USE OF QUALITY AND QUANTITY INFORMATION TOWARDS EVALUATING THE IMPORTANCE OF INDEPENDENT VARIABLES IN YIELD PREDICTION

A.M. Denton, E. Momsen, J. Xu D.W. Franzen, J.F. Nowatzki, and K. Farahmand

North Dakota State University,
Fargo, North Dakota

ABSTRACT

Yield predictions based on remotely sensed data are not always accurate. Adding meteorological and other data can help, but may also result in over-fitting. Working with American Crystal Sugar, we were able to demonstrate that the relevance of independent variables can be tested much more reliably when not only yield but also quality attributes are known, such as the sugar content and the sugar lost to molasses for sugarbeets. The problem of potentially over-fitting the data, when working with a large number of independent variables, is known as the curse of dimensionality. We show that the over-fitting problem in selecting variables can be effectively countered by increasing the number of dependent variables. An increased dimensionality on the side of dependent variables avoids that an independent variable may be considered relevant because similar values accidentally result in similar values of the dependent variable. We show that such reasoning can be very effectively applied to the problem of how to preprocess massively available data such as rainfall. Using rainfall data with a finer granularity than the aggregate over the full year can clearly hold benefits. In the absence of quantitative techniques, researchers and crop consultants have to decide how to preprocess rainfall data based on educated guessing alone. We provide a computational approach to answering such questions quantitatively.

Keywords: Yield prediction, Preprocessing, Big data

INTRODUCTION

Prediction of total annual quantities is notoriously difficult, because the number of years that can be considered comparable is typically small. These problems have been extensively known in statistics, and are aggravated when data are from production fields rather than the result of designed experiments (Glymour et al., 1997). In the case of the prediction of crop yield, it may appear that the large number of fields gives rigor to predictive approaches. However, it is important to realize that many factors, such as rainfall or temperature, affect a large number of fields similarly. The outcomes of all fields in one year may have very little bearing on the outcome of any one field in the next year.

Some approaches to this problem model plant growth depending on meteorological observations (Steduto et al., 2008). However, the growth conditions are changing systematically across years due to changes in drainage, new varieties, and other factors that cannot be easily controlled. In this paper we pursue a data-driven approach, in which the goal is to derive relevant factors entirely from the data. The availability of large quantities of satellite image and meteorological data, and increasingly data from sensors, suggest that this will become a more important approach in the future. The move from model to completely data-driven approaches follows in the footsteps of other "big data" areas. For example, in text applications that used to use grammatical structure (Smeaton, 1992), it is now common to only use the data (Wang, 2007). Similarly, we attempt to extract statistically sound information from data directly, without building models based on assumptions about any underlying structure.

Prediction problems, for which the number of observations of the dependent variable is small in comparison with the number of independent variables, have the associated risk of over-fitting. In the yield prediction problem, there may be data on several thousand fields available. However, since the weather conditions are similar in fields that are geographically close, the problem may still be ill-conditioned, corresponding to a small number of fully independent observations. In addition, the number of available climate variables, such as rainfall and temperature, each of which is available over time, is inherently large. As a result, it is important to consider what is often termed the "curse of dimensionality" (Verleysen and François, 2005). This term is based on the notion that each independent variable can be viewed as a dimension of the space of variable combinations. If the space of independent variables is large, the dependent variable, i.e. yield, offers only little evidence for making a reliable prediction.

In this paper we show that considering multiple dependent variables as well as multiple independent variables is a strategy for defeating the curse of dimensionality. Yield is not the only variable that is relevant as an outcome of the plan growth process. Many crops are characterized by quality variables as well as their weight. In particular, we consider sugarbeets for which the sugar content and the sugar lost to molasses, SLM, act as quality attributes. Considering yield and the two quality variables together, results in a space of possible outcomes that is itself 3-dimensional. Evaluating variables against this 3-dimensional space provides additional evidence for which variables are relevant to the growth process. While crops coming from two fields may accidentally coincide in their

weight, they are less likely to coincide in weight and the quality variables, unless the growth conditions are indeed equivalent with regard to the outcome.

CONCEPTS

It is common in machine learning and data mining to consider variables or attributes as dimensions in a vector space. This interpretation allows applying any mathematical techniques that are defined on vector spaces, and it can also be useful for visualizing data. Dimensions up to three can be visualized within our three spatial dimensions. Since we also consider vector spaces of higher dimensions, we will use a parallel coordinate representation (Inselberg 1991), in which each dimension is represented as if it were a time point in a time series, i.e. each dimension corresponds to a position on the x-axis of a graph, and its value is indicated in the y direction. Coordinates that belong to the same data point, i.e. in this case field, are connected through lines. Fig. 1 shows two examples of a visualization of yield, sugar, and SLM for sugarbeet fields of the American Crystal Sugar Company. A subset of fields is highlighted in black, while the remaining ones are shown with a gray color. It can be seen that when there is much rain late in the growing season (left panel), sugarbeets typically don't have a very high sugar content or SLM, and the yield is never very low. In contrast, when there is very little rain late in the growing season (right panel), low yields are more common. Overall, the distribution of the values for low rainfall late in the season (black lines) is much more similar to the overall distribution (gray lines) in the right panel than in the left panel. Knowing which subsets have a distribution that is substantially different from the overall distribution of data points, is of interest beyond the prediction question (Denton and Wu, 2009). The comparison of the subset distribution with the overall distribution is often done using the Kullback-Leibler, K-L, divergence (Denton et al., 2010), which will be discussed in the next section.

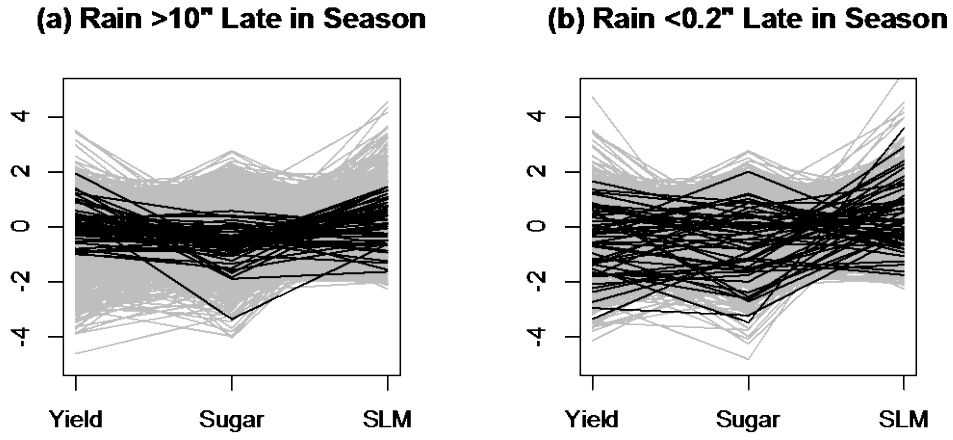

Fig. 1: Parallel-coordinate representation of yield, sugar, and SLM in sugarbeets, with subsets highlighted that correspond to (a) high or (b) low rainfall in the last two months of the growing season.

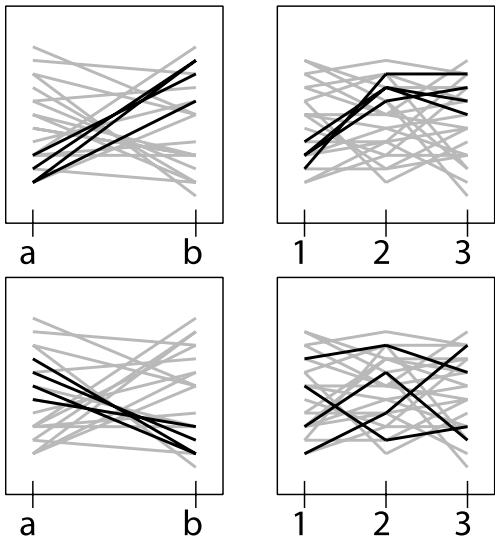

Fig. 2: Schematic of a parallel-coordinate representation of two sets of variables for the same agricultural field, where a cluster with regard to the variable in the left panels may (top) or may not (bottom) correspond to a cluster in the right-hand-side panels.

The K-L divergence can be calculated for subsets of points that are defined by having a particular property, such as high or low late season rainfall. It can also be calculated for the groups of points that are themselves constructed based on multi-dimensional data. In this paper we propose a technique for evaluating the usefulness of processing steps by grouping data points using clustering, and then evaluating those clusters using the K-L divergence. Fig. 2 illustrates that concept. On the left hand side, a parallel-coordinate representation of two variables can be seen. These two variables could, for example, be the early and late season aggregates of the rainfall in the field. In practice we will not limit ourselves to two variables but will consider as many as 180 when considering daily rainfall values. Assume that the black lines in the top and bottom panel show two clusters of similar fields. Those fields not only have rainfall values associated but also the harvest variables. Since those variables were not used in the clustering, there is no guarantee that they will show any structure. The top row shows a case in which the cluster in the rain variables corresponds to a tight set of harvest variables, corresponding to a high K-L divergence. The bottom row shows no pattern in the harvest variables (right panel) in comparison with the overall distribution. That means that the K-L divergence is small. We consider the average K-L divergence over all clusters in our evaluation of the suitability of the preprocessing strategy.

K-L Divergence based on Voronoi cells

The K-L Divergence is a measure for the degree to which a probability distribution P differs from a reference distribution Q. Eq. (1) shows the definition whereby the logarithm of the ratio of differential probability distributions of P over Q is integrated according to the distribution of P.

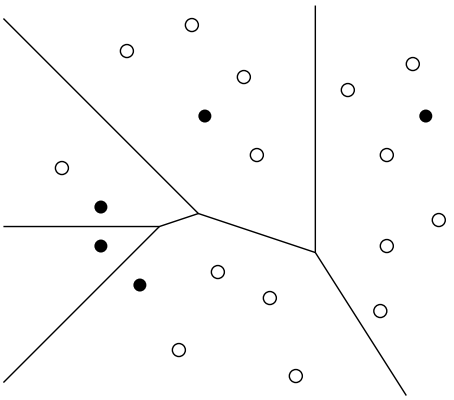

Fig. 3: Definition of cells for the estimation of probability distributions.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} ln(\frac{\mathrm{d}P}{\mathrm{d}Q})\,\mathrm{d}P \tag{1}$$

Common implementations of the K-L divergence assume that the points in distribution P and Q are independently distributed (Perez-Cruz, F. 2007). Since we want to compare a subset distribution with the overall distribution of points we estimate probabilities over Voronoi cells of the points with the item (black points in Fig. 3). A Voronoi cell contains all those points that are closer to its defining (black) point than to any other.

Comparison with Multivariate Regression

The concept of considering multiple dependent variables in regression problems is known as multivariate regression. When linear regression is used, the prediction problem can be decomposed into independent regression models for each of the dependent variables. That means that linear multivariate regression in itself is not suitable for identifying those variables that have an effect on the combination of dependent variables. Other approaches that have been developed recently to make use of the potential of a joint evaluation of multiple dependent variables include graphical models (Wang, 2013). The research area is still new since, for well-conditioned regression problems, the separability of the linear regression problem based on dependent variables means that a combined evaluation of dependent variables is not expected to improve accuracy. The yield prediction problem critically depends on the ability of the model to extrapolate beyond the training data, since a new year is likely to show rainfall pattern, for which there is not close precedent available. It is this extrapolation for which our evaluation is shown to result in more stable results.

Data Preprocessing

The results in this paper are based on sugarbeet fields in the valley of the Red River of the North, covering eastern North Dakota and northwestern Minnesota. Yield, sugar, and SLM data were provided by American Crystal Sugar Company. Rainfall is processed from data that is provided by the National Weather Service.

The processing is done using GRASS Geographic Information System software. Rainfall for a field is assigned based on the shortest distance to NWS grid points. We also use Growing Degree Days, GDD, a commonly used measure of the effect of temperature on growing conditions.

Rain data are considered relative to the planting day for each field. Further aggregation is done using bins of size 7 for weekly, 30 for monthly, and 90 for the per-season evaluation. Yields are corrected based on the average weight of sugarbeets on the respective harvest day, relative to the mean harvest day for the year. I.e., if the harvest day is early in the season, when the average yield is lower than the mean for that year, the yield is corrected to a higher effective value.

Algorithm

The algorithm can be summarized as follows: The rainfall values are clustered using k-means clustering (Jain, 2010) with k=20. For each of cluster, yield, sugar, and SLM records are extracted and the K-L divergence between the set defined by the cluster and the set of all records is calculated using the Voronoi-cell-based algorithm. The resulting K-L divergences are averaged. A high K-L divergence indicates a strong relationship between rainfall data and crop properties. The process is repeated for different rain aggregation and processing steps.

EXPERIMENTS

Fig. 4 shows an example cluster of rainfall values in the left panel together with the corresponding crop properties in the right panel. An example with a high K-L divergence is chosen to illustrate the concept. Note how the cluster members have consistently high yield, and very low SLM with intermediate to low sugar content.

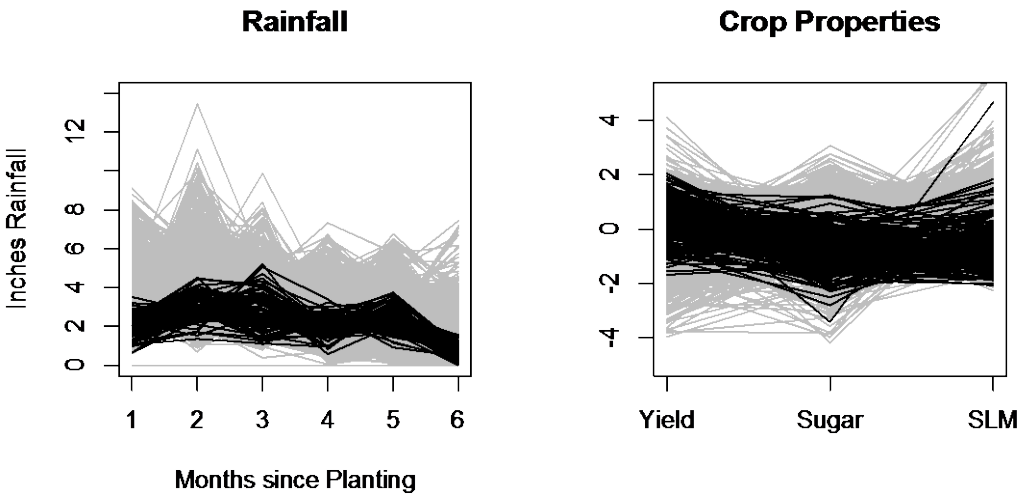

Fig. 4: Example pattern with cluster members highlighted in black and other fields shown in gray. The left panel shows rainfall aggregated to months, while the right panel shows yield, sugar, and SLM in parallel coordinate representation.

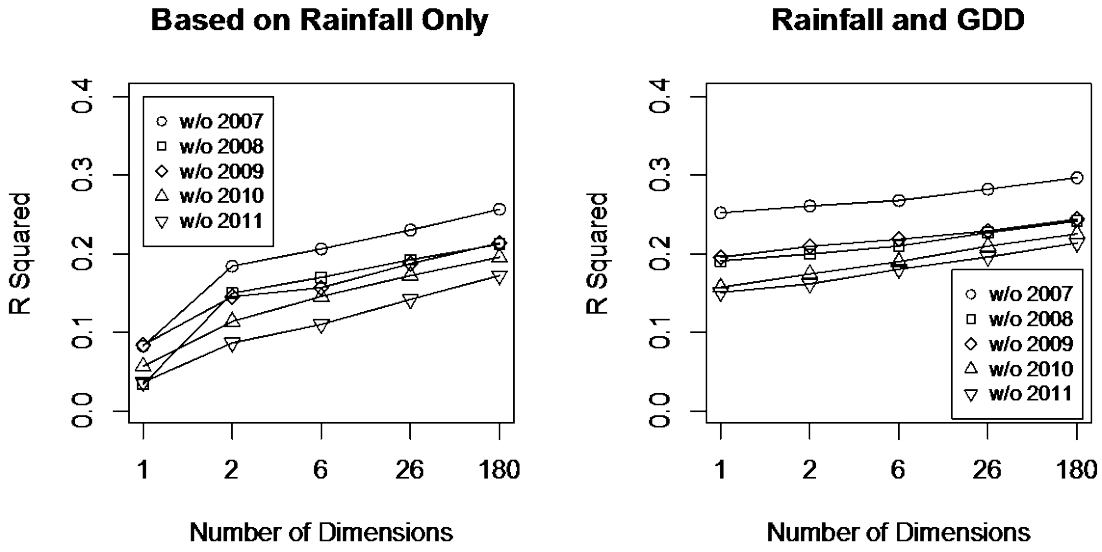

Fig. 5: R^2 values for linear regression based on rain data with varying number of dimensions, for models built by leaving out one year in each case.

To illustrate the motivation for using the K-L divergence, we will first consider traditional linear regression-based approaches for evaluating the quality of prediction. When using linear regression, it is common to use the R^2 value as a measure of how accurate the prediction is expected to be. Fig. 5 shows that R^2 values increase with the number of rain dimensions. The conclusion appears straightforward that daily values contain the most useful information for the prediction. One may argue that the prediction based on rainfall alone has such small R^2 values that a prediction is simply not reliable. The right panel shows the same linear regression, but with GDD as additional variable. As expected, the overall R^2 values are slightly higher. It is also worth noting that, even in this setting, daily rainfall values are expected to give the best model.

This result is counterintuitive, since the total yield should not matter strongly affected by whether rain falls on a particular day or on the next day during the growing season. Indeed, when doing the actual prediction, the results are different, and daily rainfall gives in the poorest predictions over most experimental scenarios. Fig. 6 shows the error in the total annual yield based on linear regression, i.e. the absolute value of the difference between the predicted overall yield per acre and the actual value. The left panel only uses rainfall, while the right panel also uses the GDDs. All but one year are used for creating the model, and the one remaining year is used for the evaluation. Each line corresponds to a different year. It can be seen that while the dependence on the number of rain bins is somewhat erratic in both cases, daily rainfall values tend to have higher error rates. Note that the overall prediction errors are not as high as one might have expected based on the R^2 values in Fig. 5. In the right panel, error rates are as low as 0.2 tons per acre, which corresponds to less than 1% of the total. This can be explained by realizing that the R^2 values are a measure of the quality of prediction at the level of individual fields, and that many of the errors in that prediction compensate each other in the prediction of the annual total.

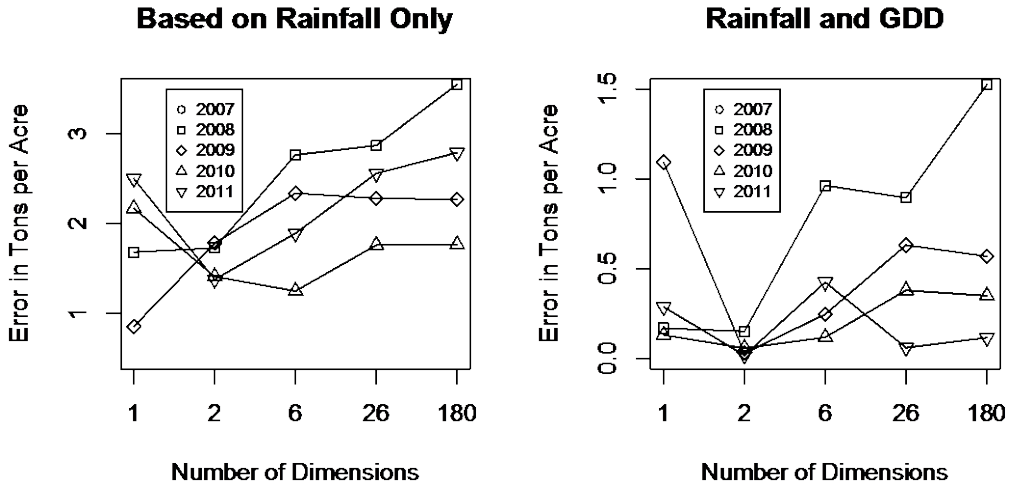

Fig. 6: Error in overall predicted yield using linear regression, with the left panel showing the regression results based on rainfall only, and the right panel the regression based on rainfall and GDD.

One could consider selecting the processing strategy by averaging over annual totals and using the processing strategy that results in the lowest average errors. However, it is important to realize that that amounts to selecting a model based on a very small number of observations, in particular one per year. Yet another alternative could be to use the per-field error rate, which is shown in Fig. 7. It can be seen that when only rainfall is used, the per-field analysis results in contradictory evidence. For 2011, the error rate consistently decreases with the number of rainfall dimensions, while for 2007 and 2008 it consistently increases. In the other two years intermediate values are optimal.

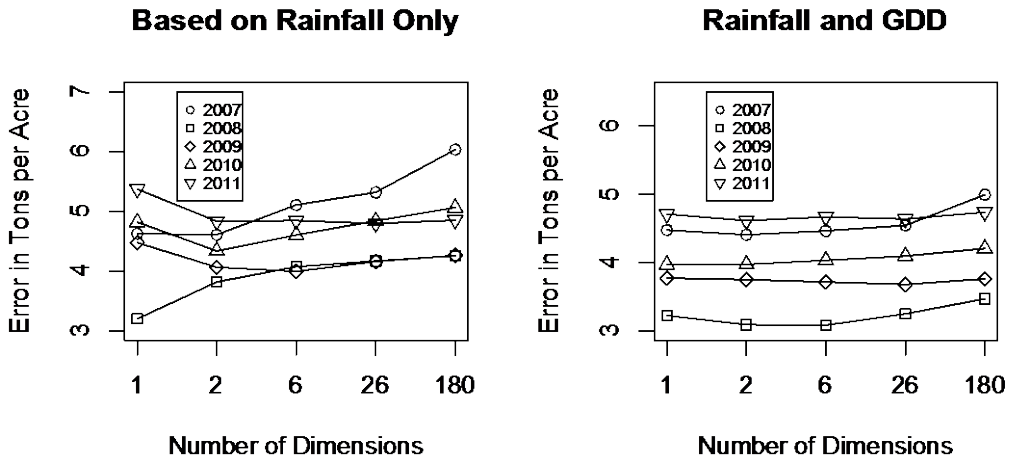

Fig. 7: Error in per-field predicted yield using linear regression, with the left panel showing the regression results based on rainfall only, and the right panel the regression based on rainfall and GDD.

When GDD is added to the independent variables, the results that are shown in the right panel of Fig. 7 are so flat that no reliable conclusions can be drawn on what processing to use. In fact, among the 5 years that are considered, 4 show a different optimal selection for the number of dimensions.

These concerns motivate using the K-L divergence as a measure for evaluating the usefulness of preprocessing strategies. Fig. 8, left panel, shows that when the K-L divergence is calculated over the same sets of data as the models in Figures 6 and 7, the results show a somewhat more consistent behavior. Monthly rainfall values have the highest K-L divergence for each data set. Note that the K-L divergence algorithm was limited to 2 or more dimensions. These results support the notion that daily or weekly rainfall aggregation is not as useful as the R^2 values of the linear model would have suggested. However, aggregation by season does not emerge to be as stable as the evaluation on annual totals would have suggested.

The right panel of Fig. 8 shows further preprocessing strategies. It can be hypothesized that high rainfall values in a single day may not help plant growth as much as rain that falls over several days. For that reason, we test the effect of only considering rain of up to 1 inch within a given day, shown as "min(x,1)" in the right panel of Fig. 8. Any further rainfall is ignored in the aggregation of results. This processing can, indeed, be seen to result in a higher K-L divergence than the default processing. If cutting off high rain values is beneficial, one may consider using rain logarithmically, shown as "log(1+x)". However, that approach can be seen to be less successful. Yet a different thought is that low rainfall is not absorbed into the soil and rainfall of less than 0.2 inches should be ignored, shown as "max(x-0.2,0)". Again, that hypothesis is not supported by the data, suggesting that low rainfall is still, in some form, absorbed by the plant, such as through the leaves. Each of the processing strategies confirm the conclusion that weekly rainfall aggregation gives the most stable estimates overall.

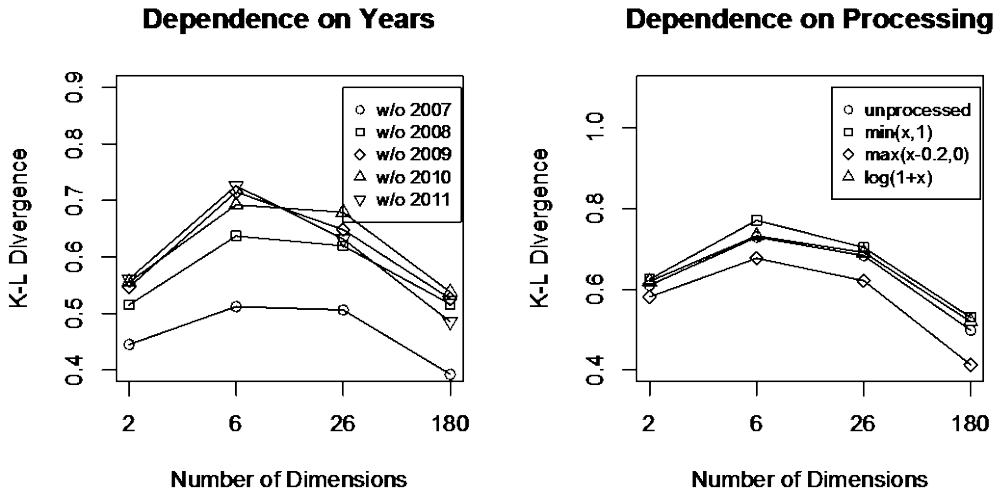

Fig. 8 K-L divergence for different subsets of the data (left panel) and different preprocessing of the rainfall data (right panel).

CONCLUSIONS AND FUTURE WORK

In summary, we have shown that crop quality information can be effectively employed in evaluating preprocessing strategies for variables that are to be used in predicting yield across years. Using high-dimensional data in prediction carries the risk of the "curse of dimensionality". We have shown that the Kullback-Leibler divergence calculated over clusters in the in the multi-dimensional data can be used to identify those preprocessing strategies that result in the most reliable predictions.

While this paper focuses on rainfall and temperature data, the extension to remotely sensed data, such as NDVI from satellite imagery, is mathematically straightforward, and the questions to be answered rather more varied: The visible biomass for sugarbeets does not correlate to beet size in a straightforward manner. Further complicating factors include the prevalence of missing data due to clouds, and complications in distinguishing leaf coverage from chlorophyll content when working with satellite data (Haboudane et al., 2002). Any of these questions lend themselves to an evaluation of different preprocessing strategies.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Partnerships for Innovation program under Grant No. 1114363. We thank the NSF and NDSU-Industry Consortium for funding of the research, especially consortium member American Crystal Sugar Company for providing a portion of the data used in this analysis.

REFERENCES

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

Denton, A.M., J. Wu, and D. H. Dorr. 2010. Point-distribution algorithm for mining vector-item patterns. In Proc. 16th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining: Useful Patterns Workshop, Washington DC.

Glymour, C., D. Madigan, D. Pregibon and P. Smyth. 1997. Statistical themes and lessons for data mining. Data mining and knowledge discovery 1(1): p.11-28.

Haboudane, D., J.R. Millera, N. Tremblay, P.J. Zarco-Tejada, L. Dextraze. 2002. Remote Sensing of Environment 81: p. 416– 426.

Inselberg, A. and B. Dimsdale. 1991. Parallel coordinates. Human-Machine Interactive Systems. Springer US: p. 199-233.

Jain, A.K. 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31(8): p. 651-666.

Perez-Cruz, F. 2007. Kullback-Leibler divergence estimation of continuous distributions. In NIPS '07 Workshop on Representations and Inference on Probability Distributions.

Smeaton, A.F. 1992. Progress in the application of natural language processing to information retrieval tasks. The computer journal 35.3: p. 268-278.

Steduto, P., D. Raes, T.C. Hsiao, E. Fereres, L. Heng, G. Izzi, and J. Hoogeveen.

2009. AquaCrop: a new model for crop prediction under water deficit conditions. In: López-Francos A. (ed.). Drought management: scientific and technological innovations. Zaragoza : CIHEAM, 2008: p. 285-292
Verleysen, M., and D. François. 2005. The curse of dimensionality in data mining and time series prediction. Computational Intelligence and Bioinspired Systems. Springer Berlin Heidelberg: p. 758-770.
Wang, J. 2013. Joint estimation of sparse multivariate regression and conditional graphical models. arXiv preprint arXiv:1306.4410.
Wang, X., A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval., 2007. Seventh IEEE International Conference on Data Mining, ICDM.