VECTOR-VECTOR PATTERNS

FOR AGRICULTURAL DATA

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

By

Eric William Momsen

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

Major Department:
Computer Science

April 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

VECTOR-VECTOR PATTERNS FOR AGRICULTURAL DATA

By

Eric William Momsen

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:

Dr. Anne Denton
Chair

Dr. Kendall Nygard

Dr. Dean Knudson

Dr. David Franzen

Approved:

4/4/13
Date

Dr. Brian Slator
Department Chair

ABSTRACT

Agriculture is increasingly driven by massive data, and some challenges are not covered by existing statistics, machine learning, or data mining techniques. Many crops are characterized not only by yield but also by quality measures, such as sugar content and sugar lost to molasses for sugarbeets. The set of features furthermore contains time series data, such as rainfall and periodic satellite imagery. This study examines the problem of identifying relationships in a complex data set, in which there are vectors (multiple attributes) for both the explanatory and response conditions. This problem can be characterized as a vector-vector pattern mining problem. The proposed algorithm uses one of the vector representations to determine the neighbors of a randomly picked instance, and then tests the randomness of that subset within the other vector representation. Compared to conventional approaches, the vector-vector algorithm shows better performance for distinguishing existing relationships.

ACKNOWLEDGMENTS

This work has been completed in no small part due to guidance from Anne Denton and my committee, discussions with John Xu and my other fellow students, and the encouragement of Jenni Momsen and my family. Thank you everyone who has been a part of this effort.

This material is based upon work supported by the National Science Foundation Partnerships for Innovation program under Grant No. 1114363. The work also had support from the NDSU-Industry Consortium, especially the American Crystal Sugar Company who providing a significant portion of the data used in this analysis. Additional data used in the work is made available by the U.S. Geological Survey and the National Oceanic and Atmospheric Administration.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Agricultural applications provide a rich and complex data source, and existing data mining techniques are insufficient for addressing the complexity. Farmers are not interested in just the yield they achieve as quantified by the weight of the crop, but also in its quality, such as protein content for wheat, or sugar content and the sugar lost to molasses in sugarbeets. That means that crop production performance in sugarbeets can be measured with a three-dimensional vector. Weather significantly affects crop development, and the associated time series can also be viewed as a multi-dimensional vector, possibly after applying dimensionality reduction techniques. Satellites and hand held sensors can be used to periodically measure the 'greenness' of plant canopies, to get a series of snap shots of the apparent plant health. Many statistical, machine learning, and data mining techniques exist that use multi-dimensional vector data for predicting a single categorical or continuous attribute, but they lack the ability of testing relationships of vector data to other vector data. Yet, in order to find interesting patterns and understand dimensionality reduction and other preprocessing questions, it is important to use the full available information.

1.1. Problem Statement

The agricultural data used in this study, in contrast to many data mining problems, has multiple characteristics described by multiple dimensions. In the agricultural domain, it is critical to observe that for some crops the importance of quality may be similar to that of quantity. For sugarbeets, increasing the sugar content (quality) may be even more important than one would estimate by calculating the total weight of sugar gained from the yield (quantity): sugarbeets that are low in sugar have a correspondingly higher water content, and energy has to be expended

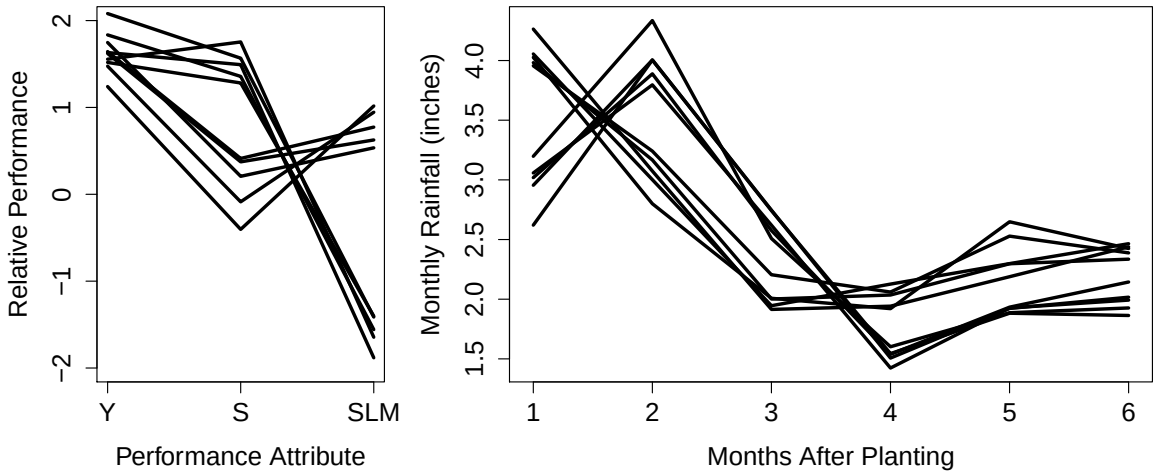

Figure 1. Toy data suggesting a potential pattern to be found in agricultural data. In the left panel, each line represents the yield, sugar, and sugar lost to molasses measured in a single field. In the right panel, each line represents monthly rainfall totals estimated for a particular field.

to remove that water. Conversely, just looking at sugar content alone would be insufficient, as a crop with high sugar content but very poor yield may not be worth the harvesting expense. Toy values for these 3 performance metrics are plotted in the left panel of Figure 1. The right panel shows toy data for a time series of monthly rainfall totals. In order to make more accurate end of season crop performance predictions, it would be interesting to know if there is a relationship between particular performance results and rainfall patterns.

Traditional statistics and machine learning techniques are able to capture relationships between vector data and one-dimensional categorical or continuous attributes. When attempting to relate two vectors to each other, it is important not to lose the multi-dimensional nature of the problem. Figure 2 illustrates the problem of determining significant relationships based on one-dimensional projections. In many situations, data is normally distributed (or relationships are very weak) with respect

to any single attribute. A projection to either axis shows both the $\times$ and the $\circ$ data points as having a normal distribution. In two dimensions, however, it can easily be seen that $\times$ and $\circ$ data points are not evenly distributed. This kind of setting is well-known in classification problems. Many techniques use multi-dimensional input to predict a value or class label for new instances. Using information from multiple attributes (a vector) is a proven method for finding interesting relationships in a complex data set.

When approaching a dataset as shown in Figure 1, the key characteristic to notice is that there is a multi-dimensional vector of attributes describing both the explanatory variables of rainfall and the response variables of crop performance. Meeting the challenge of a multi-dimensional vector of explanatory attributes can be met with existing methods. Yet the techniques would be limited in their application to the crop performance attributes. One approach could be to build a model for each of the performance attributes individually, e.g. predict the Yield (Y) based on the monthly rainfall after planting. Alternatively, the three dimensions could be projected into a single dimension. With either approach, information is discarded, and the risk of missing patterns as shown in Figure 2 is raised. Techniques have been developed to take into account all of the information in vectors of explanatory variables to find patterns, there is a similar need to develop techniques that can find patterns in a search space described by a multi-dimensional vector of explanatory *and* a second multi-dimensional vector of response attributes.

Figure 3 illustrates the solution. The two left panels can be viewed as the vector of crop performance data, with one quantity and one quality attribute. The two right panels can represent the vector of rainfall data, simplified for this example to 3 attributes. Typical statistical and data mining techniques could describe either response variable a *or* b in terms of explanatory attributes 1, 2, and 3. The proposed

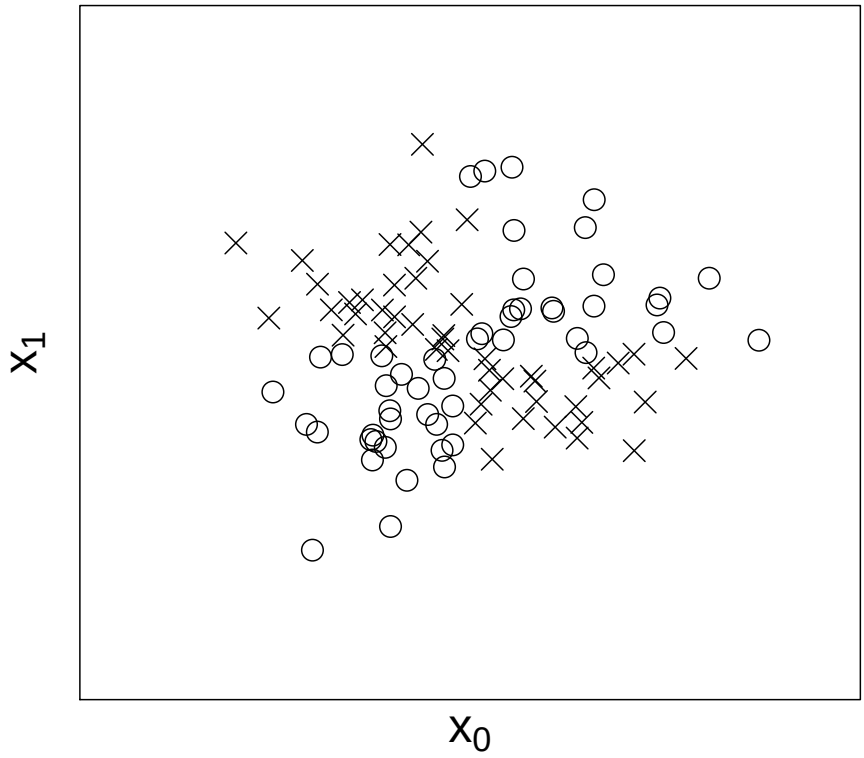

Figure 2. Example of the benefit of considering information from multiple dimensions (well-known from classification problems): The data are normally distributed in both explanatory dimensions, yet there is a clear pattern when both dimensions are considered together.

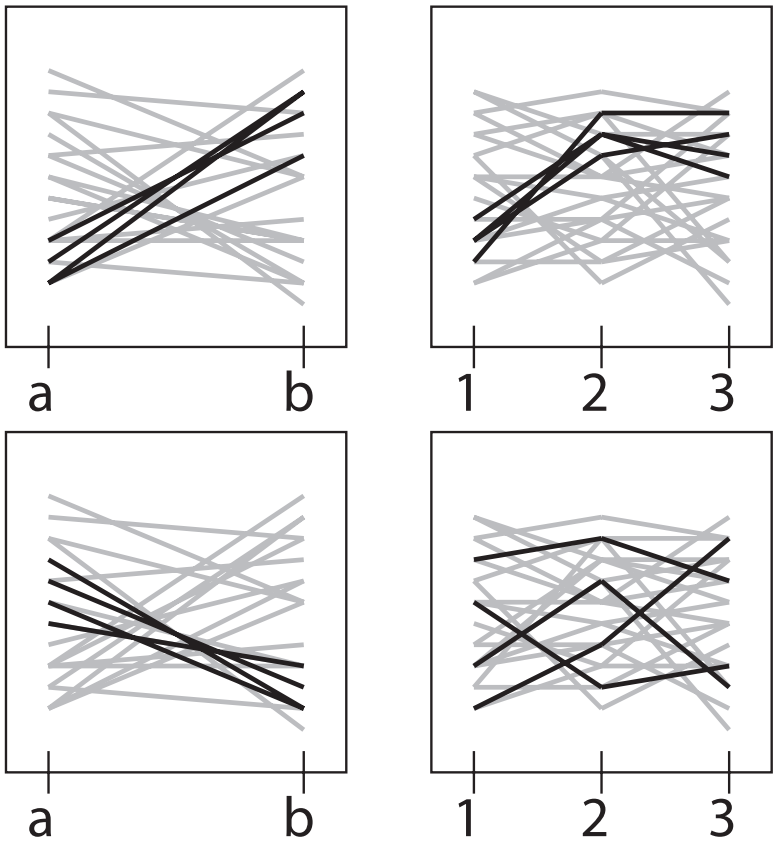

Figure 3. Conceptual question to answer using vector-vector pattern mining. If a set of instances is similar based on attributes a and b, will those same instances show similarity based on attributes 1, 2, and 3? In this toy data, there is a clear pattern in the top row but not in the bottom row.

algorithm seeks to find a relationship that is based on both a *and* b. Instead of seeking to evaluate the relationship between each individual response attribute and the explanatory attributes, an intermediate classification is first determined by selecting a set of instances that are similar according both attributes a and b. Then, using that classification as a new attribute, the distribution of data described by the explanatory attributes is evaluated according to vector-item pattern mining techniques that were adapted from [5, 6].

CHAPTER 2. RELATED WORKS

From the data mining perspective, finding a pattern in a single group of attributes is related to frequent pattern mining. Frequent pattern mining algorithms were originally developed for item sets [1] and were then generalized to include continuous data [3, 23, 19, 4]. These techniques do not, however, consider two groups of continuous attributes jointly. Recent data mining techniques have generalized such approaches to complex feature spaces that include time series and other multivariate data. One such technique is vector-item pattern mining [5, 6]. Item sets are used to define a subset of the vector data, and the distribution of the vector data in that subset are then compared with the full set of instances or with those instances that do not have the item. Similar approaches in bioinformatics are known as gene set enrichment analysis [27, 26], but lack the full multi-dimensional treatment.

A standard approach to comparing distributions is the Kullback–Leibler divergence, which was used in [6]. The current work also includes an approach that calculates the KL divergence based on nearest neighbors [32, 18], because of the availability of highly optimized code. The contribution of the current work is an extension to data mining of patterns between vectors.

An important problem in agriculture is the prediction of yield and other response variables. The traditional statistics approach is to fit the data to a linear model, using some set of available explanatory attributes. Normalized Difference Vegetation Index (NDVI) and surface temperature have been used to predict county and state wide yields [9]. Previous studies have found that using data mining techniques that deal well with a vector of explanatory variables can lead to more successful yield predictions [20]. Neural networks, decision trees, and support vector regression have all been applied to predicting crop yields. Significant work has also been directed towards dynamic models of crop growth [24, 25]. Dynamic models focus on predicting

the growth of individual plants, based on an input of many parameters expected to have an impact. To the best of our knowledge, previous work has focused on the ability to predict a total yield without regard to crop quality. The current work provides a method to include both quantity and quality measures in the pattern finding effort.

In the context of lengthy time series, much work has been done to appropriately reduce the number of dimensions before applying analysis techniques. In the case of rainfall, this typically involves a total rainfall for the growing season or counting the number of rainy days in a season. Additional statistics, such as Precipitation Concentration Index can be used to capture intra-seasonal variability [2, 16, 15]. But such summary numbers still lose information about *when* during the growing season the precipitation occurred. It has already been demonstrated that the availability of more data from over the course of the growing season can improve the yield predictions from neural networks [21]. Furthermore, it has been demonstrated that more accurate rainfall predictions can improve dynamic crop model yield predictions, as water stress (indicated in part by wet and dry periods) has a larger impact on plant growth then seasonal rainfall totals [7]. The benefit of the proposed approach is that the impact of different time series processing strategies can be quantitatively compared.

Another approach frequently used to assess plant health is to calculate a vegetation index, such as the Normalized Difference Vegetation Index (NDVI) [17]. NDVI can be obtained on a periodic basis from satellites and on demand using sensors or aerial flyovers. It has been demonstrated that NDVI provides an accurate assessment of nitrogen uptake in sugar beets. When tops are left as green fertilizer, the measured NDVI of the sugar beet canopy can be used to determine nitrogen credit for the next year's crop [10]. As with rainfall, aggregating NDVI values can provide more information than using just single data points. The sum of NDVI over a specific

portion of the growing season was useful in linear models predicting county and state wide wheat yields [8]. Others have focused on what time period during the growing season provides the best predictions. One study found that for wheat yields, NDVI at growth stage 7 was better predictor then at NDVI at growth stage 10 [11]. As with rainfall, these efforts have attempted to determine a single attribute that is most affective for future predictions. Using the proposed pattern mining technique will allow additional information to be considered when testing for relationships.

CHAPTER 3. CONCEPTS

In data mining, it is common to consider multiple attributes simultaneously. From a mathematical standpoint, these attributes are treated as dimensions of a vector. The current work considers continuous vector spaces, i.e. D_y attributes $y_j \in \mathbb{R}, 0 \leq j < D_y$ are considered as one "vector" attribute $\mathbf{y}$ with domain $\text{dom}(\mathbf{y}) = \mathbb{R}^{D_y}$. Vector representations are standard in clustering problems and, when combined with a categorical class label, in classification problems. Vector-item pattern mining approaches [6] look for patterns between vector attributes and item data. Item data can be represented as binary attributes $B^{(i)}, 0 \leq i < M$ that represent presence or absence of an item.

In this work, item data is defined as identifying those vectors $\mathbf{x}$ with domain $\text{dom}(\mathbf{x}) = \mathbb{R}^{D_x}$ that are within a neighborhood of a specific vector $\mathbf{x}_i$

$$B^{(i)} = H(d - |\mathbf{x} - \mathbf{x}_i|) \tag{1}$$

where H is the Heaviside step function that is 1 when the argument is positive and 0 when negative. The cutoff d is selected such that the relative support of item i has a predefined value, and that support is varied between 0.01 and 0.99.

For a given cutoff d, the KL-divergence is calculated for the subset P_i of vectors $\mathbf{y}$ for which $B^{(i)} = 1$ with regard to the subset Q_i for which $B^{(i)} = 0$. This definition allows defining one binary attribute for each instance in the database. That means that statements can be made about individual instances, rather than only about vectors as defined over the entire data set. It also means that for any pair of vector attributes $\mathbf{x}$ and $\mathbf{y}$, there are as many KL divergence values defined, for a single threshold, as there are instance under consideration. The ability of deriving many KL divergence values is used for testing the algorithm and for evaluating parameters:

Even for vectors that are relatively weakly related, it can be expected that KL divergences are typically larger than for randomly selected subsets.

If the vector pair is strongly related, it is expected that the KL divergence of vectors P with regard to Q is always larger than for a random subset. If the vector pair is not as strongly related, it is expected that some of the KL divergency values of randomly selected vectors may be larger than for vectors that are defined based on a neighborhood of a point $\mathbf{x}_i$. Using the KL divergences as a measure of confidence that a relationship exists, the area under the receiver operating characteristic curve (ROC) is calculated for the set of KL divergences. If a strong relationship exists, it is expected that area should approach 1. The larger the ROC, the stronger the relationship is expected to be.

In [5] it was shown that the significance of classification results can be used as an alternative to a comparison of distributions. In other words, the ability of predicting $B^{(i)}$ from P and Q is an alternative route towards a confidence measure that there is a relationship between vector attributes $\mathbf{x}$ and $\mathbf{y}$. The information gain of the contingency table of the prediction is analogous to the KL divergence, although mathematically the two are not fully equivalent, in that the information gain compares the information with knowledge of the classification result with the information without that knowledge, while the KL divergence tests the distribution of one subset ($B^{(i)} = 1$) against the other ($B^{(i)} = 0$).

Both KL-divergence and classification-based approaches use the full vector information in $\mathbf{x}$ and $\mathbf{y}$. Typical traditional statistics approaches compare distributions between individual continuous attributes. Such approaches miss information such as represented in Figure 4. After briefly discussing some terminology, we will then consider the practical implications of these three types of approaches.

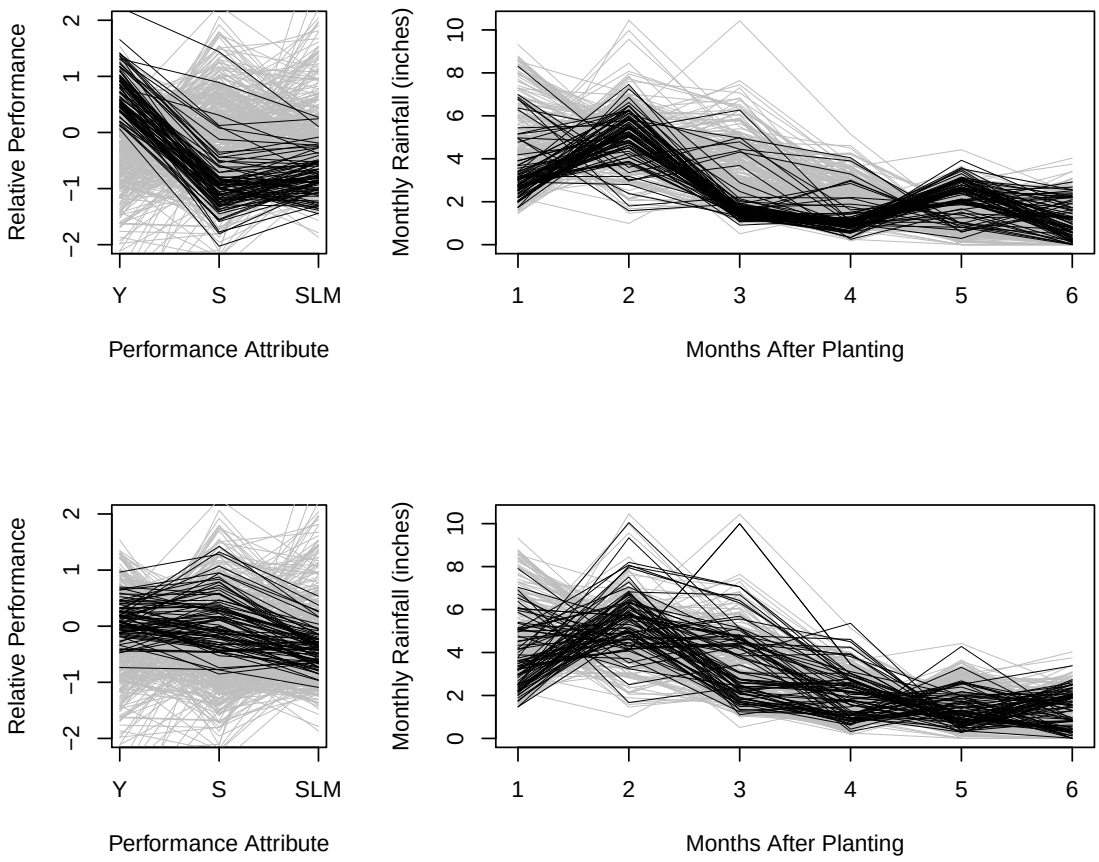

Figure 4. Examples analogous to Figure 3 but using real data: On the left side yield, sugar and sugar lost to molasses are plotted with a subset of related records being rendered in black. The grey background indicates the range of data for the remaining records. On the right hand side the monthly rainfall totals are plotted with the same subset again rendered in black. The distribution comparison indicated a strong pattern for the top row and a weak pattern for the bottom row.

3.1. Terminology

When labeling each of the vectors throughout this paper, the terms explanatory and response are often used. Explanatory variables are also commonly referred to as independent variables. In the two applications, rainfall and INSEY are considered as the explanatory variable. For the vector-vector algorithm, the vector of explanatory variables is used for the distribution comparison step. Response variables are also commonly referred to as dependent variables. For the vector-vector algorithm, the vector of response variables is used for the initial neighborhood selection step. This usage was selected since the length of our response variables vector was shorter than the length of the explanatory variables. The nearest neighbor selection according to Euclidean distance is more suitable for lower numbers of variables.

The precipitation data used for this study includes all forms of precipitation, e.g. rain, snow. For the growing season in North Dakota, the precipitation is typically rainfall. Thus precipitation and rainfall are often used interchangeably throughout this paper.

3.2. Distribution Comparisons

The first step is to carry out a neighborhood selection in the space of vectors $\mathbf{x}$. In a normalized, low-dimensional vector space, Euclidean distance is an effective means to determine similarity. The crop performance data that will be used in the evaluation consists of only 3 dimensions (yield, sugar, sugar lost to molasses). Thus, neighborhood membership determined by selecting the nearest neighbors as determined by Euclidean distance is a suitable choice for identifying a subset of instances.

A critical aspect of the vector-vector pattern mining algorithm is to compare the distribution of vector data associated with a subset of the instances (P) with the

remainder of the instances (Q). The subset (P) has already been defined based on the first vector of data. Examining the data in a second vector of attributes, it must be determined if there is a difference between the subset of instances P and Q. A number of approaches are possible, the following were evaluated:

3.2.1. KL Divergence

The Kullback–Leibler (KL) divergence can be used to evaluate the difference between two distributions. A standard KL divergence [13] has been calculated using 10 neighbors. A Voronoi cell based KL divergence has also been calculated.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} ln(\frac{\mathrm{d}P}{\mathrm{d}Q})\,\mathrm{d}P \tag{2}$$

Since most of our calculations are done for relatively sparse data, the following alternative approach towards modeling the probability distributions that are used in the KL divergence can be considered: The probability is modeled as being constant within Voronoi cells that are seeded by those instances that have a particular item. Figure 5 shows the Voronoi construction. The Voronoi cell for seed S is defined to contain all those instances that are closer to S than to any other seed. Probabilities are based on the total number of points in each cell. In practice, this means that, for every instance that does not have the item, the closest instance with item is being determined. The probability of instances with the item is taken to be 1 divided by the volume of the cell. Figure 5 shows how Voronoi cells may have widely differing numbers of instances when the distribution of instances with the item differs from the overall distribution.

3.2.2. Information Gain

Treating the membership in the subset (P) as a class label (as determined by the first data vector), now a multi-dimensional prediction model can be used to determine if the data in the second vector could be used to predict the class

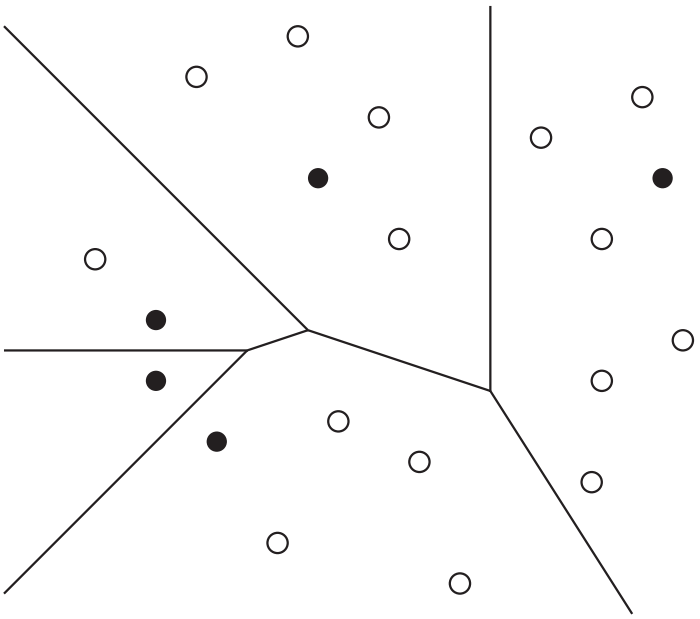

Figure 5. Voronoi construction for estimating probabilities in the KL divergence.

label. If such a predictive model is possible, it has been demonstrated that there is an underlying pattern between the two vectors. The strength of the predictive model can be evaluated by calculating the information gain (entropy - conditional entropy, Equation 5) from the confusion matrix. The confusions matrix tallies the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Many evaluations of the confusions matrix are possible (e.g. Accuracy, Precision), information gain was selected since it takes into account all 4 values of the contingency matrix and it also is analogous to the KL divergence. For this study a Naive Bayes model [14] and a Nearest Neighbors (k=10) [31] model are used for making the predictions. 2/3 of the data were used for the model generation (or as a basis for future predictions) and the remaining 1/3 of the data were used for testing.

$$H(T) = -\left(\frac{TP+FN}{total}\right)\log\left(\frac{TP+FN}{total}\right) - \left(\frac{TN+FP}{total}\right)\log\left(\frac{TN+FP}{total}\right) \tag{3}$$

$$H(T|a) = -TP\log\left(\frac{TP}{TP+FP}\right) - FP\log\left(\frac{FP}{TP+FP}\right) - FN\log\left(\frac{FN}{FN+TN}\right) - TN\log\left(\frac{TN}{FN+TN}\right) \tag{4}$$

$$Information\ Gain = H(T) - H(T|a) \tag{5}$$

3.2.3. Single Attribute

For evaluating the importance of using multi-dimensional information, two standard one-dimensional statistics tests commonly used for distribution comparisons have been selected for comparison purposes. At each time step in the time series vector, a Welch Two Sample t-test and a Two-sample Kolmogorov-Smirnov (KS) test are used to compare the distribution between the P and Q subsets. The Welch t-test is related to the Student t-test, with the addition that the variances of the two populations are not assumed to be equal. It is a test to determine if the means of the populations are equal. The KS-test is a non-parametric test, and compares the distribution of the two populations, thus it is sensitive to more than just the mean value of the two populations. After evaluating the statistical test at each time step, the mean p-value is used to evaluate the effectiveness of the test.

3.3. Evaluation Technique

3.3.1. Receiver Operating Characteristic

The final step in the process is to determine if a pattern has been distinguished by the test. A number of concerns must be met. First, the single attribute distribution comparison methods produce a p-value, but with the number of variables being tested the risk of over fitting the data is high. Second, the multi-dimensional distribution comparisons do not have a standard confidence evaluation. Finally, the results from each of the 6 methods need to be compared with each other. A typical approach to answer the question of pattern presence is to determine if the result could arise from a random set of data. The described neighborhood selection and distribution comparison is repeated for a number of starting instances. Next, a set of baseline comparison data is generated. This is done by creating a "neighborhood" of random instances, then using this subset for the distribution comparison step. The process using random neighborhoods is repeated the same number of times as it was repeated for the calculated neighborhoods. If the distribution comparison provides similar results for both the original groups and these "random" groups, there is no evidence of an underlying pattern. To make this determination, the area under the receiver operating characteristic curve (ROC) is calculated. Both groups of distribution comparisons (some from nearest neighbor and some from random neighborhoods) are now ordered together and the ROC calculated. If no pattern is present or can not be detected by the algorithm or distribution comparison, the nearest neighbor and random results will be randomly ordered, and the ROC will approach 0.5. An ROC approaching 1 indicates neighborhood defined by nearest neighbors always showed a stronger distribution difference as compared to random neighborhood selection. Intermediate values indicate that a pattern was found only for a portion of the query points or the distribution comparison test was unable to detect it in every case. As

different data preparation and algorithm parameters were varied, the changes in the ROC were used to evaluate the effect of the parameter.

3.3.2. Linear Regression

A practical method to test the usefulness of information discovered using the vector-vector pattern mining technique is to check if it improves the accuracy of future predictions. A linear model was built to predict yield as a function of other weather affects (e.g. growing degree days) and farm management decisions (e.g. cover crops, ridge tilling, starter fertilizer). This baseline model did not include in season rainfall or satellite data. To test the effect of any particular variable under study, additional models were built. For each particular value in the experiment, 5 models were built. For each of the 5 models, a different year of data was left out of the training data. Typically, the correlation coefficient (R^2) values is used to evaluate the linear model. This gives a measurement of the accuracy for each particular field. For sugarbeet farming, it is also important to know what the total yield will be in the growers' co-op, since the processing plant capacity may limit the weight of beets that can be harvested. The overall accuracy (Equation 6) of the linear model in predicting the total sugarbeet harvest is calculated by totaling the predicting tons harvested from each field for the year left out of the model building. The mean of the 5 accuracy values is then used to evaluate the experiment.

$$\text{Accuracy} = 1 - \frac{|\text{actual regional yield} - \text{predicted regional yield}|}{\text{actual regional yield}} \tag{6}$$

CHAPTER 4. ALGORITHM

The goal of our vector-vector pattern mining algorithm is to relate a vector of continuous explanatory variables to a vector of continuous response variables. The first major step is to define a neighborhood (P) based on the Euclidean distance of one vector of data. The second major step is to compare between P and Q (the remaining instances) the distribution of the second vector of data. After the algorithm is complete, the results are evaluated.

Algorithm 1 was implemented in R [30], an open source programming language typically used for statistics calculations.

The first step is to start with a selection of seed instances from the overall data set. For each of these seed instances, a neighborhood of interest (P) is identified. For this implementation, the nearest neighbors based on Euclidean distance of the 3 response variables was used, except for the evaluation of the effect of response variables on the effectiveness of the algorithm. The size of P was held constant at 10% of the total instances, except for the portion of the study where the neighborhood size was examined. The remaining instances are assigned to Q. Next vector-item pattern mining techniques are applied, using the rainfall time series as the vector data and membership in P as the item data. For purposes of evaluating the robustness of the proposed algorithm, a variety of tests are used to compare the distribution of the explanatory variables (rainfall time series) between P and Q. After making these calculations for each of the seed instances, the process is repeated for random data. Instead of using Euclidean distance to find nearest neighbors, the subset P was defined by taking a random sample from the data set. The total number of points in P was held constant for both parts of the analysis. The distribution comparisons were then calculated based on the random subsets.

Algorithm 1: Vector-Vector Pattern Mining Algorithm

Data: U /* all instances */
Data: *seeds* /* starting instances */
Data: *response* /* crop performance */
Data: *explanatory* /* time series, rainfall or INSEY */
Result: *dist_comp* /* distribution comparison */

1 **foreach** $seed \in seeds$ **do**
2 P_i = NearestNeighbor(*response*,*seed*);
3 $Q_i = U \setminus P_i$;
4 **foreach** $test \in tests$ **do**
5 $distribution\_comp_{test}$ = evaluate(test, P_i, Q_i, *explanatory*);
6 **for** *i = 1 to* $|seeds|$ **do**
7 $randomP_i$ = randomSubset(*response*);
8 $Q_i = U \setminus randomP_i$;
9 **foreach** $test \in tests$ **do**
10 $distribution\_comp_{test}$ = evaluate(test, $randomP_i$, Q_i, *explanatory*);
11 **return** *dist_comps*

The algorithm steps are illustrated in Figure 6. The first panel renders all of the data in the crop performance vector as grey lines. Each line indicates one instance, connecting the yield, sugar, and sugar lost to molasses reported for a particular sugarbeet field. The dashed line is the seed instance randomly selected for one iteration through the first loop of the algorithm. The second panel adds the information calculated during step 2 of the algorithm, the black lines indicate the particular sugarbeet fields with crop performance that are the most similar to the seed instance. Those lines are the instances assigned to the set P. In the last row, the vector of rainfall data is now considered, and is shown in the right hand panel. Data for the instances in the set P are again rendered in black; the grey lines indicate the remaining data that are assigned to the set Q. A pattern is visually apparent; the black lines are not evenly distributed among the grey lines. Step 5 of the algorithm numerically compares the distribution between P and Q.

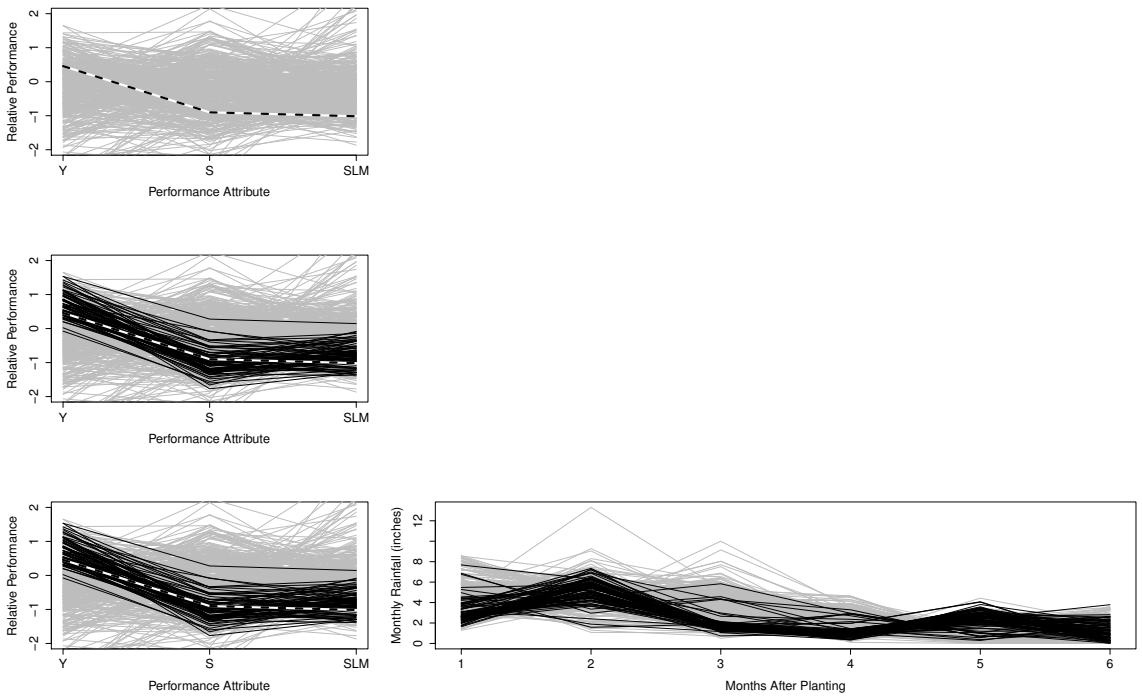

Figure 6. Illustration of the steps in the vector-vector algorithm. Performance Attributes are Yield (Y), Sugar (S), and Sugar Lost to Molasses (SLM). Top row: Step 1, a particular seed instance is randomly chosen. Second row: Step 2, the nearest neighbors of the seed instance, according to the crop performance (response) attributes. Third row: Step 5, the right panel highlights the corresponding instances that will be used for the distribution comparison.

CHAPTER 5. DATA PREPARATION

Before proceeding to the results, the data set used for the study will be described and the processing steps required to bring data from many disparate sources together into a cohesive format that can be easily examined will be discussed. The vector-vector pattern mining algorithm was applied to a data set compiled to study sugarbeet agriculture in the Red River Valley. This includes confidential data from American Crystal Sugar Company augmented by public data from the U.S. Geological Survey Landsat satellites and the National Oceanic and Atmospheric Administration (National Weather Service). The data was collected from 2007 to 2011. The spatial range covers the Red River Valley of the North, in eastern North Dakota and north western Minnesota. The precipitation data and the rest of the public data have been preprocessed and associated to the correct fields using GRASS [29] GIS software.

It is important to recognize that instances from a particular growing season are highly correlated, due to experiencing similar weather patterns. If training data is selected at random from the entire dataset, any patterns found can typically be attributed to the hidden variable of the calendar year. This difficulty has been resolved using two different approaches. For the vector-vector pattern mining algorithm, if multiple years are used at one time, the neighborhood P is dominated by instances from a single growing season. To avoid this, the vector-vector approach is applied to only single years of data at a time. The technique is applied separately to each growing season, and the results are averaged together. A second approach is used when making future predictions with the linear model. Data is from an entire growing season is withheld from the training data, and then use the withheld year's data as testing data. This procedure is also repeated for each growing season, and the results averaged together.

5.1. Field Data

Substantial data about their fields and their growing practices are collected by farmers. The American Crystal Sugar Company has compiled information from member farmers, including the geographical location of each farm. Some attributes are categorical in nature, such as what seeds were planted, the previous year's crop, and the soil type. Other attributes are continuous, such as the fertilizers applied, acreage harvested, and actual nitrogen levels measured in the soil. For vector-vector pattern mining, three continuous attributes pertaining to the crop performance are used. Yield is a measurement of total mass harvested from the field. Sugar is a measurement of how much of the yield is sugar content. Sugar lost to molasses is a measurement of the sugar lost during the beet processing. High yield, high sugar, and low sugar lost to molasses are the desirable outcomes. All farm data is at field level resolution. For purposes of this study, each attribute was z-normalized to allow for direct use of Euclidean distances between the records in the vector of crop performance attributes. Additional grower practice variables were selected and used for the linear model. Since that portion of the data is confidential, the specific variables used in the base line linear model are deliberately left unstated.

5.2. Rainfall Data

Weather data pertaining to the study area is publicly available from the National Weather Service (NWS) of the National Oceanic and Atmospheric Administration. Precipitation data [22] is preprocessed by the NWS using radar estimates biased with rain gauge measurements. The data is provided as a grid of vector points with data points roughly 4km apart. Figure 7 highlights that the sugarbeet fields are much smaller than the grid spacing. The vector format does not lend itself well to large scale distance calculations: to find the rainfall vector point that is the nearest

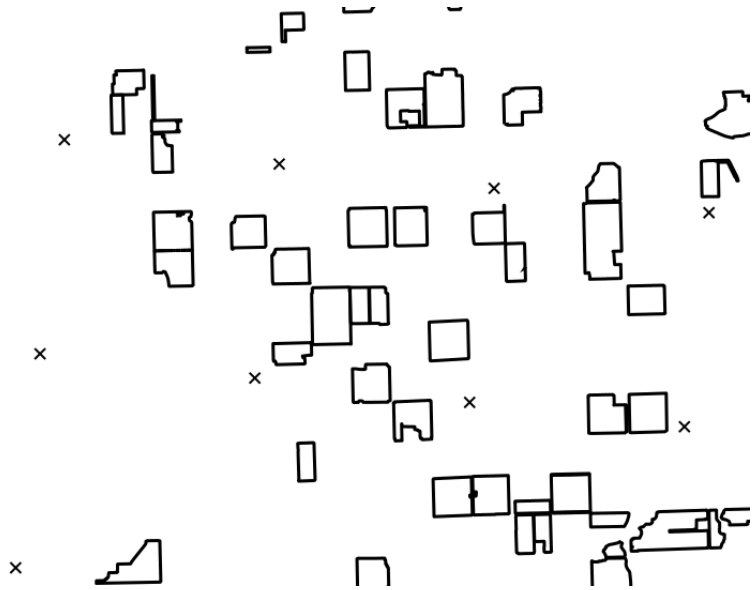

Figure 7. Physical scale of the precipitation and field data. The × represent precipitation data interspersed among sugar beet fields.

neighbor to a particular field, the distance between the field and every vector point must be calculated. Converting the vector information to a raster format (illustrated by Figure 8) provides for faster querying. The entire study area is broken into pixels on a resolution to contain one rainfall vector point, and each pixel is assigned the corresponding precipitation value. After this conversion, precipitation can be directly queried based on the latitude and longitude of any particular sugar beet field.

Each map downloaded from the NWS provides the daily total precipitation, so the above process is repeated for each day of the study period. The daily total precipitation values are used as the base time series in our work. The daily rainfall for each particular field was time shifted such that day zero of the series aligned with the day the field was planted. Additional dimension reduction was studied by further aggregating the data to weekly, monthly or other sized bins. When studying other

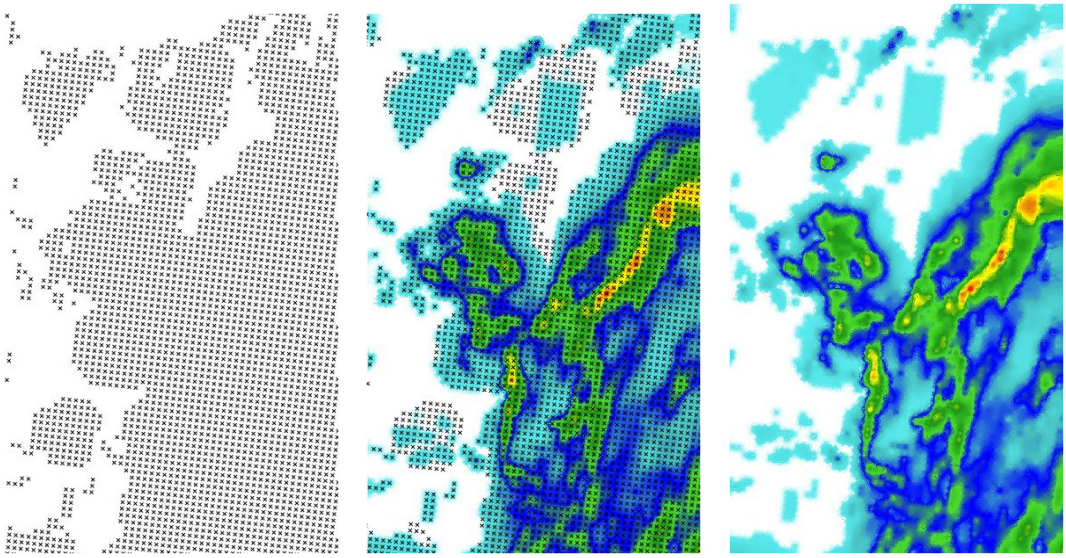

Figure 8. Conversion of the National Weather Service precipitation data from vector to raster data. The figure shows a small portion of the study area. The left panel plots the vector data points, after zero rainfall points have been removed. The center panel overlays the vector data over the raster data. The right panel depicts the final raster.

aspects of the algorithm, a consistent reduction to monthly precipitation totals was used.

5.3. Satellite Imagery

Satellite imagery is available from the Landsat program, image tiles 30026 and 30027 from the Landsat 5 and Landsat 7 satellites [28] have been used. The images were processed using standard techniques available in GRASS GIS, including converting the raw data to reflectance, identifying cloud cover, calculating the NDVI according to Equation 7, merging the two tiles, and excluding the cloud cover. Mean NDVI values were calculated for each field area, excluding the outer 60m perimeter (2 pixels) to avoid edge effects. This pre-processing work was completed by other members of our research group. Figure 9 gives a visual representation of the study area including demarkation of sugarbeet fields.

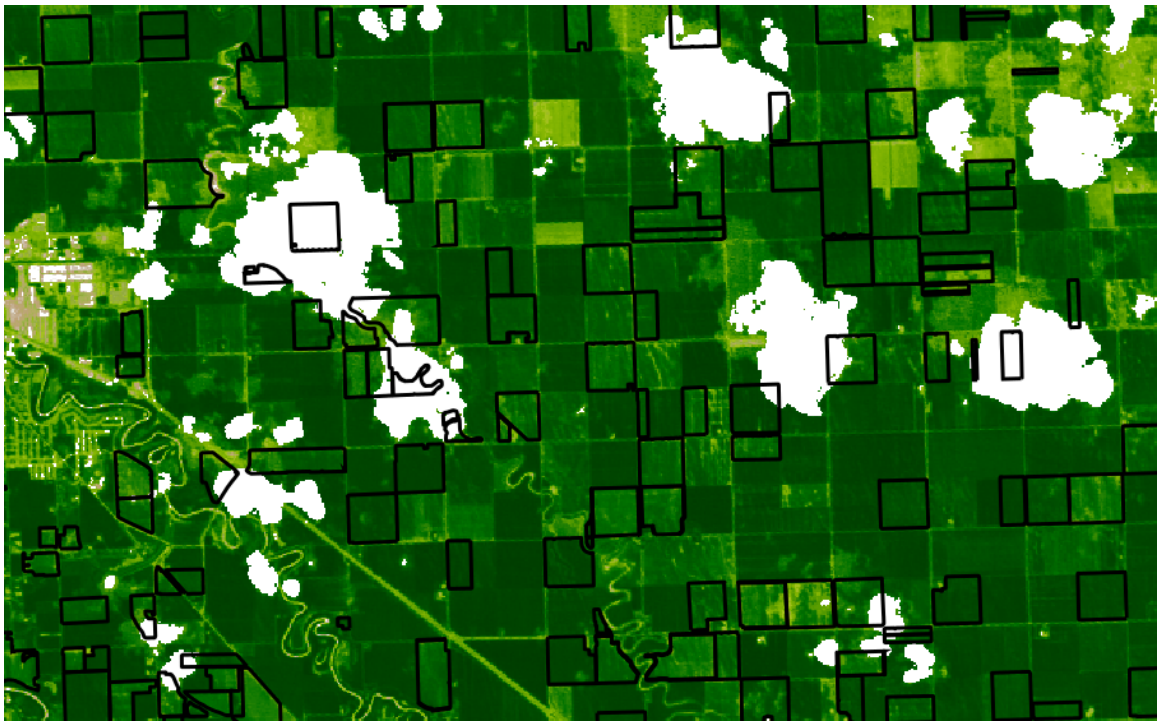

Figure 9. Calculated Normalized Difference Vegetation Index (NDVI) over a portion of the study area. Sugarbeet fields have been marked with black borders. Darker green corresponds to higher NDVI values. The large white areas are cloud cover when the Landsat image was taken.

$$\text{NDVI} = 1 - \frac{\text{near infrared} - \text{visible red}}{\text{near infrared} + \text{visible red}} \tag{7}$$

Next, the time series data was compiled for use in this study. Each of the Landsat 5 and 7 satellites passes over a particular location on 16 day intervals. Both series were used for this study, so the time series is at roughly weekly intervals. The time series for each field has numerous gaps, when clouds covered the field when the satellite image was taken. Any field with 10 or fewer NDVI values in the time series was excluded from the study. For the remaining instances, missing values were filled in by linear interpolation. If the missing value was at the beginning or end of the series, the nearest existing value was used for extrapolation. The time series of NDVI values for each field was constructed starting from the day the field was planted.

To compare NDVI values taken at different times of the year, and from different growing seasons, the timing of the image must be accounted for. In general, NDVI will increase at the beginning of the year. With a dry fall, the NDVI my decline towards the end of a growing season. To compare values from different times in the growing season, two approaches are generally used. The easiest approach is to divide the NDVI value by the number of days between planting and when the image was taken. A more nuanced approach takes into account the accumulated heat the crop has experiences. This provides for an estimation of the plant growth stage. The accumulated heat is referred to as the Growing Degree Days (GDD). The GDD are calculated for each day, and then totaled for the period of interest. GDD is based on the minimum and maximum daily temperatures, with some crop dependent baseline parameters. Temperatures are adjusted according to equations 8 and 9, and then GDD is calculated according to equation 10. All temperatures are in units of degree Fahrenheit.

$$\mathrm{T}_{min} = Max(\text{Actual Daily Min Temperature, } 34\ ^\circ F) \tag{8}$$

$$\mathrm{T}_{max} = Min(\text{Actual Daily Max Temperature, } 86\ ^\circ F) \tag{9}$$

$$\text{GDD}\ ^\circ F = \frac{\mathrm{T}_{max} + \mathrm{T}_{min}}{2} - 34\ ^\circ F \tag{10}$$

For this study, daily temperature data was prepared to allow GDD to be calculated at any point during the season. The final step in the NDVI processing is to account for the plant growth stage at the time of the image, as estimated by the GDD. This value is often referred to as the In-Season Estimate of Yield (INSEY) value. INSEY is calculated by dividing the NDVI value by the number of growing degree days the field has accumulated from planting to the day the satellite image was taken, according to equation 11.

$$\text{INSEY} = \frac{\text{NDVI}}{\text{GDD}} \tag{11}$$

CHAPTER 6. RESULTS - RAINFALL

The results of the vector-vector pattern mining algorithm have been evaluated by comparing its performance against single variable tests. In this section, a visualization of the patterns discovered with the algorithm has been provided. To better understand the accuracy of the algorithm the performance is evaluated while varying the number of attributes in the vectors as well as the size of the neighborhood subset. This includes using single attributes to confirm that using multiple attributes is advantageous. The speed impact of considering multi-dimensional data is also addressed. Finally, the findings are applied to linear models predicting future harvests to determine if the model accuracy could be improved.

6.1. Example Pattern

Figure 10 shows an example pattern that can be found for the data set. In this example some portions of the rainfall time series can be linked to a specific growing result and not be relevant to others. In the left panel the black lines correspond to a neighborhood of instances having high yield with low sugar content and low amounts of sugar lost to molasses. The first 2 months of rainfall exhibit a tight range, indicating this portion of the time series was important for high yielding crops that were also low in sugar. In the right panel, the black lines correspond to a neighborhood of similar instances having average yield with high sugar content and average sugar lost to molasses. The 3rd and 4th months show a tight range, indicating this portion of the time series correlates to high sugar content with average yield. Those same periods of rainfall exhibit no patterns for the other crop performance. These patterns only become apparent when the multi-variate nature of both the rainfall and the crop performance is considered.

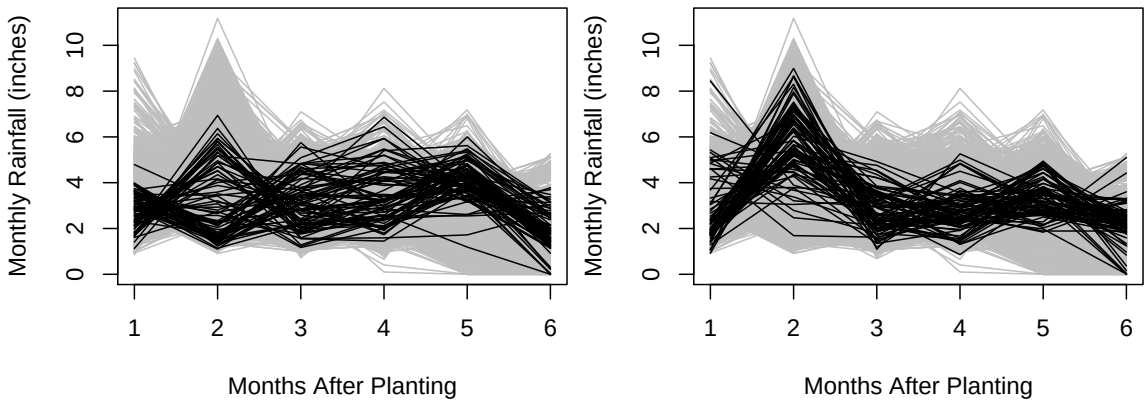

Figure 10. Examples of rainfall trends corresponding to different crop performance regimes. The grey area indicates the range of all rainfall data for the year, while the black lines in each plot are for instances with similar crop performance. The two panels highlight the rainfall associated with two different crop performance outcomes.

6.2. Accuracy

The premise of the vector-vector pattern mining algorithm is that examining multiple dimensions for both explanatory and response variables will provide a basis for discovering interesting patterns. We next will examine the effect of varying the number of dimensions (attributes) included in each of these vectors.

The neighborhood around each starting instance is defined by taking the nearest neighbors as measured by the Euclidean distance of the normalized crop performance metrics (yield, sugar, sugar lost to molasses). To confirm if using all 3 crop performance attributes was advantageous, all other parameters are held constant and repeated the analysis using alternate data for the Euclidean distance calculation. Using each attribute individually as well as each combination of 2 attributes was evaluated. Results for each dimensionality were averaged and are shown in Figure 11. All distribution comparisons show stronger correlations as the number of attributes used to determine the neighborhood increases. By considering both quantity and quality of the crops, it is possible to group similar instances together. Figure 11 also

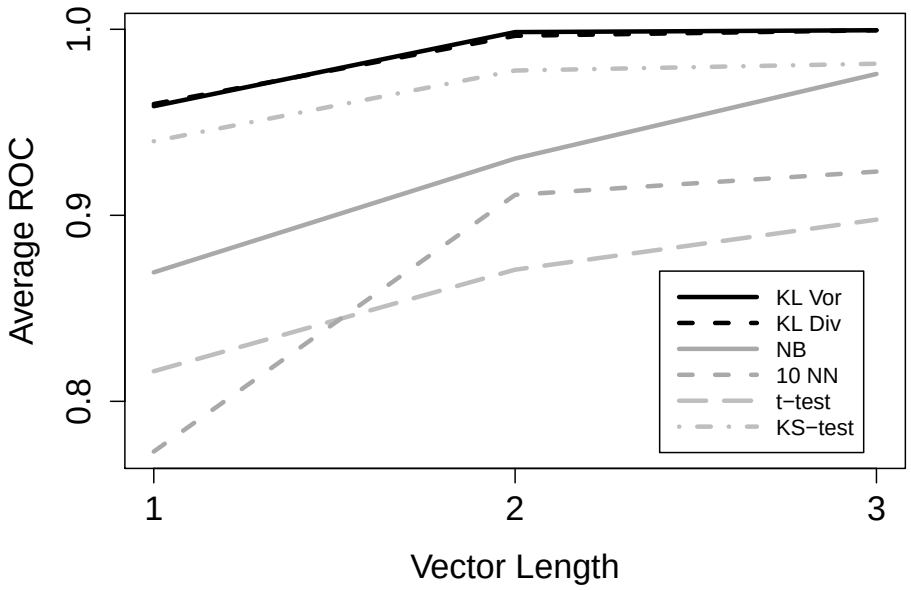

Figure 11. Evaluation of the impact of the length of the vector used to define the neighborhood (P) of similar crop performances. The response dimensions are combinations of yield, sugar, and sugar lost to molasses. A data series for each of the distribution comparison methods described in Section 3.2 is shown.

demonstrates that the divergence based distributions comparisons show the strongest performance, while the single attribute distribution comparison methods show the weakest performance.

An important decision for time series exploration is to what extent the dimensionality should be reduced. A balance should be made between reducing processing time without discarding too much information. As an additional consideration in the case of rainfall in the agricultural domain, it is expected that 1" of rain x days after planting would have the same effect as rainfall x+1 days after planting. Thus some amount of smoothing will be required to ensure a difference of one day does not create noise that would obscure patterns. Attribute reduction of the rainfall time

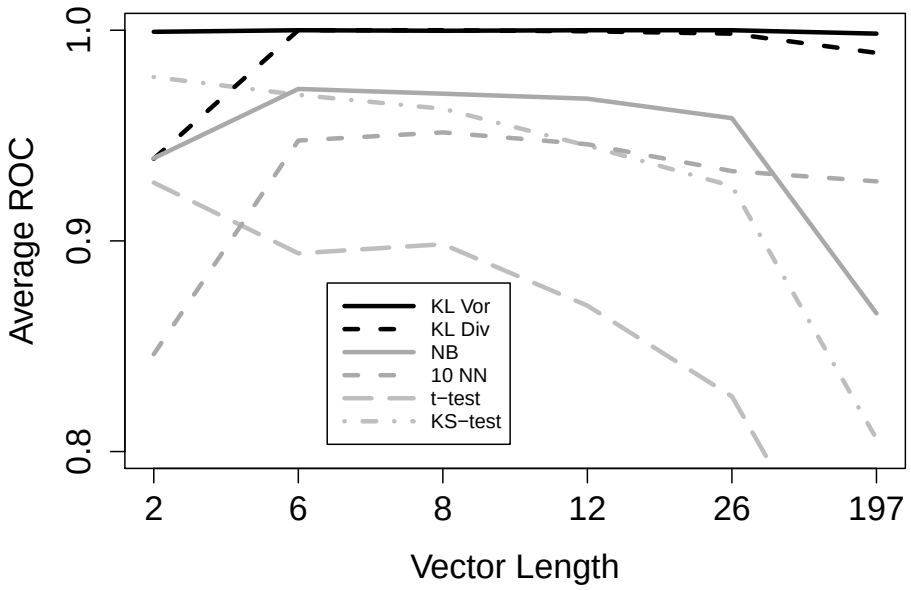

Figure 12. Evaluation of the impact on accuracy of the length of the vector used for the distribution comparison. Explanatory dimensions indicate the number of time steps the rainfall time series was reduced to.

series is accomplished by using subtotals for increasing larger time periods. The time periods ranged from 3 months (2 attributes) to 1 day (197 attributes). As shown in Figure 12, for all distribution comparison methods (except the Kullback–Leibler Voronoi method which was effective for all situations), some attribute reduction was helpful to reduce the noise found in daily data. Conversely, on the other extreme, over reduction led to a deterioration of performance.

The effect of the size of the neighborhood (P) assigned to each of the query points was examined. For a static set of 20 neighborhood starting seeds, the algorithm was run 11 times, with varying neighborhood proportions. Results are in Figure 13. Very small or large neighborhoods (1% and 99%) resulted in lower ROC scores. The discrimination of the KL divergence method more quickly approached perfect.

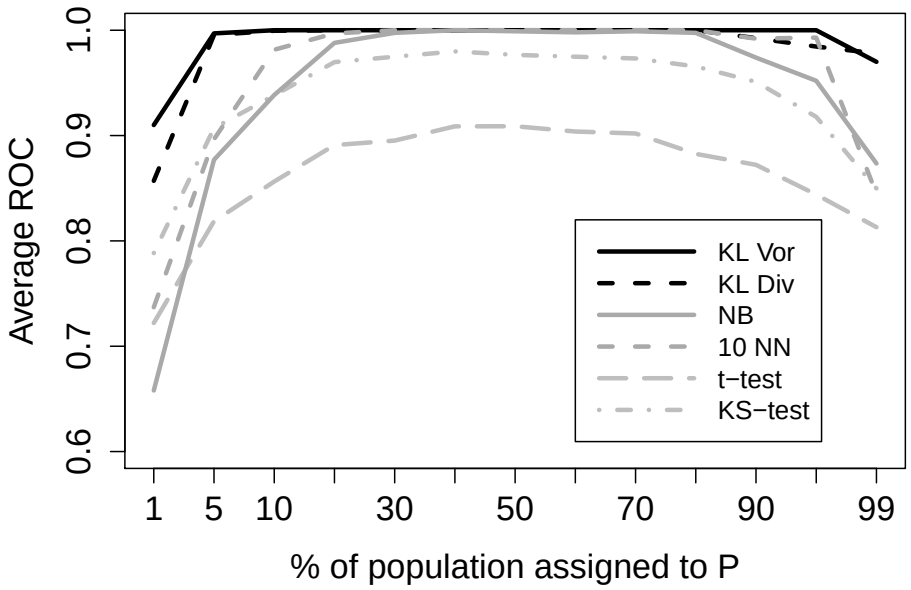

Figure 13. Influence of the number of nearest neighbors assigned to the neighborhood (P) of similar crop performance.

Information gain from the Naive Bayes and nearest neighbor methods improved as the split was closer to 50/50. Single variable methods never consistently reached perfect, indicating that using a distribution comparison that could account for the entire time series at one time was more effective in finding patterns. It is important to note that neighborhood size will change the type of pattern that is discovered. Very small neighborhoods show poor results as there are many other variables not included in the analysis (e.g. temperature, nitrogen), thus there is still a large amount of noise. As the neighborhood size increases past 50%, the more typically performing fields are compared against the outliers.

6.3. Speed

The speed impact of including additional dimensions in the time series data as well as how the algorithm would scale with an increasing number of instances is next described. For both experiments, the times reported are for 100 repetitions [12] of the distribution comparison calculation. The initial neighborhood selection using nearest neighbors will not depend on the number of attributes in the second vector, and will scale as O(log n) with the number of instances being examined. 10% of the instances were assigned to the neighborhood (P). A value of k=10 neighbors was used for the KL divergence and NN algorithms. 1/3 of the instances were reserved for testing the naive Bayes and nearest neighbors distribution comparison methods. The KL Voronoi distribution comparison method has been recently developed and not yet optimized for speed. It currently exhibits the slowest performance of all the methods used.

The timeseries of rainfall data was reduced by varying degrees, as described in Section 6.2. This data is then used as the second vector in the algorithm, evaluating the distribution comparison. A dataset was used with 3558 instances and the same length of vectors as in Figure 12. The resulting performance times are shown in Figure 14. As would be expected, increasing the number of dimensions used in the distribution comparison increases the run time. For a medium range (6-26) of dimensions, however, the speed impact is less than a single order of magnitude larger as compared to using two dimensions. Note that this intersects well with the accuracy of the algorithm. Adding enough dimensions to reach the most accurate area of performance does not impose a prohibitive time penalty.

The number of instances available in our data set is large relative to what is normally available in the agricultural domain, yet is certainly small compared to many data mining applications. Running times for this dataset on commodity hardware are

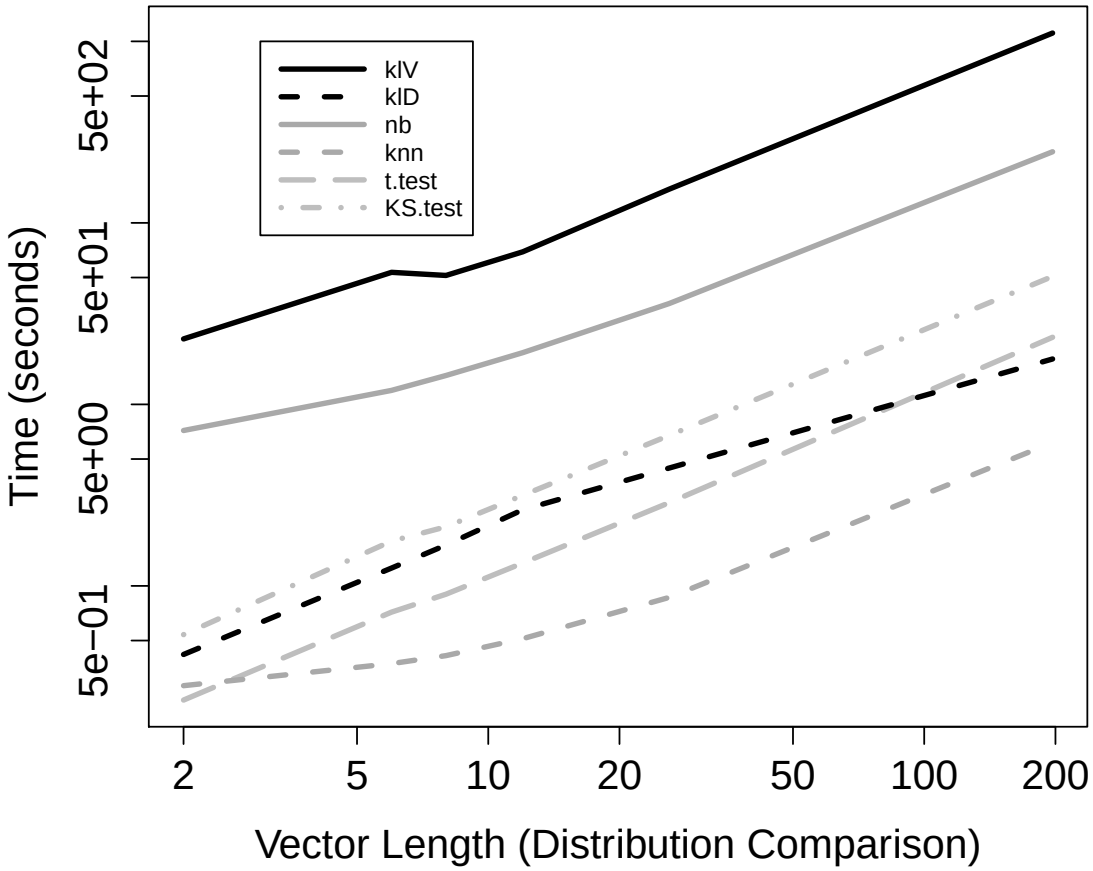

Figure 14. Evaluation of the impact on speed of the length of the vector used for the distribution comparison. Explanatory dimensions indicate the number of time steps the rainfall time series was reduced to.

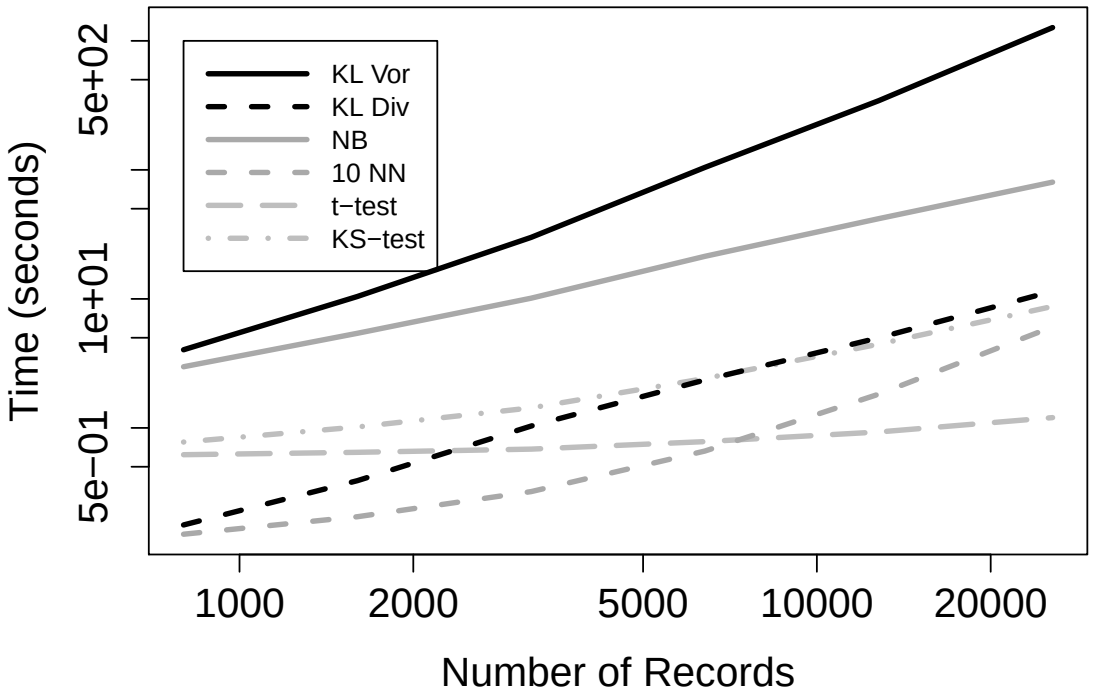

Figure 15. Time impact of scaling each distribution comparison to larger data sets.

acceptably fast for our application. Figure 15 demonstrates the scaling of the running time relative to the number of records in the data set. The single attribute t-test (based on the means of the two populations) has the best scaling, but was earlier demonstrated to provide the least information. The other single attribute test (KS-test) scales at a similar pace as the multi-dimensional distribution comparisons. In situations where the distribution is important (such as distribution of rainfall during a growing season), using a KL divergence comparison does not impose a significant speed penalty compared to a single dimension distribution test.

6.4. Predictions

The patterns and trends described in the previous sections are interesting on their own merit, but using the observations to improve future yield predictions provides a concrete benefit for the work. Future predictions have so far focused on the yield, so in this portion of the study only the yield portion of the crop performance vector is considered.

Adding rainfall data to an existing linear prediction model for sugar beet yield increases the accuracy of future predictions. As discussed in Section 6.2 and shown in Figure 12, the strength of the patterns between crop performance and rainfall were stronger for intermediate dimension reduction. The mean ROC of the four divergence and information gain based distribution comparisons is plotted in Figure 16. Two additional pieces of information have been added. First, data is added about the correlation coefficient (R^2) characterizing the linear model that includes the corresponding rainfall time series. Finally, the mean accuracy of predicting a held back year of data is added. The calculated correlation coefficient continues to increase (improve) as the number of dimensions increases. This goes contrary to our expectations, since the rainfall occurring at day 141 after planting should not have a different effect as compared to similar rainfall at day 142 after planting. This intuition is validated by the prediction results for holding back one year of data from the training, the accuracy when testing it first improves as more dimensions are added, but then begins to decrease as the highest dimensions are reached. This trend is matched in the ROC trend, indicating that the pattern strength measured by the vector-vector algorithm illuminates the physical situation.

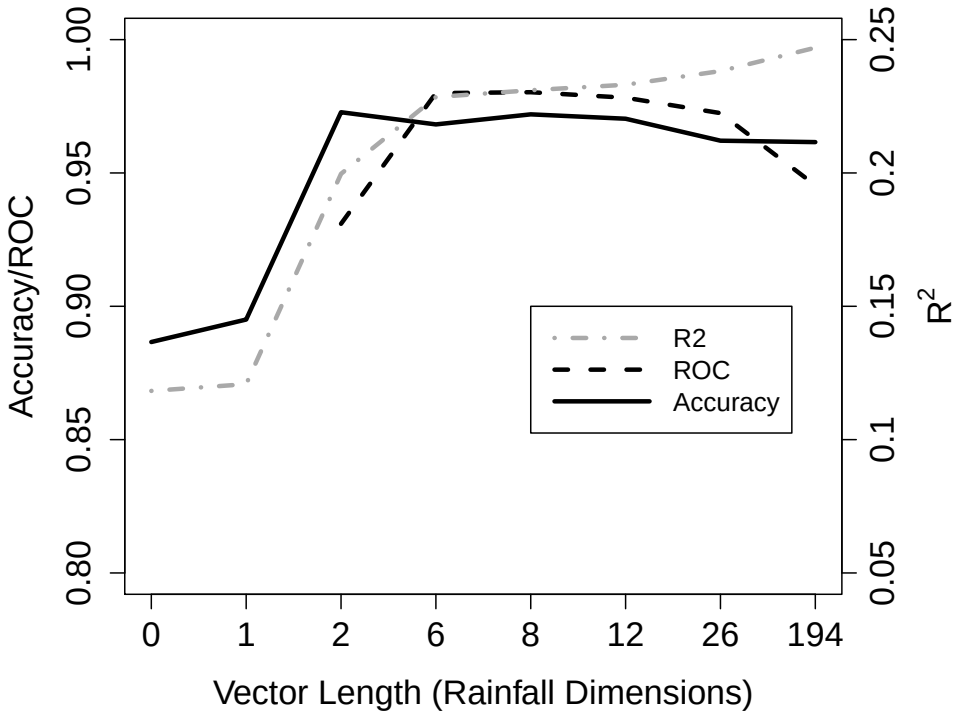

Figure 16. Impact of dimension reduction on apparent effectiveness. The correlation coefficient (R^2) measures the linear model fit on a field by field basis, the ROC measures the strength of the vector-vector patterns, and accuracy is the linear model performance when predicting total tonnage harvested in the study region in one year.

CHAPTER 7. RESULTS - INSEY

A portion of the analysis has been repeated on an independent vector of time series data, the In-Season Estimate of Yield (INSEY). The INSEY value is calculated from the Normalized Difference Vegetation Index (NDVI) and the Growing Degree Days (GDD). GDD itself is calculated from the minimum and maximum daily temperatures experienced by the crop. In this section we will report on some example patterns as well as an accuracy evaluation to determine the length and starting point of the INSEY time series that exhibits the strongest patterns.

7.1. Example Pattern

Figure 17 shows an example pattern that can be found in the data set. The left panel highlights the instances with above average yield and high sugar content. A strong INSEY value is observed for most of the growing season. The right panel highlights instances with below average yield and sugar content. As expected, a lower INSEY values is observed for most of the growing season.

7.2. Accuracy

Validating the premise that using additional attributes in the distribution comparison is helpful was explored with rainfall data in Figure 12. The same question is examined again using the INSEY data. For rainfall, the number of dimensions was changed by aggregating the data. For INSEY the number of dimensions is varied by increasing the duration of the time series used. The rainfall patterns suggested that some aggregation was helpful and Figure 18 shows an analogous result. Using only the values obtained at the end of the season (small number of time points) resulted in a lower ROC value, indicating that the short duration of the time series is missing some useful information. This is to be expected, as the crop canopy health (measured by

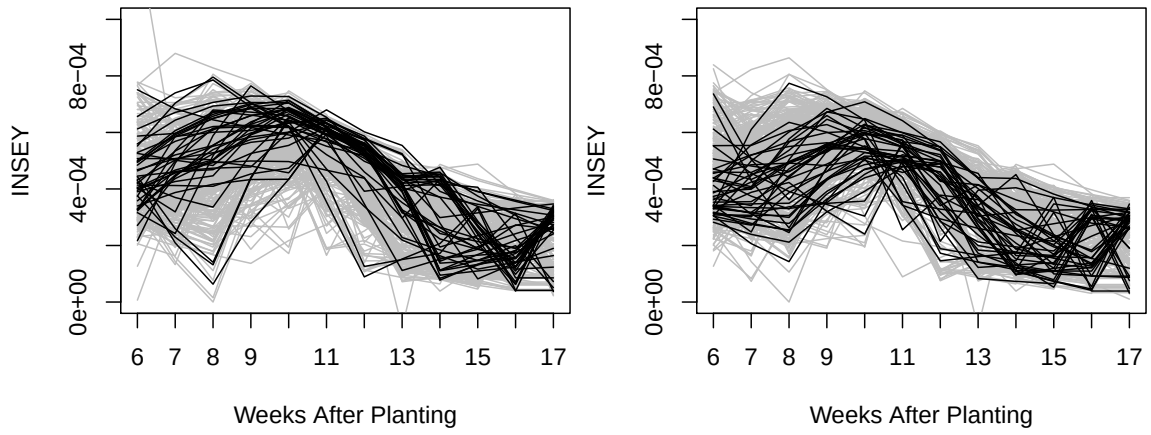

Figure 17. Examples of INSEY trends corresponding to different crop performance regimes. The grey area indicates the range of all INSEY data for the year, while the black lines in each plot are for instances with similar crop performance. The two panels highlight the INSEY associated with two different crop performance outcomes.

NDVI) might be significantly affected by late season weather, while significant yield or sugar storage was accomplished during different growing conditions in the middle of the season. Yet adding too much information also began to weaken the strength of the pattern as the highest number of dimensions was reached.

Previous research had demonstrated that NDVI values at growth stage 7 showed higher predictive ability then at growth stage 10 for particular crops. To follow up on this result using the vector-vector pattern mining technique, a small time series window is tested over a series of starting points during the growing season. A four week window was used with 2 week time steps for the starting point. Figure 19 provides the data indicating that the strongest distribution differences were found in the periods starting from week 7 to 11. It is important to note that this doesn't necessarily indicate that this is when the most sugarbeet growth is occurring, only that the INSEY measurement (using NDVI and GDD) provides the most information during that time period.

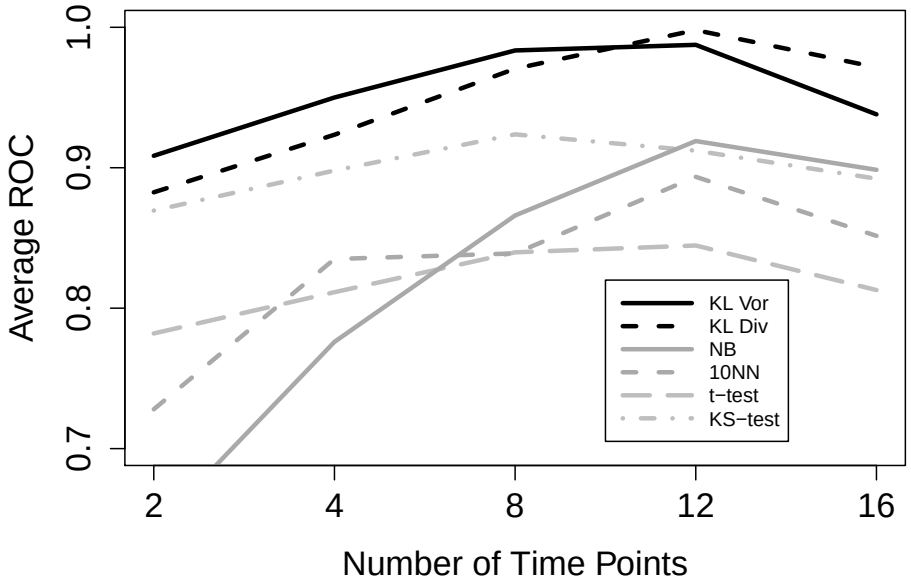

Figure 18. Effect of increasing the length of the INSEY time series in pattern strength. The INSEY time series used for the distribution comparison always ends at week 17 after planting, and the starting point is moved earlier in the season as the number of time points increases.

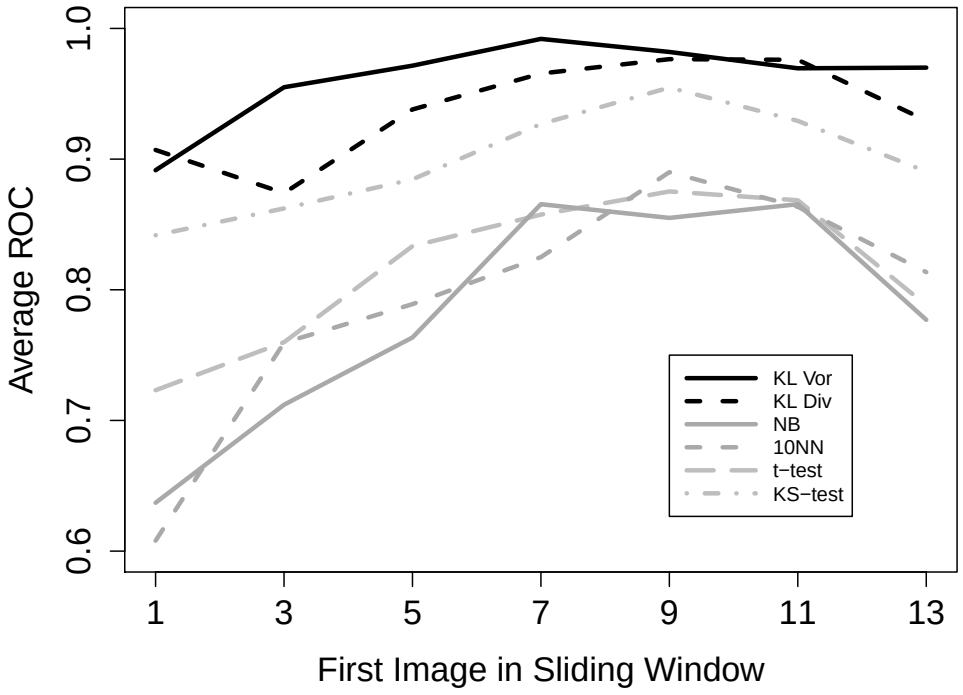

Figure 19. Relative importance of different time periods on pattern strength. The INSEY time series used for the distribution comparison is held constant at a 4 week window, while the starting point is varied.

CHAPTER 8. CONCLUSIONS

A new algorithm for finding relationships between two multi-dimensional vectors has been demonstrated to successfully find interesting patterns in a complex data set. Along with the traditional use of multi-dimensional information in explanatory attributes, this algorithm also takes into account multi-dimensional information in response attributes. The use of the vector-vector algorithm was compared with using single dimensional techniques (both as one attribute or calculated summary numbers), and the vector-vector algorithm was demonstrated to be more effective. One aspect of the vector-vector algorithm is to compare the distribution of two sets of data. A number of distribution comparisons techniques were evaluated. Kullback–Leibler divergence was the most robust distribution comparison method tested. The vector-vector algorithm was tested on data from the agricultural domain. Using multi-dimensional crop performance data in conjunction with a time series of rainfall data, the algorithm was able to find significant patterns. It was demonstrated that the algorithm successfully determined the optimal level of dimension reduction (time series smoothing) for the rainfall data.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A.N. Swami, *Mining association rules between sets of items in large databases*, Proc. ACM SIGMOD Int'l Conf. on Management of Data (Washington, D.C.), 26–28 1993, pp. 207–216.

[2] W. Bewket, *Rainfall variability and crop production in ethiopia: Case study in the amhara region*, Proceedings of the 16th International Conference of Ethiopian Studies, 2009, pp. 823–836.

[3] 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.

[4] R. Chiang, C.E. Huang Cencil, and E.-P. Lim, *Linear correlation discovery in databases: a data mining approach*, Data and Knowledge Engineering **53** (2005), 311–337.

[5] A.M. Denton and J. Wu, *Data mining of vector-item patterns using neighborhood histograms*, Knowledge and Information Systems (KAIS) journal **21** (2009), 173–199.

[6] 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.

[7] COAPS Dong-Wook S., *Assessing crop yield simulations with various seasonal climate data*, 18th Conference on Applied Climatology, 2010.

[8] Paul C Doraiswamy and Paul W Cook, *Spring wheat yield assessment using noaa avhrr data*, Canadian Journal of Remote Sensing **21** (1995), no. 1, 43–51.

[9] PC Doraiswamy, B Akhmedov, L Beard, A Stern, R Mueller, B Baruth, A Royer, and G Genovese, *Operational prediction of crop yields using modis data and products*, International Society of Photogrammetry and Remote Sensing Workshop Proceedings, vol. 36, 2007, p. 8.

[10] D.W. Franzen, *2 nitrogen management in sugar beet using remote sensing and gis*, GIS applications in agriculture **1** (2007), 35.

[11] K. Girma, K. L. Martin, R. H. Anderson, D. B. Arnall, K. D. Brixey, M. A. Casillas, B. Chung, B. C. Dobey, S. K. Kamenidou, S. K. Kariuki, E. E. Katsalirou, J. C. Morris, J. Q. Moss, C. T. Rohla, B. J. Sudbury, B. S. Tubana, and W. R. Raun, *Mid-season prediction of wheat-grain yield potential using plant, soil, and sensor measurements*, Journal of Plant Nutrition **29** (2005), no. 5, 873–897.

[12] W. Kusnierczyk, *rbenchmark: Benchmarking routine for r*, 2012, R package version 1.0.0.

[13] S. Li, *Fnn: Fast nearest neighbor search algorithms and applications*, 2012, R package version 0.6-4.

[14] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc functions of the department of statistics (e1071), tu wien*, 2012, R package version 1.6-1.

[15] P. Michiels, D. Gabriëls, and R. Hartmann, *Using the seasonal and temporal precipitation concentration index for characterizing the monthly rainfall distribution in spain*, Catena **19** (1992), no. 1, 43–58.

[16] J.E. Oliver, *Monthly precipitation distribution: a comparative index*, The Professional Geographer **32** (1980), no. 3, 300–309.

[17] Sudhanshu Sekhar Panda, Daniel P Ames, and Suranjan Panigrahi, *Application of vegetation indices for agricultural crop yield prediction using neural network techniques*, Remote Sensing **2** (2010), no. 3, 673–696.

[18] F. Perez-Cruz, *Kullback-leibler divergence estimation of continuous distributions*, NIPS '07 Workshop on Representations and Inference on Probability Distributions, 2007.

[19] R. Rastogi and K. Shim, *Mining optimized support rules for numeric attributes*, Information Systems **26** (2001), no. 6, 425–444.

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

[21] 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.

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

[23] R. Srikant and R. Agrawal, *Mining quantitative association rules in large relational tables*, Proc. 1996 ACM SIGMOD Int'l Conf. on Management of Data (Montreal, Quebec, Canada), 4–6 1996, pp. 1–12.

[24] C.O. Stockle, S.A. Martin, and G.S. Campbell, *Cropsyst, a cropping systems simulation model: water/nitrogen budgets and crop yield*, Agricultural Systems **46** (1994), no. 3, 335–359.

[25] M. Stöckle, C.O .and Donatelli and R. Nelson, *Cropsyst, a cropping systems simulation model*, European Journal of Agronomy **18** (2003), no. 3, 289–307.

[26] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, and J.P. Mesirov, *Gsea-p: a desktop application for gene set enrichment analysis*, Bioinformatics **23** (2007), 3251–3253.

[27] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov, *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, Proc. Natl. Acad. Sci. USA **102** (2005), 15545–15550.

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

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

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

[31] W. N. Venables and B. D. Ripley, *Modern applied statistics with s*, fourth ed., Springer, New York, 2002, ISBN 0-387-95457-0.

[32] 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.