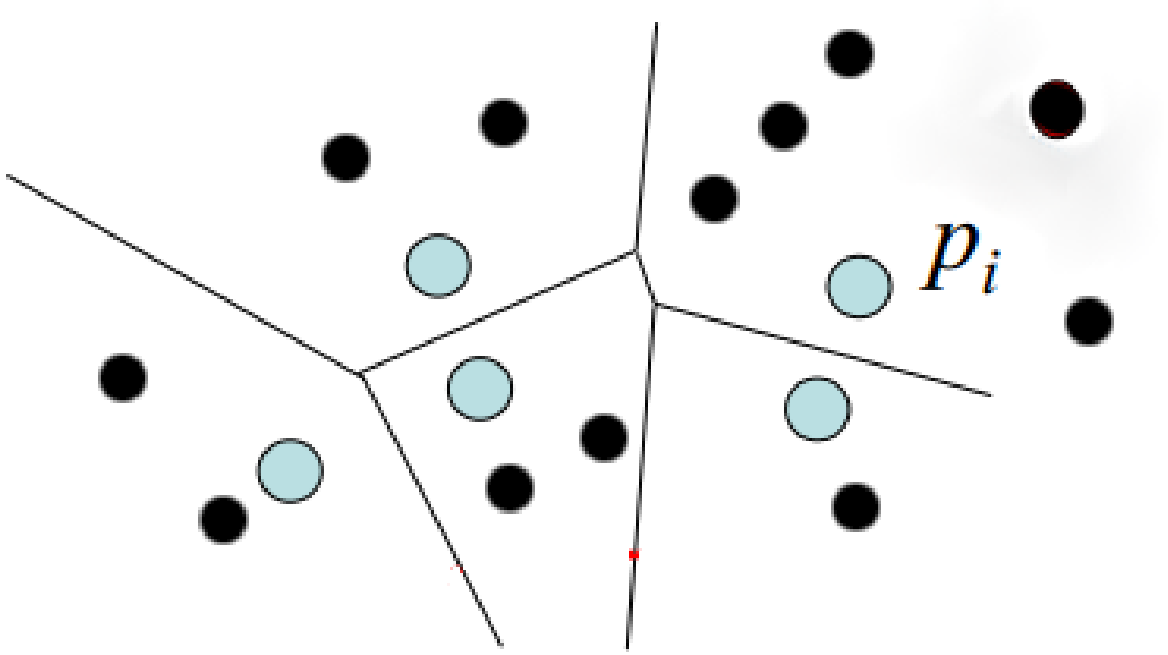

Figure 5. A finished Voronoi construction with all the P values that make all the cells

The K-L Voronoi has been used in [14], and K-L Voronoi uses the nearest neighbor [23, 7]. The Voronoi construction is shown in Figure 4 and Figure 5, first a seed is chosen, in this case they are all the P values, then the rest of the points are distributed into cells based on the seed values. They are grouped into the cell with the shortest distance to the P point, and the probability depends on the number of points in that cell. The probability of each P point is 1 divided by the total number of points of P and Q. Natural log of zero is invalid, which means that $p(x)$ has to be ensured not to be zero, so in the implementation of Voronoi function the number of points in each P cell is initialized to 1.

Once we have the tree structure based on the K-L divergence values, we will have a visualization of the top attributes. It's worth to point out that we care more about the structure of the tree than the individual leaf nodes. The individual leaf nodes are not used in our prediction model, but from the tree structure we can see what types of attributes are on the top of the tree; is it previous crops or soil type or the weather attributes etc. For example, if there are different types of previous crops showing up on top the tree structure, which means that previous crop category should be used in our prediction model; but not necessary just any specific kind of previous crop. And upon termination of a branch, it's either because all the binary attributes have been evaluated or because all the K-L divergence values are zeros on that branch.

3.2. Algorithm

The algorithm is implemented in R [32], an open source programming language. An overview of the algorithm is illustrated in the figure below. The first step is to list all possible binary attributes, more details of this step can be found in the data preparation section. For each binary attribute, P represents all the fields that contain that specific binary attribute, and Q represents all the fields on that branch of the tree. To select the most influential attribute on each branch, a KL divergence value is calculated for each binary attribute (Step 2). The more important the attribute is the bigger its KL divergence is, therefore the one with the biggest KL divergence value is identified (Step 3). The implementation of the KL divergence function is mentioned in the KL Voronoi section from the previous chapter. Once that binary attribute is identified (Step 4), it is removed from binary attribute list so that it won't be reconsidered (Step 5). Based on the selected attribute, two new branches (or two subsets) are created: one branch with all fields that contain that binary attribute (the

Algorithm 1: KL Divergence and tree structure based vector-vector pattern mining Algorithm

Data: $binary\_attributes$ /* all binary attributes on the branch */
Data: P_i /* fields contain the binary attribute */
Data: Q /* all fields on that branch */
Data: $left\_side$ /* fields have selected attribute */
Data: $right\_side$ /* fields don't have selected attribute */
Result: $tree$ /* tree structure */

1 $BuildTree(side, BinaryAttributes)$:
2 **for** i = *1 to* $|binary\_attributes|$ **do**
3 $\quad KL\_Divergence_i = \text{KLDiv}(P_i, Q)$
4 $index = \max(KL\_Divergence)$
5 $selected\_attribute = binary\_attributes[index]$
6 $binary\_attributes = \text{remove}(selected\_attribute)$
7 $tree = \text{BuildTree}(left\_side, binary\_attributes)$
8 $tree = \text{BuildTree}(right\_side, binary\_attributes)$
9 **return** $tree$

left side branch) and the other branch contains all the fields that don't have that binary attribute (the right side branch). Then the same algorithm is applied to those two child branches recursively on Step 6 and Step 7.

CHAPTER 4. DATA PREPARATION

Before proceeding to the results, the data set used in this study and its preparation will be described. Because of the uniqueness of agriculture, the data set used in this study draws upon several sources. Consequently, combination of the data types into a cohesive format was a time consuming obstacle. The proposed algorithm is applied to data pertaining to various attributes of sugar beet fields. The fields are located in the Red River Valley in northern U.S., which includes parts of Eastern North Dakota and parts of Western Minnesota. The data set includes confidential data provided by American Crystal Sugar Company, temperature and rainfall data from National Oceanic and Atmospheric Administration, and Landsat images from U.S. Geological Survey. The time frame of the collected data is from 2007 to 2011, and the spatial range covers majority of the Red River Valley. Each data type is pre-processed and then they are combined together using GRASS GIS software [31]. In this chapter, we are going to cover the processing of the data related to all the fields, weather data, and satellite data.

4.1. Field Data

Local farmers and American Crystal Sugar Company provided Substantial amount of data about each field. The data include geographical location of each sugar beet field and a few categorized and continuous data types too. The categorized attributes include soil types, previous crops, seed types, etc. Our target attributes were three continuous attributes: Yield, Sugar, and Sugar lost to molasses. Yield measures total weight of the sugar beets from each field; sugar is the weight of sugar content from the total yield; sugar lost to molasses is the weight of sugar lost during sugar-making process due to the heat, waste, etc. The desired

outcome is high yield with high sugar content and low sugar lost to molasses is desired. The three continuous target attributes were normalized in this study. All the data related to each field is stored in a vector, and each field has a spatial reference as shown in Figure 6. The different shapes of rectangles are the fields in the study.

4.2. Weather Data

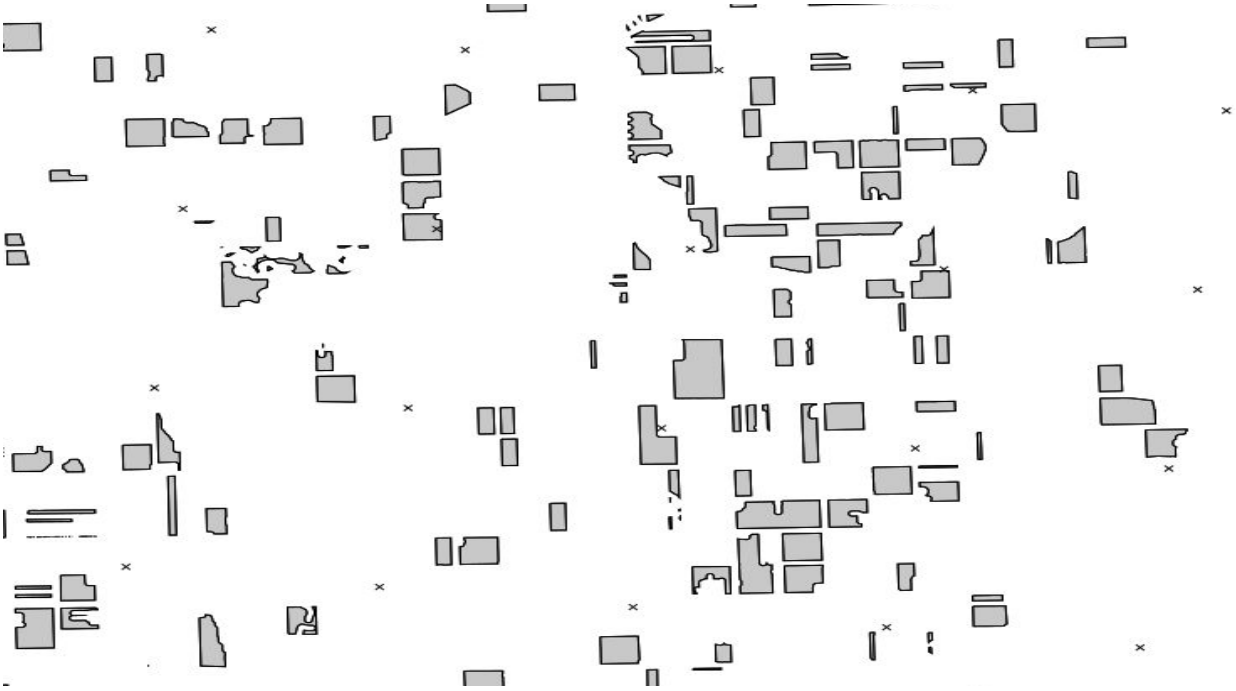

Figure 6. Fields with rainfall vector points, The x represents rain fall vector points and the rectangles represent fields

Weather data consists of rainfall data and temperature data. Raw daily precipitation data [33] was retrieved from National Weather Service of National Oceanic and Atmospheric Administration. As shown in Figure 6, this data was in vector format which produced a grid

of vector points. Since those points don't cover each field individually, each field uses the nearest grid point for its rainfall data and surface interpolation was used (interpolating data points on a two dimensional grid). Figure 6 illustrated the rainfall grid vector file displayed over the all the fields. X's designate points of measured rainfall, and rectangle shapes set field boundaries.

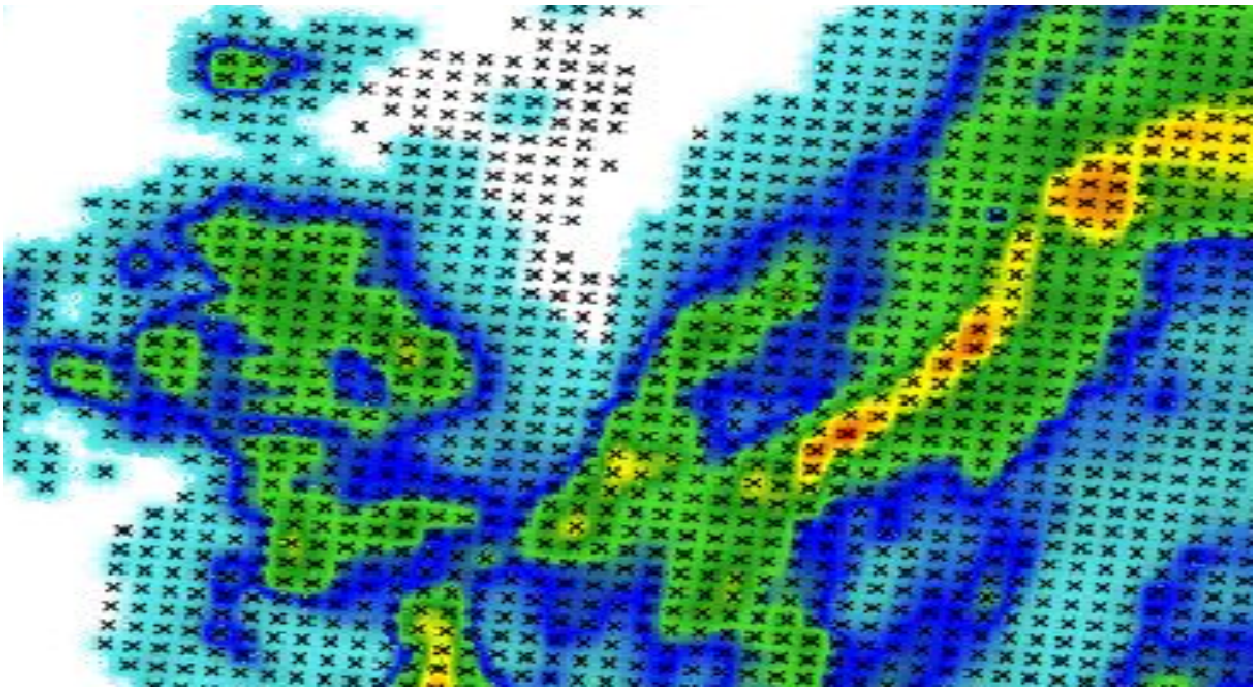

Figure 7. Rainfall in raster format displayed over the vector points after surface interpolation

Each vector rainfall file is then converted to raster file using surface interpolation Figure 7. Once we have the daily rainfall for each field, the monthly rainfall is summed up from the plant date to near-harvest date for each field.

Growing Degree Days approximates accumulated heat for scrips based on temperature. The daily min and max temperature reading can be retrieved from National Climatic Data Center, and is processed the same as the rainfall data. The minimum and maximum temperatures are adjusted accordingly in Equations 4 and 5 and GDD is calculated according to Equation 6. All temperature readings are in Fahrenheit.

$$\mathrm{T}_{min} = Max(\text{Actual Daily Min Temperature, } 34\,^{\circ}F) \tag{4}$$

$$\mathrm{T}_{max} = Min(\text{Actual Daily Max Temperature, } 86\,^{\circ}F) \tag{5}$$

$$\text{GDD}\,^{\circ}F = \frac{\mathrm{T}_{max} + \mathrm{T}_{min}}{2} - 34\,^{\circ}F \tag{6}$$

4.3. Satellite Imagery

Satellite imagery is available from Landsat 5 and Landsat 7 [30]. The areas of this study are covered by image tiles 30026 and 30027. The images from the two satellites are processed then combined using GRASS GIS [31], excluding the cloud cover and reflections from the atmosphere. Then those images are used to calculate (Normalized Difference Vegetation Index) NDVI. NDVI normalizes the difference between red and infared red, as shown in Equation 7. It is used to represent the health of the crops by measuring how green the target is, so the greener the target, the more positive its NDVI value is. The Figure 8 shows processed satellite image laid on top of the fields image.

$$\text{NDVI} = \frac{\text{near infrared} - \text{visible red}}{\text{near infrared} + \text{visible red}} \tag{7}$$

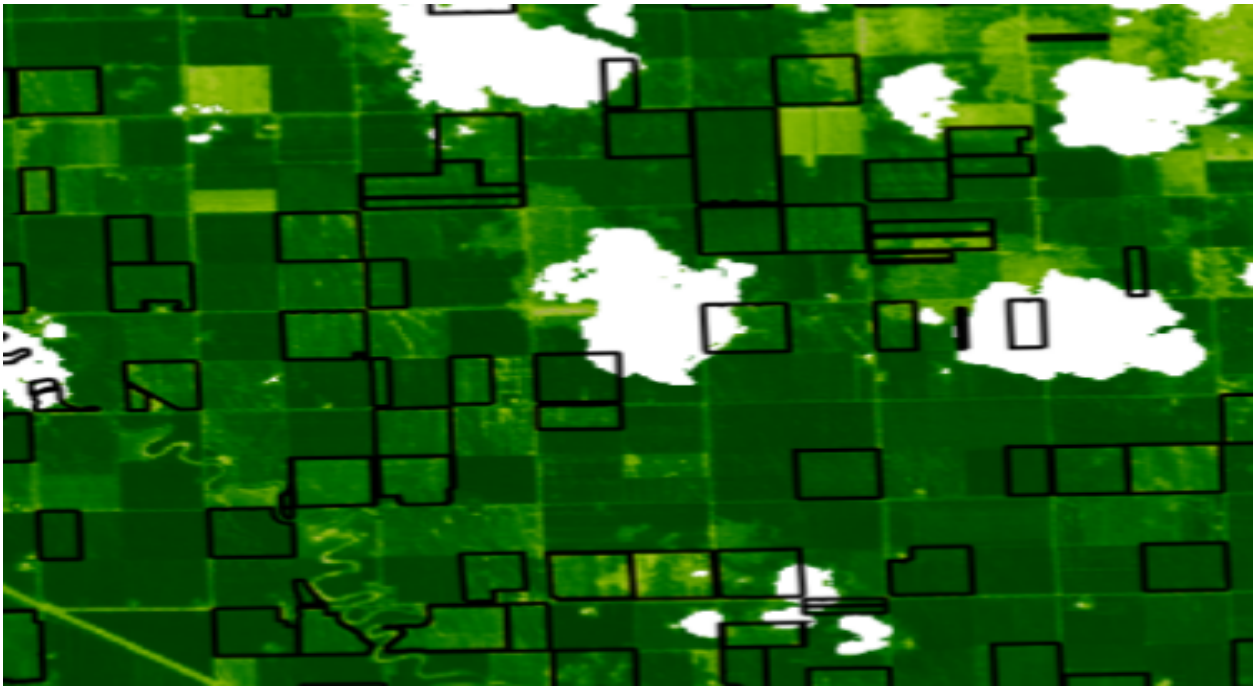

Figure 8. Calculated Normalized Difference Vegetation Index, the black rectangles represent the fields and the white areas are clouds

After all the data is processed in GRASS GIS, they are imported into R for analysis, then all of the explanatory attributes are transformed into binary attributes afterwords; there are two types of explanatory attributes: categorical and discrete. The categorical explanatory data types are separated into different types individually, then 1 represents the field has that data type and 0 represents otherwise. The numeric data (Rain, NDVI, GDD) is broken into two halves; 1 represents the value is above its mean and 0 represents below its mean. Note, some binary attributes are removed because of their small number of occurrences.

CHAPTER 5. RESULTS

The evaluation of the selected attributes from our algorithm is done by comparing results from regular multi-variate attribute selection algorithms. In this section a visualization of the tree structure from some subsets of the attributes and also a snapshot of the full tree structure is provided, we also show some of the discovered patterns from the attributes on the top branches. In order to assess the performance of the algorithm, a speed test is carried out on the algorithm. We recorded the time spent by the algorithm when various number of explanatory attributes were evaluated, starting from five to 25 in the increment of five. Then its time is compared against the time of regular multi-attribute selection algorithms on the same number of attributes. In the end, our top selected attributes are applied to linear model to predict total yield and that result is compared against the prediction from traditional multi-variate selection algorithms.

5.1. Previous Crops

Before we proceed to the results in depth, let's look at some results by running the algorithm on some subset of all the attributes. In this subset, the same algorithm is only used on all the previous crops (in order to preserve certain nutrition in the soil, farmers usually plant different crops on the same field every year); the results are presented in Figure 9. Similar with traditional decision tree graphs, the more important attributes are on the top of the tree structure and the less important ones are at the bottom. And because the nature of the binary attributes, each node in the tree only has two branches in this study; the left side branch represents the fields that have the binary attribute on that node, and right side branch represents fields don't have that attribute. Here zero is used to indicate that the end

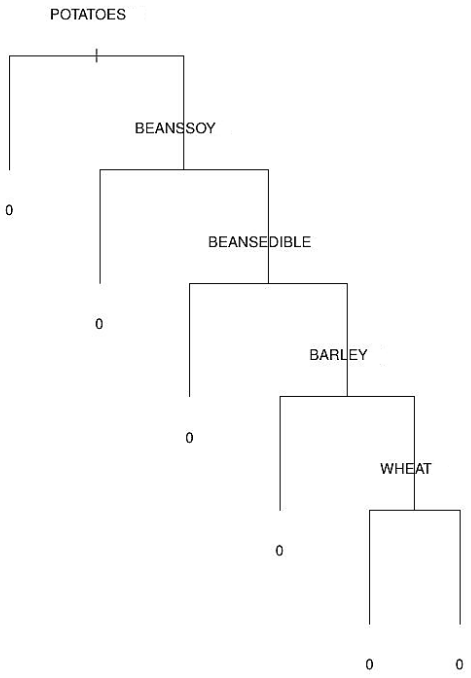

Figure 9. The same algorithm is tested on all the previous crops, the more important attributes rank higher on the tree structure. Potato is on the top branch, which validates the conventional wisdom. Zero signals the end of the branch and it appears on the left side branch is because we are only considering one category of the data.

of the branch has been reached. Since each field typically only has one type of cover crop, this tree structure looks like one sided leaning towards right. Some existing research [12] has shown that potatoes and edible beans are some very good choices for cover crop on sugar beet fields. As we can see in Figure 9, potato is on the top of the tree and edible bean is one of the top choices too. This validates that our algorithm against conventional wisdom when it comes to previous crops.

5.2. Soil Types

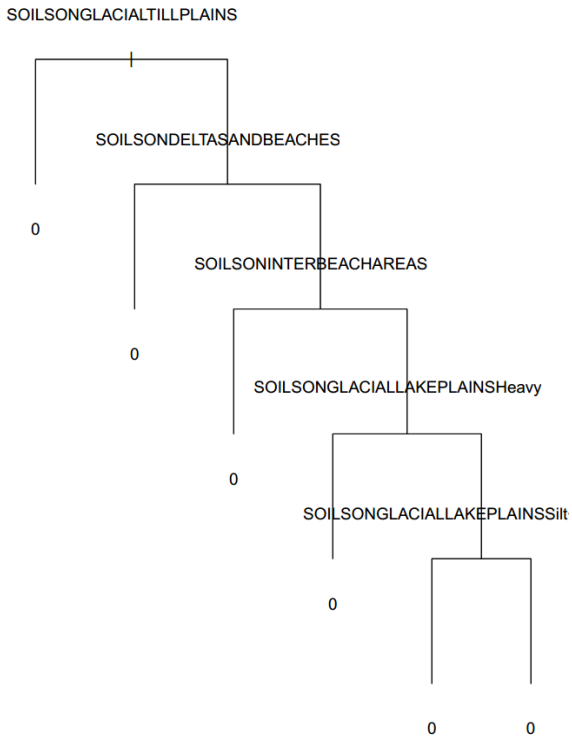

Figure 10. The same algorithm is tested on all the soil types of the fields.

The same algorithm is tested on the soil types and the result is shown in Figure 10. Since the field samples are taken from the Red River Valley region, the region's soil types are very limited. However, according to common practices soil that's well drained and no rocks is good for sugar beet; sandy soil is not considered ideal for sugar beet farming.

5.3. Full Tree Structure

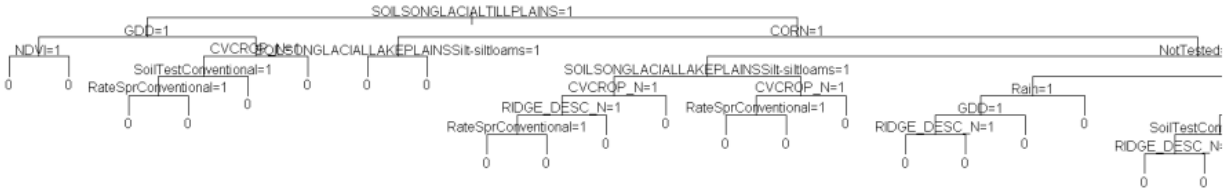

Figure 11. A snapshot of the full tree, this shows the top attributes of the tree. Soil types, previous crops and weather data dominates the top branches.

All the available explanatory attributes are used in constructing the full tree. Since there are a lot of attributes in this study, the full tree view can't fit in one page; Figure 11 depicts a portion of the top branches of the full tree. There are total 25 binary attributes used in the study, since NDVI, rain and GDD are calculated later for the field, they are split into halves by their mean. For example, if the field's NDVI is above the average NDVI, then it is considered that it contains NDVI; otherwise it doesn't contain the attribute. As expected, the top attributes are dominated by soil types, previous crops and weather data. Because agriculture depends on the weather condition, it makes logical sense that GDD, NDVI and Rain show up in many sub-branches. We are going to investigate the top branches in more detail below.

5.4. Example Pattern

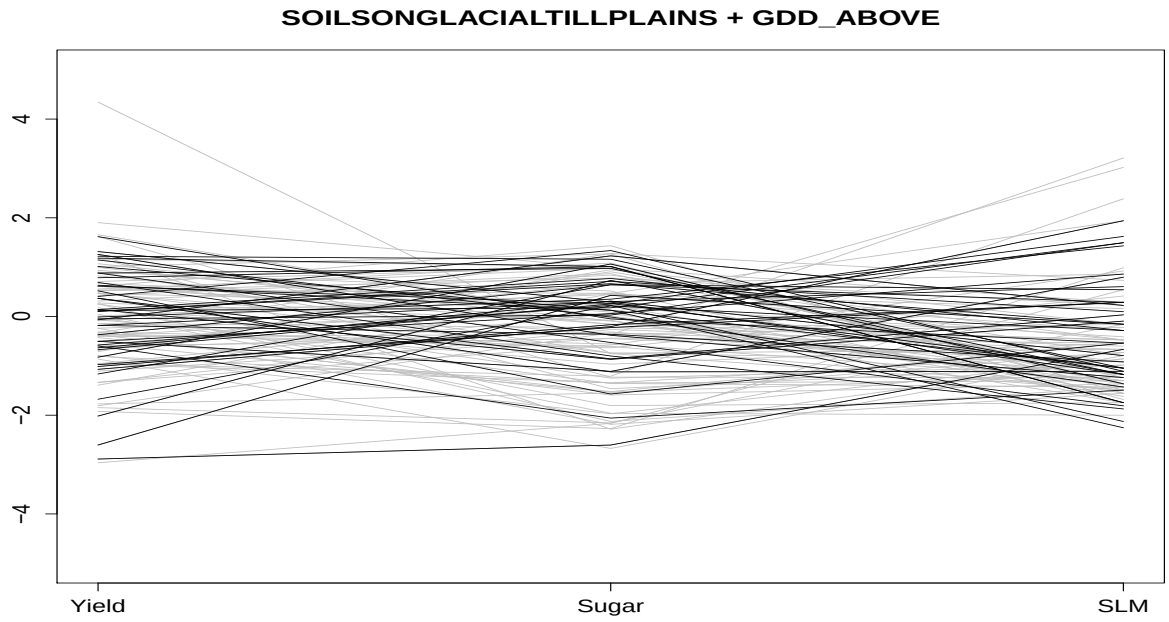

Figure 12. Pattern comparison between fields of above average GDD and fields with Glacial Till Plains.

Performance patterns reveal significance of explanatory attributes among taget attributes: Yield, Sugar, and Sugar Lost to Mollasses. Some patterns observed from our data set using the top selected attributes are shown in Figure 12 and Figure 13. In Figure 12 and Figure 13, the gray lines are the fields that have Glacial Till Plains as soil type. Since GDD is the attribute right below Glacial Till Plains, those fields are then separated into two types based on GDD; one type with GDD above its average (represented by the black lines in Figure 12) and the other type with GDD below its average (represented by the black lines in Figure 13). The black lines within each plot have distinguished performance patterns

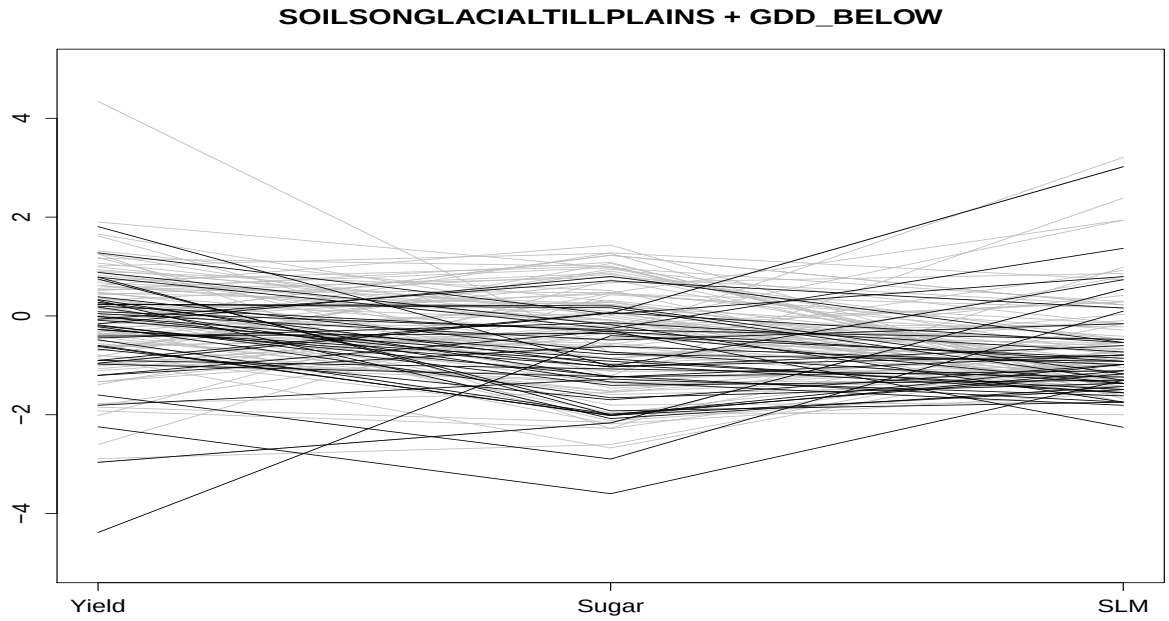

Figure 13. Pattern comparison between fields of below average GDD and fields with Glacial Till Plains.

comparing against the gray lines, and the black lines have different patterns between the two figures. This confirms that the top branches can successfully find distinguished patterns of response attributes and that they are not just randomly selected attributes.

5.5. Speed

To assess the efficiency of the algorithm, a speed test is carried out. The impact on speed due to additional independent attributes and speed comparison against traditional multi-variate attribute selection algorithms are described in this section. The number of explanatory attributes is incremented from 5 to 25 by step of 5. The same data is used on the traditional methods and the proposed algorithm, as shown in Figure 14, the dashed

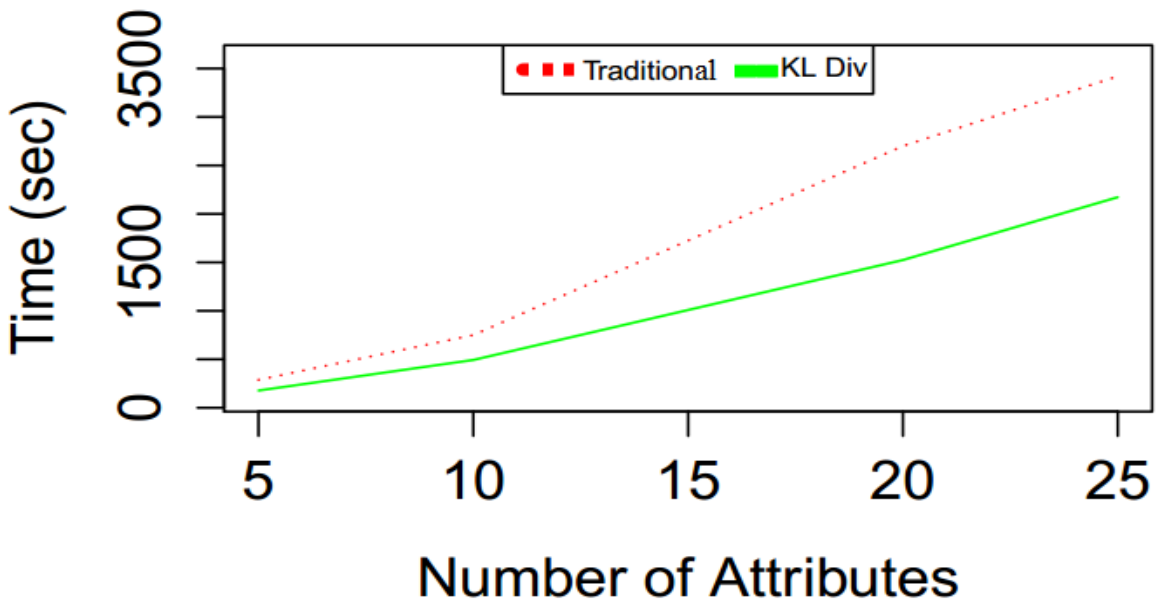

Figure 14. Comparison of run time from using our aglrithm versus using traditional multi-variate attribute selections depending on the number of explanatory attributes.

line represents the traditional multi-variate selection algorithms (they include: StepWise, Linear Forward selection and Best First, then the mean of their time is taken) and the solid line represents our algorithm (the same hardware is used on all algorithms). The multiple traditional multi-variate attribute selection algorithms are evaluated using Weka. As we can see, KL divergence is faster than the traditional algorithms all the time, and the gap increases as the number of dimensions increases. The KL divergence algorithm takes about half of the traditional algorithms' time at the end of the plot. And as expected, with the number of attributes increases as both algorithms take more time to finish. Most of the time in KL divergence algorithm is spent on calculating the KL divergence value (KL Voronoi

value) for each binary attribute and comparing all the values to come up with the selected attribute. At this point the KL Voronoi value calculation function is not optimized for speed yet.

The run time for KL divergence algorithm increases linearly with the number of explanatory attributes in a steady slope. However, the traditional algorithms' slope gets steeper as the number attributes increases. The number of dimensions in our study is relatively large in agricultural data sets, but it is definitely not large in comparison with other data mining data sets. Therefore, KL divergence algorithm definitely has a big advantage on speed over other traditional multi-variate attribute selection algorithms, especially when a large number of explanatory attributes are evaluated. The speed test is carried out on a Linux machine with 32GB total RAM and CPU Intel(R) Xeon(R) at 3.33GHz.

5.6. Predictions

It is interesting to observe the patterns described in the previous section, but using the selected attributes to improve yield prediction gives a bigger financial benefit to farmers and crop processors. So far future prediction has been focusing on yield only, so in this section of study the selected attributes are evaluated by yield prediction. In our previous researches, several models were explored and linear model achieved the best results in future prediction; therefore linear model is used in our evaluation process.

$$\text{Error Percentage} = \frac{\text{actual regional yield} - \text{predicted regional yield}}{\text{actual regional yield}} \tag{8}$$

As described previously in the generalized linear model Equation 1 in related works section, each X_i represents individual explanatory attribute and Y is the target attribute. A

baseline linear model is built to predict the total yield from year 2007 to 2011, so there are total five linear models; for each year's model, all the other years' data is used for training and then the model is tested on that year's data alone. In the baseline model, some collected farming attributes (e.g. previous crop, seed type, Nitrogen fertilizer) and weather data (e.g. GDD, NDVI, Rain) are included. However, in order to improve the prediction accuracy, based on our top selected attributes that baseline model is modified. Traditionally, the correlation coefficient is used in linear model, but we care the most about the total yield (total weight) of the sugar beets; so we will show the error percentage of the model based on yield. The overall error percentage (Equation 8) of the linear model is calculated by taking the difference between the actual total harvested sugar beets weight from all the fields and the predicted total weight of sugar beets for each year divided by the actual total harvested yield. Below we are going to compare the yield prediction results from our algorithm against the yield prediction results from traditional multi-variate selection algorithms.

Figure 15 shows the error percentages of yield prediction from year 2007 to 2011. The dotted line is the error percentage from the original model. The dashed the line is from new model without the selected attributes and the solid line is the new model with selected attributes. The original model is the model to be used, but from observation of the tree, some of the attributes in the old model are not on the top branches of the tree structure, so they are removed from that model. The new model does a better job predicting yield (dashed line vs dotted line), and that model is used to compare the impact of using our top selected attributes (the sold line vs the dashed line). In all three different approaches, the new model with our selected attributes achieves the best yield prediction; the biggest

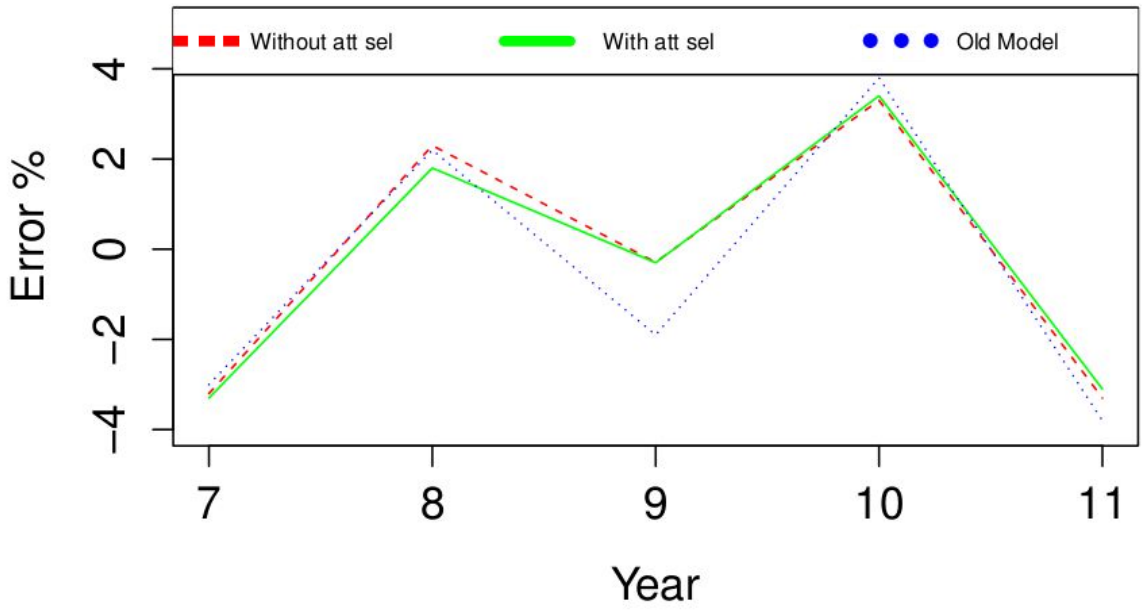

Figure 15. Error percentages of the same model with selected attributes and without selected attributes.

difference is in year 2009. It's about a two percent difference between the new and old linear models.

Another prediction comparison is between conventional multi-variate attribute selection and our algorithm. To do that, the selected attributes from the two approaches are tested in the same linear model. Those selected attributes are used to predict the yield for a whole year, and since we are mostly concerned about the total error percentage, Figure 16 shows the absolute error percentages of using conventional multi-variate attribute selection algorithms and our algorithm. The dotted line is the absolute prediction error percentage from using regular attribute selections and the other line represents the absolute error percentage from using our algorithm. As we can see, both lines have the similar trend but the our algorithm

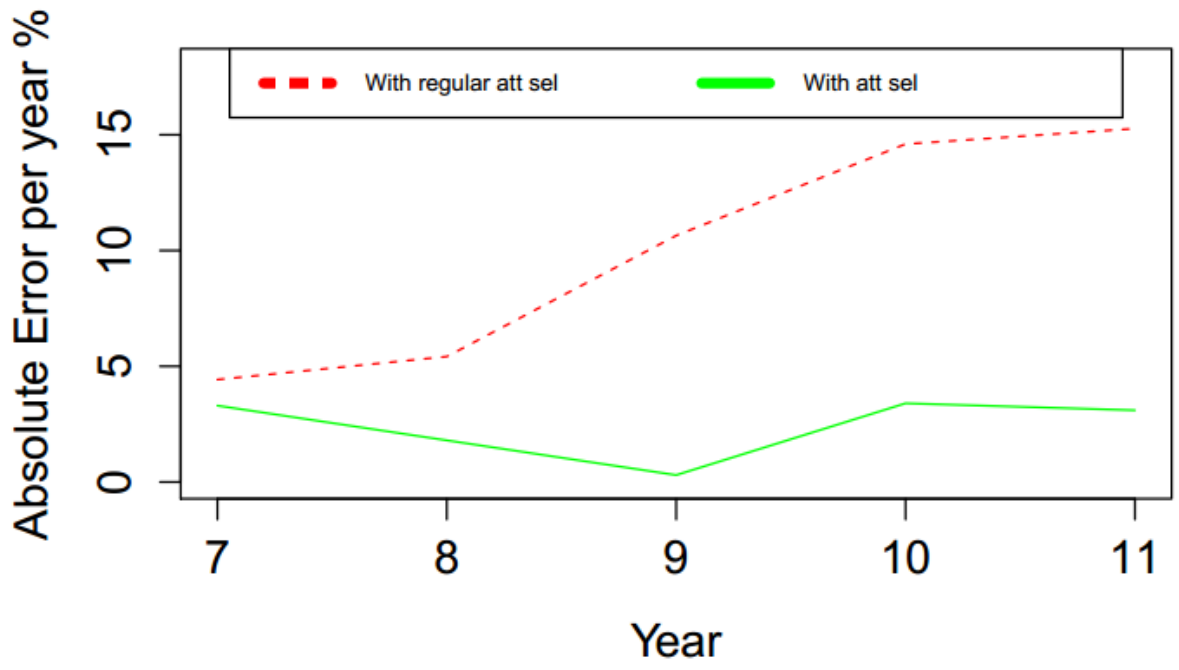

Figure 16. Absolute error percentage of using regular multi-variate attribute selection aglorithms and our algorithm. Both lines have similar trend, but the result from our algorithm has much smaller error percentage.

has much smaller error percentage. The smallest error percentage difference is in 2007, it's about 1.5 percent difference; however, it's about 10 percent difference for the rest of the years. This further shows the effectiveness of the proposed algorithm. The prediction should be better if there were other years of data, as right now there are only five years worth of data, so the training data for the model is very limited. However, with the limited data, the prediction overall is fairly good. The minimum absolute error percentage is 0.16, and the maximum absolute error percentage is 3.4.

CHAPTER 6. CONCLUSIONS

This project demonstrated a new algorithm for multi-variate attribute selection, the algorithm is tested in an agricultural data set. It successfully identifies some significant patterns in the multi-dimensional explanatory attributes and multi-dimensional response attributes. It also ranks the selected attributes in a tree structure by the order of importance. K-L divergence is used in the algorithm for identifying meaningful pattern, and the top resulting attributes are compared with the traditional multi-variate attribute selection results. The selected attributes from traditional methods and the proposed algorithm are evaluated in the same prediction model, the yield prediction from the proposed algorithm has higher accuracy than the other algorithms, and the speed of this algorithm is much faster. Therefore, it is successfully demonstrated in this study that Kullback-Leibler divergence can be used to find distinguishing patterns in a multi-variate data set.

6.1. Future Works

The current project only predicts yield, but many other response attributes or combinations of them remain as viable research interests. For our prediction, a linear model is selected but more models can be explored for improvement in the future. One disadvantage of using five year agriculture data is a shortage of training data, because we want to predict yearly yield, it means that we only have five data points in our prediction model. Therefore, it makes the prediction more difficult. It would be ideal if more data can be collected in the future. In this study sugar beet is the target crop, but this algorithm is not limited to sugar beets; it can be used and improved upon evaluating different crops or any other multi-variate data sets. At this point the time taken to run the algorithm is $O(\mathrm{n})$, it would be ideal to increase the speed so that it only takes $O(\ln(\mathrm{n}))$.

REFERENCES

[1] S. Brin, R. Motwani, and C. Silverstein, *Beyond market baskets: generalizing association rules to correlations*, SIGMOD '97: Proc. of the 1997 ACM SIGMOD Int'l Conf. on Management of Data (New York, NY, USA), ACM Press, 1997, pp. 265–276.

[2] K. Younghee, K. Wonyoung, and K. Ungmo, *Mining Frequent Itemsets with Normalized Weight in Continuous Data Streams*, JIPS 6.1, 2010, pp. 79–90.

[3] K. Jin Oh, S-H. Chung, and Y-M Suh, *Prediction of hospital charges for the cancer patients with data mining techniques*, Journal of Korean Society of Medical Informatics 15.1, 2009, pp. 13–23.

[4] W.M. Campbell, and K. Zahi N, *Simple and efficient speaker comparison using approximate KL divergence*, INTERSPEECH. 2010.

[5] L.D. Baker, and A.K. McCallum, *Distributional clustering of words for text classification*, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1998.

[6] B. Biggi, *Using Kullback-Leibler Distance for Text Categorization*, ECIR 2003, LNCS 2633, 2003, pp. 305–319.

[7] T.M. Cover, P.E. Hart, *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, January 1967, vol.13, no.1, pp. 21–27.

[8] J. Shao, *Linear model selection by cross-validation*, Journal of the American statistical Association 88.422, 1993, pp. 486–494.

[9] Hagan, M.T.: Neural Network Design (Electrical Engineering). Thomson Learning, 1995.

[10] S. Haykin: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall, Englewood Cliffs, 1998.

[11] M. Teittinen, T. Karvonen, and J. Peltonen, *A dynamic model for water and nitrogen limited growth in spring wheat to predict yield and quality*, Journal of Agronomy and Crop Science 172.2, 1994, pp. 90–103.

[12] A. L. Sims, *Sugar Beet production after previous crops of corn, wheat and soybean*, University of Minesota, Northwest Research and Outreach Center, 2006.

[13] A.M. Denton and J. Wu, *Data mining of vector-item patterns using neighborhood histograms*, Knowledge and Information Systems (KAIS) journal, 2009, 173–199.

[14] A.M. Denton, J. Wu, and D. Dorr, *Point–distribution algorithm for mining vector-item patterns*, Proc. 16th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining: Useful Patterns Workshop (Washington, DC), July 2010.

[15] D.W. Franzen, *2 nitrogen management in sugar beet using remote sensing and gis*, GIS applications in agriculture, 2007, 35.

[16] S. Panda, D. Ames, and S. Panigrahi, *Application of vegetation indices for agricultural crop yield prediction using neural network techniques*, Remote Sensing, 2010, no. 3, 673–696.

[17] G. Ruß, *Data mining of agricultural yield data: A comparison of regression models*, Advances in Data Mining. Applications and Theoretical Aspects, 2009, 24–37.

[18] G. Ruß, R. Kruse, M. Schneider, and P. Wagner, *Data mining with neural networks for wheat yield prediction*, Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, 2008, 47–56.

[19] G. Isabelle, A. Elisseeff, *An introduction to variable and feature selection*, The Journal of Machine Learning Research 3, 2003, 1157-1182.

[20] A. S. Moucheshi, E. Fasihfar, H. Hasheminasab, A. Rahmani, and A. Ahmadi, *A Review on Applied Multivariate Statistical Techniques in Agriculture and Plant Science*, International journal of Agronomy and Plant Production. Vol. 4, 127-141, 2013

[21] J. W. Grice, *A Truly Multivariate Approach to MANOVA*, Applied Multivariate Research, Volume 12, No. 3, 2007, 199-226.

[22] French, Aaron, et al, *Multivariate analysis of variance (MANOVA)*, Retrieved January 18, 2002: 2009.

[23] Q. Wang, S.R. Kulkarni, and S. Verdu, *A nearest-neighbor approach to estimating divergence between continuous random vectors*, 2006 IEEE International Symposium on Information Theory, 2006, pp. 242–246.

[24] H.P. Yan, M.Z. Kang, P.D. Reffye, and M Dingkuhn, *A Dynamic, Architectural Plant Model Simulating Resource-dependent Growth*, Annals of Botany, 2004, pp. 591-602.

[25] E. Heuvelink, *Evaluation of a Dynamic Simulation Model for Tomato Crop Growth*, Annals of Botany, 1999, pp. 413-422.

[26] D.W. Shin, G.A. Baigorria, Y.-K. Lim, S. Cocke, T.E. LaRow, James J.O'Brien, and James W. Jones, *Assessing Crop Yield Simulations with Various Seasonal Climate Data*, Proc. 7th NOAA Annual Climate Prediction Application Science Workshop (Norman, OK), October 2009.

[27] P. C. Doraiswamy, B. Akhmedov, L. Beard, A. Stern, and R. Mueller, *Operational prediction of crop yields using modis data and products*, International Archives of Photogrammetry Remote Sensing and Spatial Informatioon Sciences, 2007.

[28] B. Bigi, *Using Kullback-Leibler Distance for Text Categorization*, ECIR, 2003. pp. 305-319.

[29] P.J. Moreno, P.P. Ho, N. Vasconcelos, *A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications*, 2003

[30] U.S. Geological Survey, *Satellite images - landsat*, http://glovis.usgs.gov.

[31] GRASS Development Team, *Geographic resources analysis support system (grass gis) software*, http://grass.osgeo.org, 2008.

[32] R Development Core Team, *R: A language and environment for statistical computing*, http://www.R-project.org, 2008, ISBN 3-900051-07-0.

[33] National Weather Service, *Precipitation data*, http://water.weather.gov/precip/download.php.