2016

A Bayesian network approach to county-level corn yield prediction using historical data and expert knowledge

Vikas Chawla
Iowa State University

A Bayesian network approach to county-level corn yield prediction using historical data and expert knowledge

by

Vikas Chawla

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:

Baskar Ganapathysubramanian, Co-Major Professor

Soumik Sarkar, Co-Major Professor

Jin Tian, Co-Major Professor

Samik Basu

Chinmay Hegde

Iowa State University

Ames, Iowa

2016

DEDICATION

I would like to dedicate this thesis to my research advisor, Dr. Baskar Ganapathysubramanian and Dr. Soumik Sarkar. Without their constant support and expert guidance, I would never have been able to complete this work with enthusiasm and become more well-versed in this field.

I would also like to dedicate this thesis to my family for their unconditional love and persistent support when completing this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to use the opportunity to express my gratitude to everyone who has inspired me and helped me with various aspects of conducting the research.

First and foremost, I am especially grateful to my family for constantly offering emotional support while enrolled in the graduate program. I lost my father during this journey at Iowa State University but I'm proud that I have fulfilled his dream of completing the graduate degree.

I would like to express my sincerest gratitude and appreciation to Dr. Baskar Ganapathysubramanian and Dr. Soumik Sarkar for their immeasurable amount of guidance, patience and support throughout this research and the writing of this thesis. Their insights and words of encouragement are truly a blessing to me and have often inspired me to become an excellent, a well-rounded practitioner of science.

I would also like to thank my committee members, Dr. Jin Tian, Dr. Samik Basu and Dr. Chinmay Hegde, for providing insightful advice, efforts, and contributions to this work. I would also like to thank my friends in the Lab of mechanics and Computer science department who have been best friends to share the up and downs of the life. Thank you all for being there for me during both good and bad times, and for making my time as a graduate student extremely delightful and memorable. I would like to thank all of my collaborators. Without their invaluable opinions and expertise, I would never have produced tangible results from my research efforts. Along with my collaborators and co-authors, we sincerely acknowledge the support of ISU PSI through PSI faculty fellow grant.

ABSTRACT

Machine learning has become a popular technology that has not only turbo-charged the existing problems in the AI but it has also emerged as the powerful toolkit to solve some of the interesting problems across the various interdisciplinary domains.

The availability of food is the biggest problem of the 21st century and many experts have raised their concerns as we continue to see a rise in the global human population. There have been many efforts in this direction which include but not limited to improvement in the seeds quality, good management practices, prior knowledge about the expected yield, etc.

In this work, we propose a data-driven approach that is 'gray box' i.e. that seamlessly utilizes expert knowledge in constructing a statistical network model for corn yield forecasting. Our multivariate gray box model is developed on *Bayesian network analysis* to build a *Directed Acyclic Graph (DAG)* between predictors and yield. Starting from a complete graph connecting various carefully chosen variables and yield, expert knowledge is used to prune or strengthen edges connecting variables. Subsequently, the structure (connectivity and edge weights) of the DAG that maximizes the likelihood of observing the training data is identified via optimization. We curated an extensive set of historical data ($1948 - 2012$) for each of the 99 counties in Iowa as data to train the model. We discuss preliminary results, and specifically focus on (a) the structure of the learned network and how it corroborates with known trends, and (b) how partial information still produces reasonable predictions (predictions with gappy data), and show that incorporating the missing information improves predictions.

CHAPTER 1. INTRODUCTION AND RELATED WORK

1.1 Introduction

The United Nations in 1948 recognized the *Right to Food* in the declaration of human rights and formed the Food and Agriculture Organization on *Food security* which defined the food security as the condition in which all the people at all time have[1]:

- Physical
- Social
- Economical, access to sufficient safe and nutritious food

But since last few years experts have raised their concerns over the factors (not limited to) that will have significant yet highly uncertain impacts on food security:

- Population growth
- Global water crisis
- Land degradation
- Climate change
- Agricultural diseases
- Dictatorship and kleptocracy
- Food sovereignty

There have been many efforts to protect the food security and various approaches have been adopted in this direction which includes but not limited to improvement in the seeds

quality, good management practices, prior knowledge about the expected yield, etc. There are strong, direct relationships between agricultural productivity, hunger, poverty, and sustainability[1]. Therefore, it's important to increase the agricultural productivity such that the demand and supply can be met. But, making changes as increasing productivity in areas dependent on rainfall; soil quality; expanding cropped areas; improving irrigation techniques; increasing agricultural trade between countries; and reducing gross food demand by influencing diets and reducing post-harvest losses[1].

Agricultural or Crop insurance allows deprived farmers to compensate for their unexpected losses by contributing premium to an insurance fund. This approach reduces the risk for an individual by spreading the risk across multiple fund allocations.

The United State of America, is the largest producer of corn in the world and produces on averag around 15,000 million busels per year. It has a market worth $80 billion and is roughly 0.1% of the total GDP. Corn is not only used for food but it has various other uses such as:

- Ethanol (in oil)
- Plastic production
- Gas industry
- Animal bedding

On the other hand, Iowa is the largest producer of corn in USA. Since, corn has such a huge market because of it the government is forced to make and formulate good agricultural policies that will benefit the farmers and overall production of the corn. Moreover, people also trade corn as a commodity in the share market and make profit out of it.

This requires one to have strong knowledge of the market trend and historical data. But, it's not possible for everyone to make the prediction of the future corn prices, analysis of the market and ability to recover from losses. This has allowed private players to create information asymmetry in the market by making better prediction model and selling it further to make profit. In this work, our goal is to build a publically available county level corn yield prediction model at par with private players.

1.2 Related Work

Crop yield forecasting is the methodology of predicting crop yields prior to harvest. The availability of accurate yield prediction frameworks have enormous implications from multiple standpoints, including impact on the crop commodity futures markets, formulation of agricultural policy, as well as crop insurance rating. The focus of this work is to construct a corn yield predictor at the county scale. Corn yield (forecasting) depends on a complex, interconnected set of variables that include economic, agricultural, management and meteorological factors. Conventional forecasting is either knowledge-based computer programs (that simulate plant-weather-soil-management interactions) coupled with targeted surveys or statistical model based. The former is limited by the need for painstaking calibration, while the latter is limited to univariate analysis or similar simplifying assumptions that fail to capture the complex interdependencies affecting yield.

Charles L. Hornbaker[2] have built a spatial model of maize yields in the US Corn Belt that uses the Bayesian prior estimation method for every state in the belt region which induces spatial smoothness among the regression coefficients to mitigate the effects of noisy data across regions and to improve yield forecasting. This helps in formulating an in-season forecasting model.

Nathaniel Newlands[3] have shown that the crop yield is strongly coupled to climate and soil environmental variables. The planting date and harvesting date also plays a significant role and have an appreciable impact on optimal annual yield as the efficiency that crops can use available water. This work is highly focused towards the soil texture and formulates a model to track sensitivity of yield to this variable.

CHAPTER 2. A BAYESIAN NETWORK APPROACH TO COUNTY-LEVEL CORN YIELD PREDICTION USING HISTORICAL DATA AND EXPERT KNOWLEDGE

This chapter is an article titled"A Bayesian Network approach to County-Level Corn Yield Prediction using historical data and expert knowledge " published In Proceedings of the 22nd ACM SIGKDD Workshop on Data Science for Food, Energy and Water, San Francisco, CA, 2016 authored by V. Chawla, H. Naik, A. Akintayo, D. Hayes, P. Schnable, B. Ganapathysubramanian, S. Sarkar[4].

2.1 Abstract

Crop yield forecasting is the methodology of predicting crop yields prior to harvest. The availability of accurate yield prediction frameworks have enormous implications from multiple standpoints, including impact on the crop commodity futures markets, formulation of agricultural policy, as well as crop insurance rating. The focus of this work is to construct a corn yield predictor at the county scale. Corn yield (forecasting) depends on a complex, interconnected set of variables that include economic, agricultural, management and meteorological factors. Conventional forecasting is either knowledge-based computer programs (that simulate plant-weather-soil-management interactions) coupled with targeted surveys or statistical model based. The former is limited by the need for painstaking calibration, while the latter is limited to univariate analysis or similar simplifying assumptions that fail to capture the complex interdependencies affecting yield. In this paper, we propose a data-driven approach that is 'gray box' i.e. that seamlessly utilizes expert knowledge in constructing a statistical network model for corn yield forecasting. Our multivariate gray box model is developed on

Bayesian network analysis to build a *Directed Acyclic Graph (DAG)* between predictors and yield. Starting from a complete graph connecting various carefully chosen variables and yield, expert knowledge is used to prune or strengthen edges connecting variables. Subsequently the structure (connectivity and edge weights) of the DAG that maximizes the likelihood of observing the training data is identified via optimization. We curated an extensive set of historical data ($1948 - 2012$) for each of the 99 counties in Iowa as data to train the model. We discuss preliminary results, and specifically focus on (a) the structure of the learned network and how it corroborates with known trends, and (b) how partial information still produces reasonable predictions (predictions with gappy data), and show that incorporating the missing information improves predictions.

2.2 Introduction and Related Work

Crop yield forecasting is the methodology of predicting crop yields (at various scales: from farms to counties, to countries and to global scale) prior to harvest. Accurate crop yield predictions have enormous implications from multiple standpoints. These include: the impact on the crop commodity futures markets, timely interventions for crop management, unraveling genetic-environment interactions (GxE) for plant breeding, and appropriate policy decisions in both developing countries where food shortages remain a threat and in US where improved yield forecasting can improve targeting of conservation funding from major federal programs such as the Conservation Reserve Program.

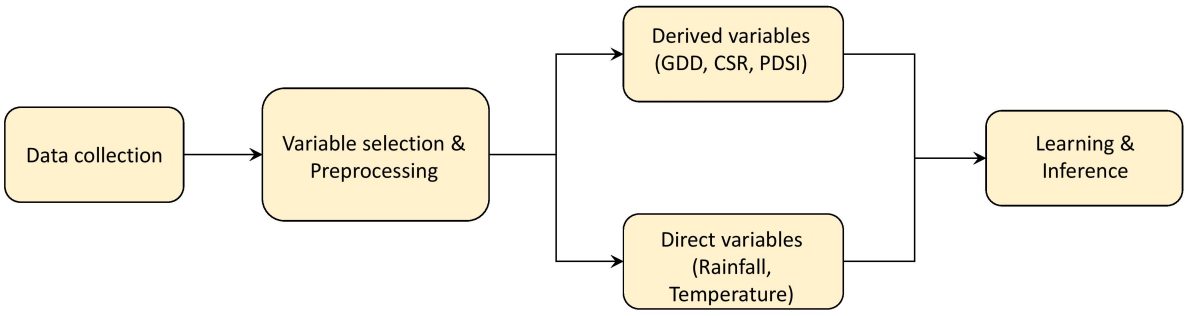

Figure 2.1 Schematic of the yield prediction workflow

The United States is the largest producer of corn in the world. Exports of corn alone account for approximately 10-20% of annual revenue in the trade market. In the United States corn is grown nationwide, but production is mainly concentrated in the heartland region which includes Iowa and Illinois. Government and insurance companies have established a compensation system that insures farmers to support them against natural causes that have adverse effects on yield, but their premium rates are reported to be too high [5; 6]. On the other hand, any fluctuations in the corn futures market can have a debilitating impact on farmers. Therefore, the U.S. Department of Agriculture (USDA) invests an enormous amount of time and financial resources to making periodic county level yield predictions. This helps keep market participants equally informed about events that influence cash and futures prices for major commodities in an effort to prevent market failure due to non-participation by uninformed groups. The intellectual foundation behind this effort, described in a Nobel Prize winning paper on "The Market for Lemons" by George Akerlof, is that markets will fail if one set of participants have more information than other participants. Recent developments in the way agricultural information is collected and shared suggests that companies and big data firms may now be able to beat the USDA at this activity leading to detrimental asymmetric markets. A publicly available high quality yield prediction tool will enable the producers to make informed decisions thereby ensuring a symmetrical market. This is the motivation for the current work.

Conventional crop forecasting relies on a combination of knowledge-based computer programs (that simulate plant-weather-soil-management interactions) along with soil and environment data and targeted surveys or is based on statistical black-box approaches. The former is limited by the need for painstaking calibration, while the latter is limited to univariate analysis or similar simplifying assumptions that fail to capture the complex interdependencies affecting yield [7; 8; 9]. In this paper, we tread a middle ground between so-called 'black-box' and 'white-box' approaches. We present a novel, knowledge-based statistical forecasting approach to predict county-wide corn yield in the state of Iowa. Our multivariate 'gray box' model is based on *Bayesian Networks* and is utilized to build a *Directed Acyclic Graph (DAG)* between predictors and yield. This mathematical construct is implemented in a freely available reasoning engine for graphical models, SMILE, along with its graphical user interface (GUI),

GeNIe[10]. We curated an extensive set of historical data ($1948 - 2012$) for each of the 99 counties in Iowa for use as training data for the model. This historical weather data ($1948 - 2012$) was tediously collected from several public sources such as the National Agricultural Statistics Service (NASS), and included weather, topographic/soil, and some management traits. We utilize expert knowledge for variable selection and for graph pruning, and present promising initial results. Results include yield forecasts for all counties and a discussion of prediction accuracy; an illustration of how prediction is possible with incomplete information, and the possibility of a probabilistic graphical model to perform what-if scenario analysis.

2.3 Methodology

Corn yield depends on a complex set of economical, meteorological, agricultural and financial inputs. These inputs are most likely interdependent. Formulating a 'mechanistic model '(i.e. 'knowledge–based' models, or those based on mathematically defined equation(s)) relating inputs with output seems (currently) intractable. However, there is a large amount of historical data across geographical regions available that can be used to make future yield prediction. The availability of a corpus of historical data along with advances in 'gray box' machine learning models motivate us to utilize this approach to yield prediction. Probabilistic graphical models (PGM's) are an example of such 'gray box' machine learning (ML) models that are helpful in capturing conditional and causal dependencies; spatially, temporally and spatial-temporally. PGM's naturally allow for incorporation of expert knowledge and derive scientific understanding form the learnt models. Inference process in such Bayesian networks can be used for prediction and also for exploring *What-if* scenarios; thus allowing us to perform inference on specific explanatory variables and observing changes in trends. PGM's are also scalable and are capable of handling large data sets. More attractively, they are capable of working with missing and conflicting data, and can inherently handle uncertainty. We outline a schematic of our workflow in Figure. 2.1.

2.3.1 Data collection and curation

The focus of the data collection was getting a historical record of various explanatory variables and county yields for the 99 counties of the state of Iowa. We divided this task into two stages: 1) Collecting raw data from a variety of sources, and 2) Data curation, to organize the collected raw data in a form that is compatible with the machine learning framework, GeNIe. The weather data is taken from the Global Historical Climatology Network (GHCN) database which is hosted by the National Climatic Data Center (NCDC). We chose to utilize weather data from the months of May - September. This choice simply tracks the corn growing season over most of the corn belt region across Iowa. We assume that explanatory variables of time periods outside the growing season have negligible effect on end-of-season yield harvest. Relaxation of such assumptions will be explored in the future. The county scale soil data is taken from the Soil Survey Geographic (SSURGO) database that is hosted by the USDA. The collected data was then post-processed into expert knowledge derived variables – specifically, aggregating daily temperatures into monthly averages, converting daily temperature into Growing Degree Days (GDD), an agronomic means of keeping track of heat. Further details of the data set, along with descriptions of each derived variable are provided later in the text. Data is curated for 99 counties over a time period of 64 years (1948 to 2012). The total dataset collected has an approximate size of 500 MB and is stored in comma-separated values (CSV) file format. Our preliminary results are based on a subset of this data. We focus on a recent six year duration of 2005–2010, with 5 years used as training data, and the data from 2010 used as testing data to explore the model's predictive capability.

2.3.2 Variable selection and preprocessing

Variable selection is critical to the construction of a viable yield predictor. We utilize expert knowledge (via agronomic arguments) to chose a subset of all possible inputs affecting yield in order to construct our probabilistic graphical model. We detail each variable and the rationale for the specific choice next.

2.3.2.1 Growing Degree Days (GDD) or Heat Units

The growth rate of corn is highly dependant on temperature. Ideal temperature conditions for robust growth is between a minimum temperature of 50°F (10°C), upto an optimum temperature of 86°F (30°C). Growth rates have been observed to decline if temperatures do not fall within this range. The Growing Degree Days (GDD) is an agronomic variable that represents the relationship between temperature and growth rate[11]. GDD is a heuristic tool in phenology that measures heat accumulation to predict development rates. GDD is given by

$$GDD = (T_{max} + T_{min})/2 - T_{base}$$

where,

- T_{max} is the maximum daily temperature or equal to 86°F (30°C) when temperature exceed beyond 86°F (30°C).
- T_{min} is the minimum daily temperature or equal to 50°F (10°C) when temperature falls below 50°F (10°C).
- T_{base} is the base temperature required to trigger the optimum growth.

An additional motivation to choose this variable is the possibility of integrating seed type as an explanatory variable in the future. Seed companies typically report hybrid maturity in days and in terms of GDD. These reports are linked to the expected number of days necessary to reach enough GDD (about 2700 to 3100 GDD to reach *R6* (physiological maturity)) to complete growth and development. For example, the commonly used 111 day hybrid requires approximately 111 days to attain enough GDD for harvest maturity.

2.3.2.2 Palmer Drought Severity Index (PDSI)

Drought has a critical impact on farming and yield. The Palmer Drought Severity Index (PDSI) measures the availability of moisture after precipitation and recent temperature changes. It is based on the supply and demand concept of the water balance equation and considers multiple meteorological parameters (including water content in the soil, rate of evapotranspiration, soil recharge and moisture loss from the surface layer). The PDSI has also been

used to perform spatial, and temporal correlations analysis[12]. The PDSI [1] takes a value of 0 to indicate the normal conditions, negative values indicate drought severity and positive values indicate wetland or flooded conditions.

2.3.2.3 Corn Suitability Rating (CSR2)

Soil type impacts productivity potential, and combined with weather conditions, is considered a dominant factor influencing yield. Corn Suitability Rating (CSR2) is an integrated measure based on soil mineral content, topographic features like slope gradient and slope length that indicate the suitability of the soil to grow corn. CSR2 ratings [1] varies minimally over time and usually range from 5 - 100, with higher ratings correlating to better growing conditions.

2.3.2.4 Rainfall

Precipitation is a factor that strongly affects yield. During the growing season, moisture requirements have to be met by rainfall, or through water held within the soil prior to growing season. High yield harvest within the corn belt region of the US has been due to the amount of precipitation available (>45cm) throughout the growing season. The demand for water utilization increases when the corn plant nears the tasseling stage, usually around mid-July, extending to mid-August. Note that both inadequate as well as over abundant rainfall reduce corn yields.

2.3.2.5 Data Discretization

Before any network or structure is learnt, the available dataset is first categorized into a set of bins. This data transformation is necessary since our model is based on discrete Bayesian networks where modeling of the relationship is required in a parsimonious manner. The goal is to retain the underlying relationship between the variables while reducing the effects of external disturbances that may distort the relationship. We chose to use a hierarchical discretization[14] over uniform width or uniform count. This enables automatic determination of the optimal number of bins and their widths, given the multivariate distribution of the variables.

[1] In Figure. 2.2 and 2.3, "*DI_Avg*" represent annual average PDSI values[12] and "*Soil_WA*" represent weighted average CSR2 ratings[13] for each of the 99 counties in Iowa.

2.3.2.6 Incorporating Background Knowledge

The ability to include domain knowledge in the construction of a model is one of the strong points for the probabilistic graphical modeling technique. This allows domain experts to provide quality input regarding known correlations between variables, as connections (or edges) in the graph. Domain expertise enabled us to specify a strong link between rainfall and yield. This approach also allowed domain experts to forbid connections between specific variables (either through intuition or where such lack-of-correlation has been previously shown). This is extremely useful when working with temporally-sensitive data, allowing one to forbid connections from future observations to past observations. It is also important for the scalability of the structure learning stage. Furthermore, it allows the sorting of variables in temporal tiers, which also forbids future to past connections. Figure. 2.2 displays the implemented background

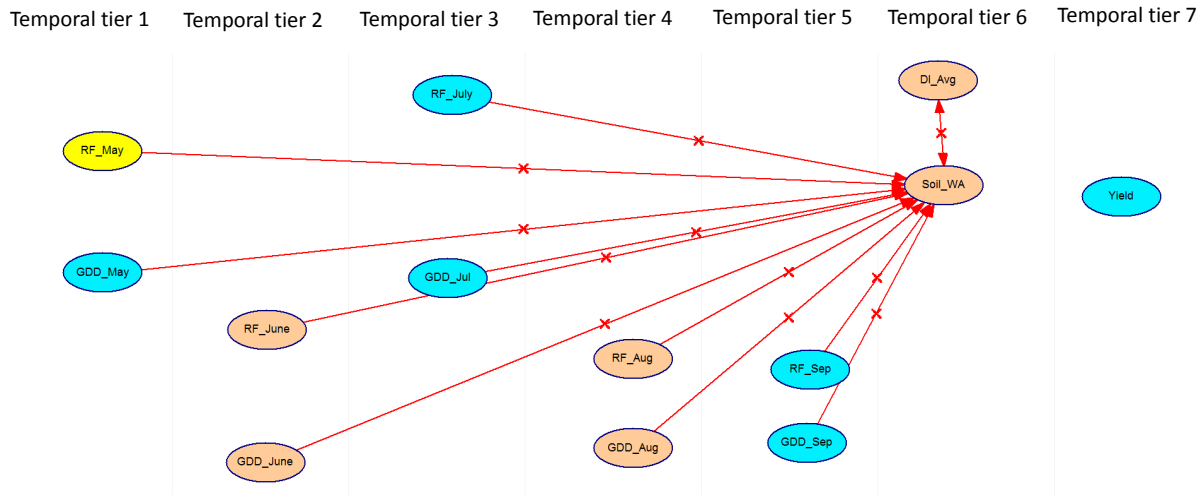

Figure 2.2 Tiering and partial enforcing of Bayesian Network Structure with Prior Background Knowledge

2.3.3 Learning and inference

Learning and inference are the two main steps associated with graphical models such as Bayesian networks. Learning refers to training the probabilistic graphical model with the training data and the inference step involves decision making using the trained model and testing

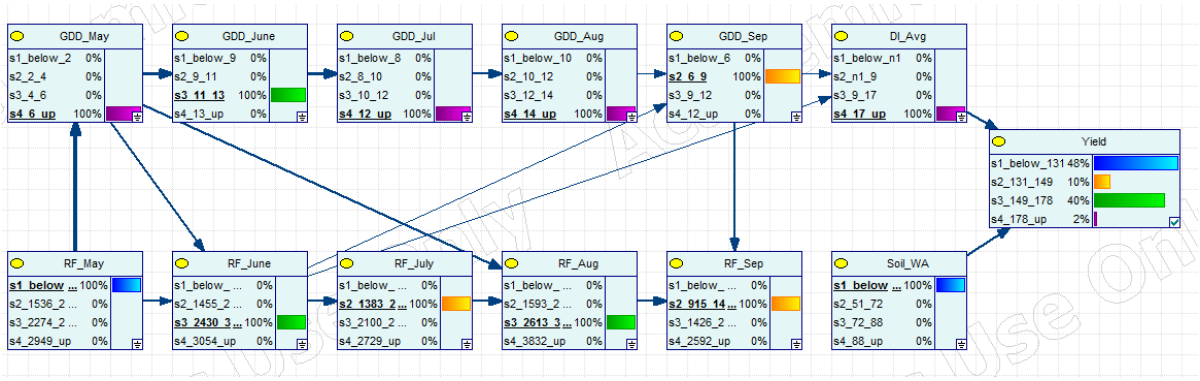

Figure 2.3 Illustration of the learnt Bayesian Network Structure based on Background knowledge

data/evidence. Learning/training involves identifying the structure (the DAG, or the edges of the graph) and learning the parameters (the edge weights), i.e., the conditional probability densities. The goal is to identify the structure and the associated parameters that best explain the given training data.

Given a Markovian set of variables $\mathbf{x} := (x_1, \cdots, x_l)$, a DAG, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a $\mathbf{P}_\theta$ where $\mathcal{V}$ describes the set of nodes in the model, $\mathcal{E}$ gives the edges connecting nodes. $\mathbf{P}_\theta(\mathbf{x})$ represents the joint probability distribution factored on the variables given their parent nodes and θ describes the parameters learnt in the factoring process. More detailed descriptions of such models are available in vast amount of literature[15; 16]. Mathematically, the aim of the learning task is to determine the optimal set of $(\mathcal{V}, \mathcal{E})$ as well as θ that describes the relationship embedded in the factors and the class variable (in this case, yield). Finding the optimal Bayesian network structure is an NP-hard problem, but efficient algorithms are available that often yield near optimal solutions[17]. Bayesian networks support learning in supervised as well as in unsupervised settings, and thereby can be used with both labeled and unlabeled data sets (such as for knowledge discovery).

In this study, after discretizing the training data, we learned a network structure (Directed Acyclic Graph) that maximizes the likelihood of observing the training data. As mentioned earlier, finding such a DAG is an NP-hard problem, hence we used efficient heuristics to approximate the underlying structure. Also, we sought expert knowledge in order to make the structure search more efficient. This knowledge elicitation helps the algorithm to streamline

its connectivity search since we forbid some unreasonable links and force links where we have information related to conditional dependencies among variables. It is important to penalize dense structures as they typically lead to over-parameterization and hence, over-fitting (bias-variance tradeoff). To address this tradeoff, we track the Bayesian Information Criterion (BIC) to drive our search for the best DAG. A set of scoring functions such as minimum description length, MDL, Bayesian-Dirichlet functions and their variations[3] for learning DAG structures were introduced in[18]. Figure. 2.3 shows the Bayesian Network structure that was learned via GeNIe toolbox on the so far curated training dataset. Note, the thickness of an edge between a pair of nodes reflects the degree of statistical dependency between those nodes i.e., strength of influence[17].

Inference pertains to finding probabilistic answers to user specified queries. For example, a user may seek the joint distribution of a subset of random variables given the observed values of other independent subsets of the random variables. Since Bayesian networks only encode node-wise conditional probabilities, finding answers to such queries is not straightforward. However, efficient algorithms exist that allow one to find the exact answer to an arbitrary query using a secondary structure (such as junction tree) and a message-passing architecture[17].

GeNIe has in-built support for various learning algorithms. In this paper, we employed the Bayesian search algorithm to train the model. It is a general purpose graph structure learning algorithm that makes use of the Bayesian search procedure to explore the full space of graphs, $\mathcal{G}$. In this case, the posterior probability tables are filled out using expectation maximization algorithm,

$$\arg\max_{\mathcal{G}} P(\mathcal{G}|D)$$

given the data, D. The aim of the algorithm is to run partial search over Markov equivalence class of the data instead of directly searching over the full DAGs space to reduce the computation time. Note that a Markov equivalence class[16] is a subset graph class that contains both directed and undirected edges, i.e., it is a set containing all the DAGs that are Markov equivalent to each other.

In the implementation of Bayesian search in GeNIe[2], we added background knowledge by

[2] http://www.bayesfusion.com/

forbidding 20 edges. The tiering edges ($i- > tier$) that associates nodes with particular tier in the 7–tier model is shown in Figure. 2.2.

2.3.3.1 Expected yield prediction

Given that the model structure and the parameters of a DAG have been learnt, it is necessary to make inferences on the model by getting forecast of yield in terms of expected yield. Accuracy of the model is tested based on the available evidence to calculate the difference in the predicted and actual yield. Given, historical values of yield Y (in bu/ac), we define $\hat{Y}$ as the expected yield prediction provided that we have computed the posterior distribution $P(b_n)$ during the inference process where b_n is the n^{th} bin signifying a certain range of yield. With this setup, we have

$$\hat{Y} = \sum_{n} P(b_n) \cdot E(Y|b_n)$$

where,

- $n \in \{1, \cdots, 4\}$ denotes the discrete bin for the yield variable.
- $P(b_n)$ denotes the probability of yield being in the range marked by bin b_n.
- $E(Y|b_n)$ represents the expected yield in the bin b_n computed based on the training data.

2.4 Results and Discussion

In this section, initial results are presented for the Bayesian network based county level yield prediction approach. We used 2005–2009 data in this study and the data set was divided into a training and testing set. While 75% of the data was used for learning the Bayes Net structure and parameters, the remaining 25% was used to provide an in-sample validation for the model. The validation set is used to determine the effectiveness of the model; to estimate its accuracy and the confidence level; to analyze performance with incomplete and complete evidence and to examine various 'what-if' scenarios as described below.

2.4.1 Yield prediction

The effectiveness of our model is described using a confusion matrix shown in the Table 2.1. It shows the overall capability of the model to correctly categorize predicted yields in the validation set into the appropriate bins, i.e., yield prediction ranges.

Table 2.1 Confusion Matrix with four yield level classes

	Predicted yield (in Bu/ac)			
True yield (in Bu/ac)	0–131	131–149	149–178	178–Above
0–131	6	0	0	0
131–149	4	11	0	0
149–178	0	1	14	7
178–Above	2	0	6	46

While most of the data is in the diagonal (i.e., correct prediction), some of the estimated yields fall into the wrong bins. However, in most cases the miss-predictions fall into neighboring bins which suggests small errors. Moreover, this current study uses an incomplete set of explanatory variables and we are currently expanding the set of variables to utilize cumulative effects of temperature and localized effects of rainfall.

Table 2.2 Difference between Predicted and Actual Yield at a county level

County	**Actual Yield Bu/ac**	**Predicted Yield (Bu/ac)**	**Difference (%)**
Shelby	171.6	171.71	0.06
Bremer	174.6	174.39	0.12
Palo Alto	174	174.39	0.22
Calhoun	173.3	174.39	0.63

Table 2.2 displays sample results of expected yield (as described in 2.3.3.1) obtained from the model. The model was used to predict yield in all 99 counties of Iowa in 2010 and overall, predicted yield for 70 out of the 99 counties had an accuracy of 80% or more. This illustrates the yield prediction potential of a Bayesian Network model with reasonable explanatory variables

and domain knowledge embedding. However, this is still an on-going effort and we are working to include more key variables and domain knowledge for better prediction accuracy.

2.4.2 Prediction with partial and complete evidences

Table 2.3 Table showing the effects of gradual addition of evidence on selected counties yield prediction accuracy

Evidences	Time Period	**County**	**Actual Yield (Bu/ac)**	**Predicted Yield (Bu/ac)**	**Difference (%)**
GDD & RF	May–June	Polk	139.40	167.91	30
GDD & RF	May–July	Polk	139.40	167.91	30
GDD & RF	May–August	Polk	139.40	167.91	30
GDD & RF	May–September	Polk	139.40	165.55	29
GDD, RF, PDSI & CSR2	May–September	Polk	139.40	140.88	2

The ultimate goal of this research is a publicly available high quality yield prediction tool that will enable the producers to make informed decisions. From this perspective, the tool needs to start predicting yield estimates from early part of the season and aim to improve the prediction as season moves forward and more observations are used as evidence. In this context, Bayesian network is an ideal inference framework as it can function with missing variables/data unlike many other approaches such as standard regression. We investigated the yield prediction performance in the absence of complete evidence–that is, before the end of the growing season, where information on future weather conditions is unavailable. Note, in such a scenario, a model can still use future weather predictions which can potentially help such a tool positively. However, we did not consider availability of any such predicted weather conditions in this study. In this case study, initial (incomplete) evidence includes only the growing degree days (GDD) and rainfall (RF) for the months of May–June. Then as the season progresses, we added evidence from months of July, August and September respectively. Furthermore, we added key variables such as PDSI and CSR2 at the final stage to examine the improvement in yield prediction performance.

The effect of incomplete evidence for Polk county is shown in the Table. 2.3. With initial limited evidence, the model is capable of providing a reasonable estimate of yield and as expected, performance improves with added evidence and finally with complete evidence[3], the computed yield comes very close to the actual yield (lagging the actual by only $\approx 1(Bu/ac)$). This is an illustration of how a Bayesian Network based tool can be leveraged seamlessly for continuous yield prediction throughout the growing season.

2.4.3 What-If Scenarios

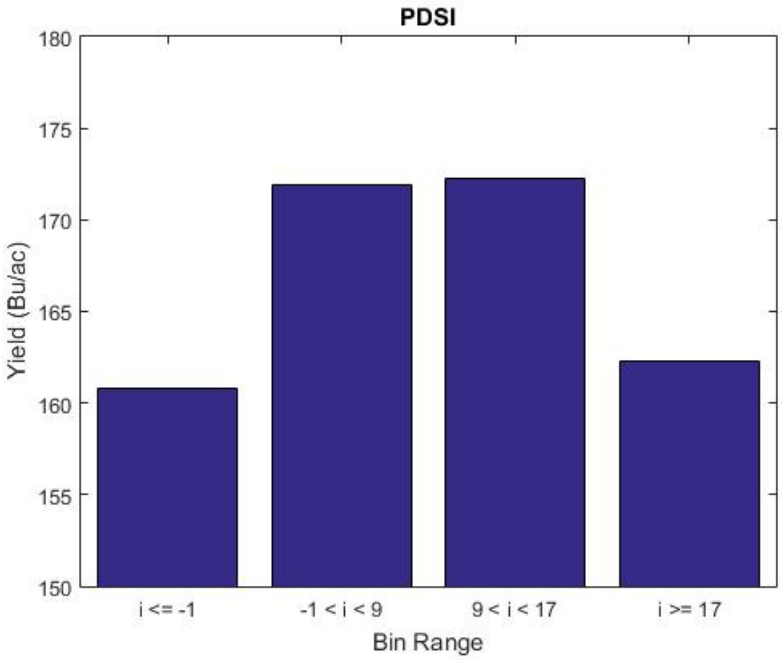

Figure 2.4 Histogram of inference on expected yield of PDSI

Farmers and plant scientists are extremely interested in learning key driving variables and parameters that affect yield. In this context, a probabilistic graphical model such as Bayesian Network can be an effective tool to understand the impact of different variables (e.g., weather) on a certain target variable (e.g., yield). Such an inference exercise is called simulation of 'what-if' scenarios and a few examples are provided below:

[3]Note that the term complete evidence in this case is based on the data available for this study which is far from being exhaustive.

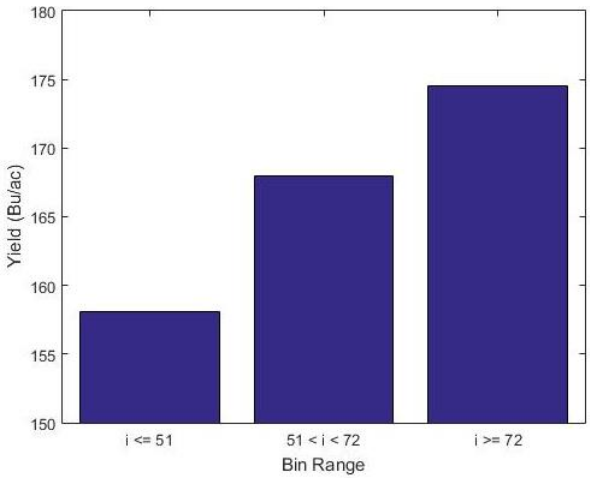

Figure 2.5 Histogram of inference on expected yield of CSR2

It is known that a host of the climatic factors lead to drop in expected corn yields at extreme conditions. A good example to support that is the effect that PDSI, described in subsection 2.3.2.2, has on the estimated yield. Figure 2.4 shows the result of a 'what-if' scenario simulation where bins 1 and 4 for PDSI lead to lower yield compared to bins 2 and 3. Note, bins 1 and 4 suggest highly negative or highly positive PDSI values which indicate extreme drought or extreme wet conditions respectively whereas bins 2 and 3 contain PDSI values that are around zero which indicate a close to ideal condition. Thus the Bayes Net inference result conforms with the scientific knowledge that extreme dry or extreme wet conditions are both bad for corn yield.

In addition to PDSI, the effect of CSR2 on yield is examined and the result is shown in Figure. 2.5. There is a reasonable positive correlation between the CSR2 values and expected yield confirming the domain knowledge of farmers and plant scientists.

Another example is shown in Figure. 2.6 where increased rainfall in July tends to help corn production slightly. In summary, a Bayesian Network model is not only useful for yield prediction but also effective for understanding various causal effects (unlike different black box models) that can enhance the scientific knowledge in this domain.

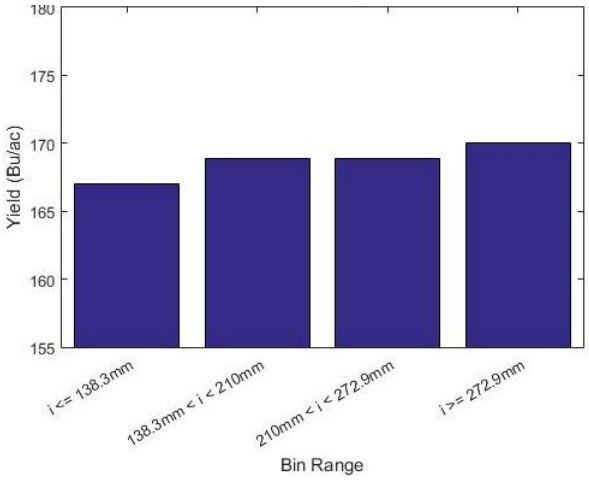

Figure 2.6 Histogram of inference on expected yield of rainfall in july

2.5 Summary, Conclusions and Future Work

In this paper, we demonstrated a Bayesian Network approach in order to predict county-wide yield in the corn belt state of Iowa, primarily utilizing historical weather data. Apart from the yield prediction capability with incomplete and complete evidence, key advantages of such an approach include ability to incorporate domain knowledge, enhance scientific understanding via 'what-if' scenario simulation and naturally provide a prediction confidence. In the case study presented here, the model performed reasonably well based on its validation accuracy. Example 'what-if' scenarios involving PDSI, CSR2 and rainfall in July show effectiveness of this approach in enhancing scientific understanding. We also demonstrated the capability of yield prediction based on incomplete and complete evidence which makes it a useful tool for continuous yield prediction throughout the season. While the main future goal of this research is to be able to accurately predict yield within 5 Bu/ac of the actual yield in every county, many other technical aspects are being pursued as well such as (i) incorporation of cumulative weather variables, (ii) handling different time-scales of different explanatory variables and (iii) establishing a model adaptation mechanism along with climate change patterns.

2.6 Acknowledgments

Vikas Chawla and Baskar Ganapathysubramanian thank ISU PSI for support through the PSI faculty fellow. All authors thank the ISU PIIR DDSI funding for partial support.

CHAPTER 3. CONCLUSION AND FUTURE WORK

In this thesis, we have demonstrated a Bayesian network approach in order to predict county-level yield in the corn belt state of Iowa by primarily utilizing the historical weather data and expert knowledge. This approach has some key advantages:

- Ability to incorporate domain and expert knowledge:

  Finding a solution to a graphical model is considered to be an NP-hard problem thus it's important to have a model (like Bayes net) which is capable enough to include the domain specific knowledge prior to training the model. This not only optimizes the problem but also improve the prediction accuracy.

- Enhance scientific understanding via what-if scenario inference:

  Example what-if scenarios involving PDSI, CSR2, and rainfall in July show effectiveness of this approach in enhancing scientific understanding.

- Naturally provide strength of influence between parameters:

  The model not only predict the yield given other parameters but it also learns the weights for each edge it draws between different nodes (i.e. variables). Higher the value of weight shows the strong dependence of variables on each other and this gives us the information about the parameters that has a strong influence on the yield prediction.

The future work is focused towards accurately predicting the yield for every county within 5 Bu/Ac of its actual value. Many other technical aspects are being pursued as well such as:

- Incorporation of cumulative weather variables. Some of the potential variables are:

  - Humidity:

    The relative humidity can affect the flow of water through the plant and affect the transpiration rate.

  - Irradiation:

    Different type of radiations also plays an important role in optimal plant growth.

  - Wind speed:

    Wind speed directly influences the kernel development in a corn plant. Extreme conditions can not only destroy the plant but it also left ears partially filled with ripe kernels.

  - Snowfall:

    The Amount of snowfall helps in determining the ground level water.

- Handling different time-scales of different explanatory variables

  - Exploring Markov symbol dynamics

    To study the dynamic nature of certain parameter (like hourly temperature) in order to analyze the data often requires the need to study the topologically equivalent system using symbol dynamics representing trajectories by infinite length sequences using a finite number of symbols[19]. The state space is divided into different finite bins and assigns a symbol to each one. This kind of transformation is known as *markov transformation.*

- Establishing a model adaptation mechanism along with climate change patterns

BIBLIOGRAPHY

[1] Wikipedia. Food Security. https://en.wikipedia.org/wiki/Food_security#Approaches. [Online; accessed 11-27-2016].

[2] Charles L. Hornbaker II and J. Benjamin Cook. *Predicting Yield in the Corn Belt*. [Online; accessed 05-07-2016].

[3] Nathaniel K. Newlands and Lawrence Townley-Smith. Predicting energy crop yield using bayesian networks. *Proceedings of the Fifth IASTED International Conference, Computational Intelligence*, pages 107–112, August 2010.

[4] Vikas Chawla, Hsiang Sing Naik, Adedotun Akintayo, Dermot Hayes, Patrick S. Schnable, Baskar Ganapathysubramanian, and Soumik Sarkar. A bayesian network approach to county-level corn yield prediction using historical data and expert knowledge. *In Proceedings of the 22nd ACM SIGKDD Workshop on Data Science for Food, Energy and Water, 2016*, abs/1608.05127, 2016.

[5] Wyatt Thompson Scott Gerlt and Douglas J. Miller. Exploiting the relationship between farm-level yields and county-level yields for applied analysis. *Journal of Agricultural and Resource Economics*, 39(2):253–270, August 2014.

[6] Tian Yu. *Three essays on weather and crop yield*. PhD thesis, Economics, Iowa State University, Ames Iowa, 2011.

[7] Chad Lee and Jim Herbek. Estimating Corn Yields (AGR-187). http://www2.ca.uky.edu/agcomm/pubs/agr/agr187/agr187.pdf. [Online; accessed 05-07-2016].

[8] Charles L. Hornbaker II and J. Benjamin Cook. Predicting Yield in the Corn Belt. http://iacs-courses.seas.harvard.edu/courses/iacs_projects/BenCookCorn/ac299r/assets/pdf/cook-stat225.pdf. [Online; accessed 05-07-2016].

[9] Kefaya Qaddoum. Modified naive bayes based prediction modeling for crop yield prediction. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, 8(1):36–39, 2014.

[10] Marek J. Druzdzel. Smile: Structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. *AAAI '99/IAAI '99 Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, (ISBN:0-262-51106-1):902–903, 1999.

[11] Lance R. Gibson. Growing degree-day calculation. http://agron-www.agron.iastate.edu/Courses/agron212/Calculations/GDD.htm. [Online; accessed 05-08-2016].

[12] Brian Fuchs. Palmer Drought Severity Index (PSDI and scPDSI). http://drought.unl.edu. [Online; accessed 05-26-2016].

[13] Iowa State University Extension and Outreach. Iowa CSR2 Weighted Means by County. http://www.extension.iastate.edu/soils/suitabilities-interpretations. [Online; accessed 05-07-2016].

[14] Randy Kerber. Chimerge: Discretization of numeric attributes. *Proceedings of the tenth national conference on Artificial intelligence*, pages 123–128, 1992.

[15] Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham M. Kakade. Learning linear bayesian networks with latent variables. *Proceedings of 30th International Conference on Machine Learning*, 28:1 – 9, 2013.

[16] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3921–3962, November 2014.

[17] Sudha Krishnamurthy, Soumik Sarkar, and Ashutosh Tewari. Scalable anomaly detection and isolation in cyber-physical systems using bayesian networks. In *ASME 2014 Dynamic Systems and Control Conference*, pages V002T26A006–V002T26A006. American Society of Mechanical Engineers, 2014.

[18] Brandon Malone. Scoring functions for learning bayesian networks. Online, February 2014.

[19] Erik M. Bollt and Joe D. Skufca. *Markov Parititions.* [Online; accessed 05-07-2016].