12-2012

Analysis of Irrigation Decision Behavior and Forecasting Future Irrigation Decisions

Sanyogita Andriyas
Utah State University

ANALYSIS OF IRRIGATION DECISION BEHAVIOR AND FORECASTING FUTURE IRRIGATION DECISIONS

by

Sanyogita Andriyas

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

Approved:

Dr. Mac McKee Major Professor	Dr. Wynn R. Walker Committee Member
Dr. Christopher M. U. Neale Committee Member	Dr. David Stevens Committee Member
Dr. DeeVon Bailey Committee Member	Dr. Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2012

ABSTRACT

Analysis of Irrigation Decision Behavior and Forecasting Future Irrigation Decisions

by

Sanyogita Andriyas, Doctor of Philosophy

Utah State University, 2012

Major Professor: Dr. Mac McKee
Department: Civil and Environmental Engineering

Farmers play a pivotal role in food production. To be economically successful, farmers must make many decisions during the course of a growing season about the allocation of inputs to production. For farmers in arid regions, one of these decisions is whether to irrigate. This research is the first of its kind to investigate the reasons that drive a farmer to make irrigation decisions and use those reasons/factors to forecast future irrigation decisions. This study can help water managers and canal operators to estimate short-term irrigation demands, thereby gaining information that might be useful in management of irrigation supply systems. This work presents three approaches to study farmer irrigation behavior: Bayesian belief networks (BBNs), decision trees, and hidden Markov models (HMMs). All three models are in the class of evolutionary algorithms, which are often used to analyze problems in dynamic and uncertain environments. These algorithms learn the connections between observed input and output data and can make predictions about future events. The models were used to study behavior of farmers in the Canal B command area, located in the Lower Sevier River Basin, Delta, Utah. Alfalfa, barley, and corn are the major crops in this area. Biophysical variables that are measured during the growing seasons were used as inputs to build the models. Information about crop phenology, soil moisture, and weather variables were compiled. Information about timing of irrigation events was available from soil moisture probes installed on some agricultural fields at the site. The models were capable of identifying the variables that are important in forecasting an irrigation decision, classes of farmers, and decisions with single and multi-factor effect regarding farmer behavior. The models did this across years and crops. The advantage of using these models to study a complex problem like behavior is that they do not require exact information, which can never be completely obtained, given the complexity of the problem. This study uses biophysical inputs to forecast decisions about water use. Such forecasts cannot be done satisfactorily using survey methodologies. The study reveals irrigation behavior characteristics. These conform to previous beliefs that a farmer might look at crop conditions, consult a neighbor, or irrigate on a weekend if he has a job during the week. When presented with new data, these models gave good estimates for probable days of irrigation, given the past behavior. All three models can be adequately used to explore farmers' irrigation behavior for a given site. They are capable of answering questions related to the driving forces of irrigation decisions and the classes of subjects involved in a complex process.

(134 pages)

PUBLIC ABSTRACT

Analysis of Irrigation Decision Behavior and Forecasting Future Irrigation Decisions

by

Sanyogita Andriyas, Doctor of Philosophy

Utah State University, 2012

Major Professor: Dr. Mac McKee
Department: Civil and Environmental Engineering

Farmers play a pivotal role in food production. To be economically successful, farmers must make many decisions during the course of a growing season about the allocation of inputs to production. For farmers in arid regions, one of these decisions on any given day is whether to irrigate. This research is the first of its kind to investigate the probable reasons that lead a farmer to make irrigation decisions and use those reasons/factors to forecast future irrigation decisions. This study can help water managers and canal operators to estimate short-term irrigation demands, thereby gaining information that might be useful to more efficiently manage irrigation supply systems. This work presents three approaches to study farmer irrigation decision behavior: Bayesian belief networks, decision trees, and hidden Markov models. All three models are in the class of evolutionary algorithms, which are often used to analyze problems in dynamic and uncertain environments. These algorithms learn the connections between measured input and output data and can make predictions about future events. The models were used to study irrigation decisions of a set of farmers in the Canal B command area, located in the Lower Sevier River Basin, Delta, Utah. Alfalfa, barley, and corn are the major crops in this area. Biophysical variables (plant, soil, canal flow, and weather conditions) that are measured during the growing seasons were used as inputs to build the models. Information about crop phenology (growth stages), soil moisture, and weather variables were compiled. Information about timing of irrigation events was available from soil moisture probes (which measure soil moisture content) installed on some irrigated fields at the site. The models were capable of identifying the variables that are important in forecasting an irrigation decision, classes of farmers, and decisions with single and multi-factor effect regarding farmer behavior. The models did this across years and crops. The advantage of using these models to study a complex problem like behavior is that they do not require exact information, which can never be completely obtained, given the complexity of the problem. This study uses biophysical inputs to forecast decisions about water use. Such forecasts cannot be obtained satisfactorily or in a cost-effective manner using survey methodologies. The study reveals irrigation behavior characteristics. These conform to previous beliefs that a farmer might look at crop conditions, consult a neighbor, or irrigate on a weekend if he has a job during the week. When presented with new data, these models gave good estimates for probable days of irrigation, given the past behavior. All three models can be adequately used to explore farmer irrigation decision behavior for a given site. They are capable of answering questions related to the likely driving forces behind irrigation decisions and the classes of subjects involved in a complex process.

(134 pages)

ACKNOWLEDGMENTS

I would take this opportunity to express my gratitude to my major professor, Dr. Mac McKee, for his continuing help, patience, and guidance throughout the course of this work. In spite of his busy schedule and travel, his invaluable inputs throughout the program only took me up the ladder. My gratitude is also due to Dr. Walker for helping us in this project and giving us valuable information about the area. I would also thank my committee members, Drs. Bailey, Neale, and Stevens, for their help time and again.

This work would not have been possible without the telephonic conversations with Shane Clark, Water Master, and Megan Greathouse, DMAD. I would also not forget to mention Jan Urroz, who helped me get out-of-schedule appointments to meet Dr. McKee, and other UWRL staff, Tamara Peterson, and Andrea Carroll.

I would also like to thank my mentors Dr. Harold Schroeter and Dr. Ali H. Sajid from Canada who had a lot of belief in my capabilities.

Staying away from home would have been difficult without the unstoppable mental support from my brother Tushar Andriyas. I am highly obliged to my parents, Mithlesh and Fredrick Andriyas, who made many sacrifices to see me where I am today. Their encouragement and undying faith in my strengths was a huge motivation to continue my work. There were many hard times and it was God on high who helped me pull through them successfully. It was He who gave me health and mind to accomplish this work. At the end, I find myself lucky to have been aligned with so many good things at a given time.

-Sanyogita Andriyas

To God Almighty, my parents - Mithlesh and Fredrick Andriyas, brother - Tushar,
pet parrot, and my grandparents - Alfreda and the late Harold Singh.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Agriculture dominates global water use. According to a World Bank report, irrigated agriculture accounts for 18 % of all the cropped land area, which is about 277 million ha. This land area produces approximately 40 % of all the crops in the world. With a steady population growth worldwide and limited land area, it will become more difficult in the future to meet food production requirements, especially when competing water uses are given priority in times of scarcity. This makes it imperative to utilize these scarce resources to the fullest, and to save water for future use. Laser leveling, mulching, and fallowing are some ways to save water in situ for irrigated lands. There are various interventions, in form of participatory stakeholder workshops, to help in better water management. But there is a huge difference in the objectives of those workshops with that of the farmers which creates a big gap between the two. For the success of such mediation, it is important to know the preferences of the farmers and their attitudes towards water use. According to Uphoff (1986), irrigation decision-making cannot be adequately represented, in isolation, by a single disciplinary perspective. Various fields need to be integrated to impart a collective perspective to the decision and policy makers.

The on-farm allocation of water, i.e. deciding whether or not to irrigate, when, and determining the amount, is only a part of irrigation behavior. Fahim and Rady (1989) found that the behavior farmers is affected by educational level, size of land owned, costs and delivery of water to the farm, economic status, and the experience of the farmer with irrigation practices. The farmer, soil, plant and water are the main pillars of food production. Another aspect to this is the availability of natural free-water, in the form of precipitation, which is random and non-stationary. Coupled with this is the uncertainty of water demand at a given time, and the evaluation of whether the crop stage is critical with regard to water stress. All these factors are important, but it is the farmer who controls all the inputs in this dynamic system. Hence the decisions made by farmer on the operation of the system are important to understand.

It is important to recognize that farmers are faced with a constrained multiple-criteria objective function. The intervals between irrigations and the amounts; uncertain time for harvest of the crop from the field; constrained and legally binding water rights for the season; and imposition of land preparation, inter-culture and labor costs as required make the process stochastic. Thus it is difficult to model the decision process surrounding irrigation using a deterministic framework. Past studies, such as by Becu et al. (2006) and Le Bars et al. (2005), simulated farmers behavior by using multi-agent systems and proved that stochastic models are an alternative to handle these problems well.

The analysis of irrigation decisions is important because this can help in the estimation of short-term irrigation demands. If the probable farmer decisions are known, it can help canal operators to better manage water deliveries and avoid unexpected delays and operational conditions that increase canal losses. Information about these demands can also be helpful for the evaluation of expected future agricultural supplies.

In the first place, very limited information is available which can be used to study farmers irrigation decisions. Biophysical data is widely available to help farmers make informed decisions, but little socioeconomic data are available about irrigation decisions. Biophysical data can be used to build models and make forecasts about irrigation decisions. The main advantage with such information is that they are representative of the conditions under which farmers irrigate or do not irrigate, and can therefore provide inferences into irrigation behavior if appropriate models of such behavior can be formulated. It can never be possible to know the exact reasons why a farmer decided to irrigate, however, and all farmers are different and prefer their own decision processes over those of others. This study used biophysical data to infer probable farmer actions. This data was used to build the models and these learnt frameworks were used to predict irrigation decisions.

1.1 Study Objectives

The purpose of this study was to investigate potential evolutionary algorithms to discover whether they could give useful insights into farmer irrigation behavior. The specific objectives for this work were to:

1. Identify the important variables contributing to a farmer's irrigation behavior by training the models with relevant data from that farmers field.

2. Group the irrigation decisions into distinct classes and categorize the farmers into different types using the selected models.

3. Identify the decisions taken on the basis of a single critical factor as well as those where multiple factor interactions lead to the decision.

4. Discern the patterns in decisions from irrigation-to-irrigation, crop-to-crop, year-to-year using evolutionary algorithms.

5. Infer a farmers likely future irrigation decisions using the information and modeling tools, above.

1.2 Research Rationale

Past work in irrigation behavior aims only at simulating scenarios and recreating irrigation behavior from generic principles. There is no study in the literature which analyzes farmers irrigation actions and develops a tool to predict future actions.

The study site selected has relevant information that shows potential to infer indicators of irrigation decisions. The site has real-time information for canal flow rates and soil moisture measurements from some fields located in the Canal B command area. The productivity in the region is high and depends mostly on irrigation. The farmers are well-mechanized and up-to-date with present technologies in agriculture.

The irrigation decisions can help to estimate short-term irrigation demands which can be important for canal operators to manage resources efficiently and deliver water to the agricultural fields without delays.

1.3 Research Significance to the Field of Agriculture

This research presents an innovative effort using decision analysis approaches to predict future water management actions that contain significant uncertainty. This could potentially serve to help improve canal operation and describe how farmers make irrigation decisions. The study contributes to the field of water management in agriculture by introducing probabilistic models to identify the reasons why a farmer irrigates on certain days as compared to others. It finds the crop, soil and allied conditions the farmer uses as indicators for a decision as to when to irrigate his crops. It helps investigate the possibility of categorizing farmers into groups based on various conditions of crop and type of water year.

1.4 Description about the Data Used

The study uses data from a variety of sources as documented below. The analysis of irrigation decisions in the Canal B command area have been done for years 2007-2010. Several variables have been selected for this study as possible predictors of the irrigation decision. Some of these are surrogates of other possible conditions that any given farmer might consider in determining whether to irrigate. For example, a farmer irrigating when he/she sees a neighbor irrigate is represented in this study by using the daily canal flows assuming that if a farmer irrigated on a high flow day, he might have irrigated when his neighbor irrigated.

1.4.1 Soil Moisture Content Time Series

One of the variables used in the study as a predictor of the irrigation decision was daily soil moisture content. In 2007, the Utah Water Research Laboratory (UWRL), Utah State University (USU) established 44 stations with 88 sensors to record soil moisture at 1 and 2 ft depths on various farms in Delta, Utah to study agricultural water use. The sensors are maintained by personnel at the UWRL. Soil moisture content measured at these stations was used to determine the day of irrigation and the approximate amount of irrigation. These were obtained from: http://odm.usu.edu/odmmap/default.aspx?NetworkName=Delta. Hourly measurements are available on the website, so daily average values were estimated for the first day of the season, for starting the soil moisture balance calculations. For assessing the data quality, the measured soil moisture contents were compared with standard values of porosity and field capacity from FAO-56 (Allen et al., 1998). It was found that the values were relatively higher than the literature values for silt clay loam and silt clay soil types. The possible reasons could be one of the following:

1. During the installation of probes at two different depths or subsequent maintenance, a pit is dug. This can disturb the soil layering and open up pores to

allow for more water movement which is picked up by the sensors. This can result in high water content values.

2. Subsurface flow from adjoining fields could occur.

3. Finer soil texture leads to smaller pores and more water is retained in the soil in general (Dr. Scott Jones - Environmental Soil Physics Group, USU, personal communication).

During the initial analysis of the soil moisture content data, we found some issues which raised questions about the quality of the soil moisture data. Figure 1 shows the observed data, which makes it clear that the soil moisture probes did not function well towards the onset of the next irrigation. Hence, the measurements were judged to be erroneous during the drying phase of the irrigation cycle. Shock et al. (2003) stated that neutron probes and soil water potential sensors are more sensitive to the drying phase (refer to Figure 2) than are soil moisture probes.

A soils map (Figure 3), available from the AGRC website (`http://gis.utah.gov/data/geoscience/soil/`) indicated three types of soils in the study area: silty clay loam, silty clay, and loam. The major soil characteristics governing water movement and retention are porosity, field capacity and wilting point. Porosity defines the saturation limit of a specific type of soil. Field capacity and wilting point put limits on the plant available water, which was important in our case since we were considering crop growth as well as soil water extraction.

Standard values from Allen et al. (1998), for these parameters (soil characteristics) were considered. In some cases, however, where the literature values (Allen et al., 1998) were found to be either too low or too high relative to the measurements, the values which were reasonably close to soil moisture probe readings were taken

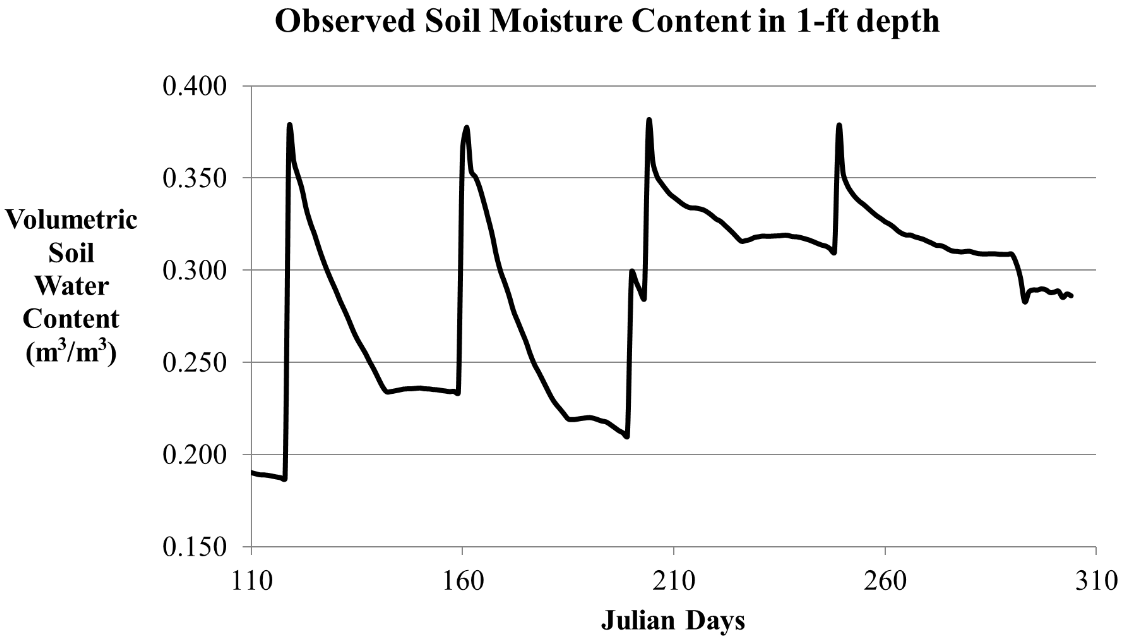


Fig. 1: Observed soil moisture content from one of the fields where soil probe was installed.

into account. This was done to closely model the field conditions. The values for soil characteristics and the day of planting used in this study are presented in Table A.1.

Taking soil-water hysteresis effects into account, the soil moisture readings were corrected by calculating a soil water balance by considering the dates of irrigation. Thus, new soil moisture curves were constructed which were reasonably close to the soil probe readings but did not have questionable data in portions of the drying phase after irrigation.

1.4.2 Meteorological Data

The weather data for the study came from the station located in Delta, Utah, which is available on the website: `http://www.cemp.dri.edu/cgi-bin/cemp_stations.pl?stn=delu`. The station was established by National Climatic Data Center (NCDC), NOAA and has historical weather data since 1965. The station metadata are as fol-

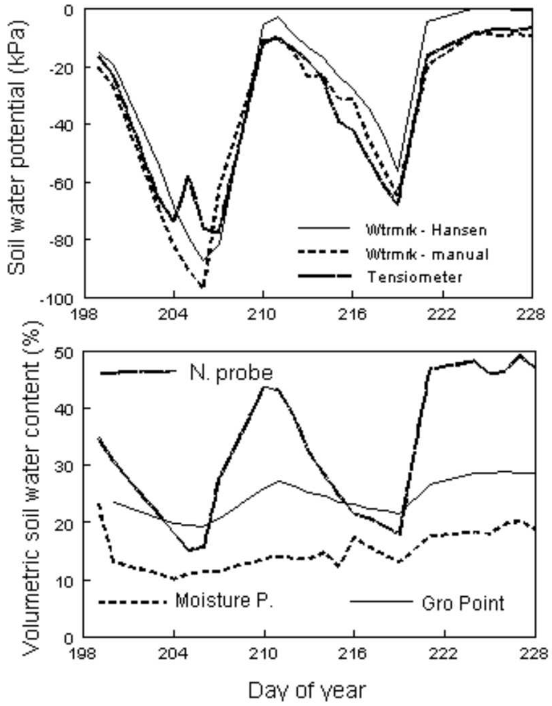


Fig. 2: Accuracy of volumetric water content read by tensiometer (top) vs. soil moisture probes (bottom). Courtesy: Shock et al. (2003)

lows and can be located on the NCDC-NOAA website (`http://www.ncdc.noaa.gov/`) using them:

GHCND ID : USW00023162, COOP ID : 422090, WMO ID : 72479, and NCDC ID : 20026236

Daily values were obtained for average air temperature (deg C), average relative humidity (percent), average wind speed (m/s), Penman evapotranspiration (mm), and precipitation (mm). For variables where cumulative values were used in the study (such as cumulative crop evapotranspiration, or CumET$_c$), the starting point was the day of planting, with cumulative values calculated up to the day of irrigation. New accumulations began the next day and continued to the subsequent irrigation decision day. For variables such as growing degree days, the cumulative values of temperature continued until the end of the season. Again, growing degree days were used as a surrogate of the progression of the crop growth stages through the season.

1.4.3 Canal Flow Rates

USU tested and introduced low-cost automated systems of canal management in 1992 in Delta according to Walker (1993). SCADA (Supervisory Control and Data Acquisition) systems maintain the past records of water discharge across Sevier River Basin (Berger et al., 2002) and real-time conditions are accessible on `http://www.sevierriver.org/`. The site is sponsored by Sevier River Water Users Association (WUA). Daily canal flows in cubic feet per second (cfs) in Canal B were obtained from the URL: `http://www.sevierriver.org/rivers/delta/b-canal/`. The flows included irrigation to some part of the area which has pastures, but since we did not know when the pastures were irrigated or what percentage of the area was covered with pastures, we did not make any adjustments to the data.

1.4.4 Market Prices

The USDA annual statistical bulletin at the following website (`http://www.nass.usda.gov/Statistics_by_State/Utah/Publications/Annual_Statistical_Bulletin/index.asp`) regularly posts the prices for the crops in a region for the 15th of the month. The prices for previous years (2007-2010) were compiled from the site. Since a single value for the whole month would not be practical to use, these values were linearly interpolated between months to obtain daily values.

1.4.5 Soil Moisture Balance and Miscellaneous Variables

Bayesian belief networks and decision tree models in this study used the components of a soil moisture balance, while hidden Markov models did not use all the components. Hence wherever they have been used, a detailed procedure for estimating the values of those components has been described. All the derived variables (with “IrrigNeed” suffix) mime the farmer’s assumed thought process. These derived variables are based on the data itself. The logic in each of the derived input variables is as follows:

1. ‘SoilIrrigNeed’ - Soil condition is one of the most important criteria for an irrigation decision. If the soil is moisture stressed (the plant available water is nearly exhausted) the ‘SoilIrrigNeed’ indicated this condition with a ‘Yes’.

2. ‘StressIrrigNeed’ - This variable resembled the stress imposed on the crop due to accumulated evapotranspiration (CumET$_c$).

3. ‘WkEndIrrigNeed’ - This variable was used to indicate a ‘Yes’ if a farmer prefers to irrigate on a weekend.

4. ‘WaterSupplyIrrigNeed’ - This variable was used to indicate if the farmer irrigates when his neighbor irrigates.

5. ‘GrowStageIrrigNeed’ - This variable indicated a crop growth stage vulnerable to moisture stress. This factor is different for different crops.

6. ‘EconIrrigNeed’ - Some farmers might irrigate to improve crop quality and hence maximize profits. This variable was used to account for this condition.

7. ‘CropIrrigNeed’ - This variable incorporates the rooting depth of a crop to help determine the depth from which the plant can extract water. This can be important since newly planted crops, such as alfalfa, stop root growth after the development stage and before first cutting, and have already stopped rooting further if they have been developing from previous years.

More details have been provided in the chapters where these variables have been used.

1.5 Dissertation Outline

The dissertation is composed of five chapters. Chapter 1 provides a foundation to this work, including the objectives and rationale behind the work, along with the significance of the research.

All the models presented in the chapters have been applied to data from the Canal B region of Delta. The classical learning and testing procedure has been used for building and assessing model capabilities.

Chapter 2 presents the first approach used to study irrigation behavior. Bayesian belief networks (BBNs) have been built using variables such as soil water balance, market prices, and canal flows.

Chapter 3 demonstrates the use of various tree algorithms to study behavior. The models have been cross-validated to refute any possibilities of over-fitting given limited amount of data. The same variables used in BBN development have been used for tree building.

Chapter 4 introduces a hidden Markov model (HMMs) to explore irrigation behavior. The observed variables were discretized into states using common irrigation principles. The output states were the irrigation decision sequence. Four factors, soil stress coefficient, depletion, canal flow, and cumulative crop ET, were identified to have some information on the irrigation behavior.

Chapter 5 provides a summary of the models developed and the lessons learned from the case study. Research challenges are discussed, some of which were overcome and others not. Final conclusions on the irrigation behavior learned from the farmers of the study site are discussed, followed by some suggestions for future work.

References

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration - guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper no. 56.

Becu, N., Sangkapitux, C., Neef, A., Kitchaicharoen, J., 2006. Participatory simulation sessions to support collective decision: the case of water allocation between a Thai and a Hmong village in northern Thailand. In: March 7-9, 2006, Chiang Mai, Thailand, International Symposium, Towards Sustainable Livelihoods and Ecosystems in Mountainous Regions.

Berger, B., Hansen, R., Hilton, A., 2002. Using the World-Wide-Web as a Support System to Enhance Water Management. In: Workshop Proceedings - Irrigation Advisory Services and Participatory Extension in Irrigation Management - Workshop organised by FAO - ICID, July 2002, Montreal, Canada.

Fahim, W., Rady, A., 1989. Influence of farmer's behavior on water management practices. In: Rydzewski, J.R. Ward, C.F. (Eds.), Irrigation Theory and Practice. Pentech Press, London, pp. 721-731.

Le Bars, M., Attonaty, J.M., Pinson, S., Ferrand, N., 2005. An agent-based simulation testing the impact of water allocation on farmers' collective behaviors. Simulation 81 (3), 223–235.

Shock, C.C., Akin, A.I., Unlenen, L.A., Feibert, E.G.B., Tschida, A., Nelson, K., 2003. A comparison of soil water potential and soil water content sensors. M. E. Station, Trans. Oregon State University, pp. 235-240.

Uphoff, N., 1986. Getting the Process Right: Improving Water Management with Farmer Organization and Participation. Westview Press, Boulder, CO, USA.

Walker, W., 1993. USU develops automated system of canal management. Utah Sci. 54 (4), 106–109. URL: http://digitalcommons.usu.edu/utscience/vol54/iss4/1.

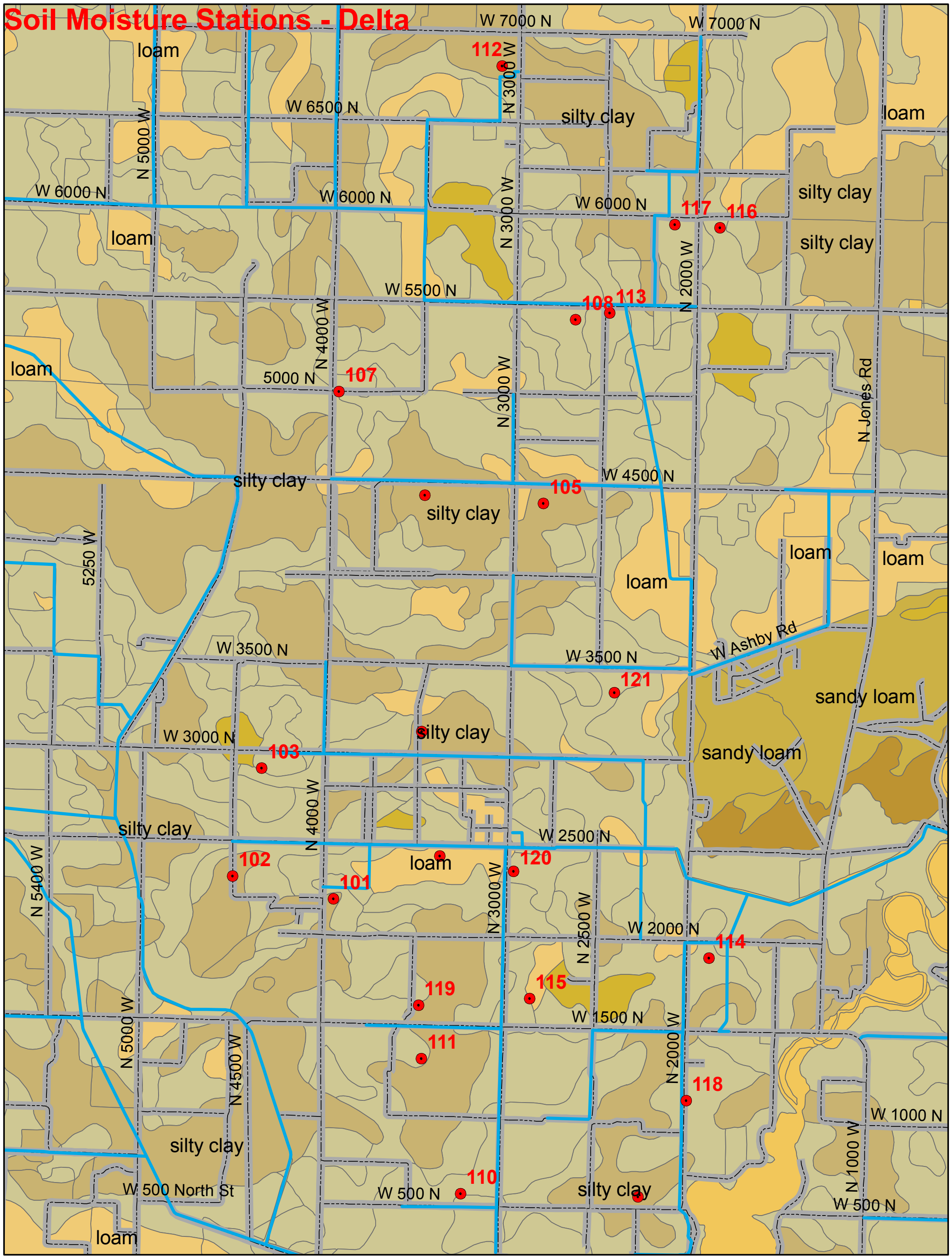


Fig. 3: Soils map for Canal B agricultural area where soil moisture probes are installed.

CHAPTER 2
DEVELOPMENT OF A BAYESIAN BELIEF NETWORK MODEL FRAMEWORK FOR ANALYZING AND FORECASTING IRRIGATION BEHAVIOR

Abstract

Canal operators need information to manage water deliveries to irrigators. Short-term irrigation demand forecasts can potentially valuable information for a canal operator who must manage an on-demand system. Such forecasts could be generated by using information about the decision-making processes of irrigators. Bayesian models of irrigation behavior can provide insight into the likely criteria which farmers use to make irrigation decisions. This paper develops a Bayesian belief network (BBN) to learn irrigation decision-making behavior of farmers and utilizes the resulting model to make forecasts of future irrigation decisions based on factor interaction and posterior probabilities. Models for studying irrigation behavior have been rarely explored in the past. The model discussed here was built from a combination of data about biotic, climatic, and edaphic conditions under which observed irrigation decisions were made. The paper includes a case study using data collected from the Canal B region of the Sevier River, near Delta, Utah. Alfalfa, barley, and corn are the main crops of the location. The model has been tested with a portion of the data to affirm the model predictive capabilities. Irrigation rules that might be predictive of observed irrigation decisions were deduced in the process of learning and verified in the testing phase. It was found that most of the farmers used consistent rules throughout all years and across different types of crops. Soil moisture stress, which indicates the level of water available to the plant in the soil profile, was found to be one of the most

likely, significant predictive variables of the irrigation decision. Irrigation decisions appeared to be triggered by a farmer's perception of soil stress, or by a perception of combined factors such as information about a neighbor irrigating or an apparent preference to irrigate on a weekend. Soil stress resulted in irrigation probabilities of 94.4% for alfalfa. With additional factors like weekends and irrigating when a neighbor irrigates, alfalfa irrigation probabilities were found to be 90.0 and 92.3%, respectively. Prediction accuracy of the date for irrigations of alfalfa was observed to be 81.0%, and 61.0% for barley and corn. The study shows that BBNs can be a prospective tool to forecast likely decisions about irrigation in an on-demand system with good accuracy.

2.1 Introduction

Irrigation is an integral part of agriculture. Crop water demand fluctuates throughout the growing season, with high demands occurring during warmer conditions. This brings an uncertainty in farmers' irrigation decisions. A reliability ability to predict a farmer's future actions could be useful in providing valuable information for better operation of irrigation canals to respond to fluctuations in short-term water demand.

Decisions that farmers make about when to irrigate are difficult to predict because they could be based upon the perceived importance of many different factors, such water rights, individual preferences, neighbor's irrigation decisions, crop type and expected future market prices. These factors make it difficult to distinguish which factors contributed to a farmers decision to irrigate, and when.

Data availability presents another difficulty in forecasting irrigation decisions. Models of irrigation decision behavior must discriminate antecedent conditions on the days leading to the day of the irrigation decision, and such discriminators are difficult to identify. A simple deterministic, physically based model of crop water requirements can prescribe when and how much a farmer should have irrigated on a certain day, but it typically cannot shed light on the inherent reasons why a farmer decides to irrigate. A soil moisture balance model would suggest irrigation occur as soon as there is stress that would be indicative of deterioration in the crop condition. Deterministic models also need estimated amounts of water deliveries, conveyance system design, system efficiencies, etc., to be able to make a reasonable forecast of irrigation practices.

To anticipate future irrigation actions, an analysis of previous irrigation practices and identification of patterns in them would be necessary. A wide range of data sources is available. These constitute scientific measurements and involve expert judgment about variables which are derived using prior experience. This is also a problem involving fields such as economics, hydrology, sociology/anthropology, and irrigation engineering. This means a model for forecasting future irrigation decisions must combine categorical and continuous variables, which is not possible in conventional approaches such as in soil moisture balance calculations. Studies have been conducted individually in all these fields with regards to farm operations, but there is no study in the literature which combines all these fields into a model for forecasting irrigation decisions.

Bayesian belief networks (BBNs) can be used to study problems that involve decision-making under uncertainty and make inference about the related behavior. Figure 4 shows a BBN with three nodes and illustrates the modeling of cause-effect types of relationships. These models can make use of available data and provide information to infer the reasons which led to the decision being modeled. Bayesian models are characterized by their simplicity, ease of interpretation, and viability. Such methods are cost-effective since they can provide results with available information about the problem. Bayesian models have been applied in ecology (Cohen, 1988; Haas, 1991, 1992; Crome et al., 1996) and environmental management (Oliver and Smith, 1990; Dixon and Ellison, 1996; Ellison, 1996; Wolfson et al., 1996).

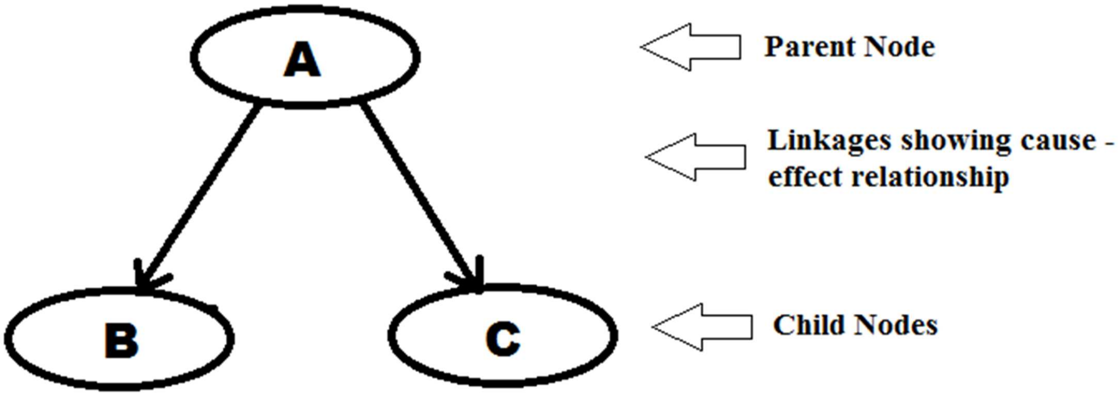


Fig. 4: Framework of a Bayesian Belief Network with two child and one parent nodes.

The main focus of this study is to analyze the factors affecting farmers irrigation decisions. Some studies have been reported in the literature that focus on farmer:

Becu et al. (2006) developed a multi-agent model to understand water sharing between two villages, one upstream and the other downstream. Farmer behavior in making decisions regarding planting crops, irrigation, harvesting, etc. was studied. Since different cropping patterns were identified in the region, agent farmers were divided into sub-classes. An irrigation decision was made on the basis of an irrigation schedule for each type of crop. The agent in this case had to decide the amount of water to be supplied to each plot, which was computed as the biophysical requirement for water. Bontemps and Couture (2002) studied farmers water consumption while being charged minimally for water use. The farmers did not bear the full cost of irrigation supplies. The study formulated a sequential decision model to analyze farmers irrigation behavior. Le Bars et al. (2005) developed a discrete event simulator called MANGA using a multi-agent systems paradigm. Two types of agents were considered: (a) cognitive, the human element, representing farmers and water supplier, and (b) reactive, which modeled crops, information suppliers, and climate. The objective was to simulate evolving farmer-agents over years, given a limited water resource. The model was useful for analyzing water use and its effects on yields at both individual and global levels. It could also be used to verify various scenarios in a given problem without having to contend with them in the field.

Overall, these studies built representative farmers and created scenarios of how farmers might act. They did not study how target farmer groups actually behave. They did not look at the variables that might be affecting farmer behavior. The work reported here is a first attempt at studying farmer irrigation decision behavior for which information is, or can be made available. The objective was to answer question: why farmers decide to irrigate on certain days as opposed to others. Is profit maximization one of the goals for irrigation? Which measured variables in the soil-plant-water system best account for the decisions that are actually observed? We study irrigation decisions by using plant, weather, and soil conditions, on and before the day of irrigation. Representative variables have been used to construct a modeling framework for the problem. Learning capabilities of BBNs have been exploited here. Since learning is data-intensive, we have used data from years 2007-2010. The model was tested with a subset of the data and used to make inferences about future irrigation decisions.

2.2 Learning Bayesian Belief Networks

The problem involves classifying decisions into two mutually exclusive classes on any given day during the growing season, i.e., a decision to irrigate or a decision not to irrigate. BBNs were selected to model farmers' irrigation decision behavior.

BBNs represent a system as connections between variables (nodes) and defines the relationships between variables with probabilities, denoting the magnitude of effect of one variable on the other (Jensen, 2001). This makes it very easy to visualize and interpret the relationships between variables. The network input parameters are prior probabilities, conditional probabilities and the posterior (outputs) probabilities. The likelihood of an input variable to be in a certain state is called the prior or unconditional probability. If a node has inputs from two or more other nodes, then the likelihood of the state of that variable depends on the state of the input nodes affecting it and is called conditional probability. Posterior probability is the probability that a variable is in a certain state resulting from the combined effects of the input variables, conditional probabilities and linkages.

The variables of a BBN are known as nodes. A BBN is based on Bayes' probability rule and updates existing beliefs with new evidence and finds marginal posterior probability for each node/variable. It can use a combination of the following at the same time: a) continuous and categorical variables, b) empirical and variables based on expert judgment, and c) deterministic or stochastic or the probabilities learned from data. BBNs can evaluate the outcome of an event by forward propagation and learning, and they can find the probabilities of factors contributing to an output of a natural system through backwards propagation.

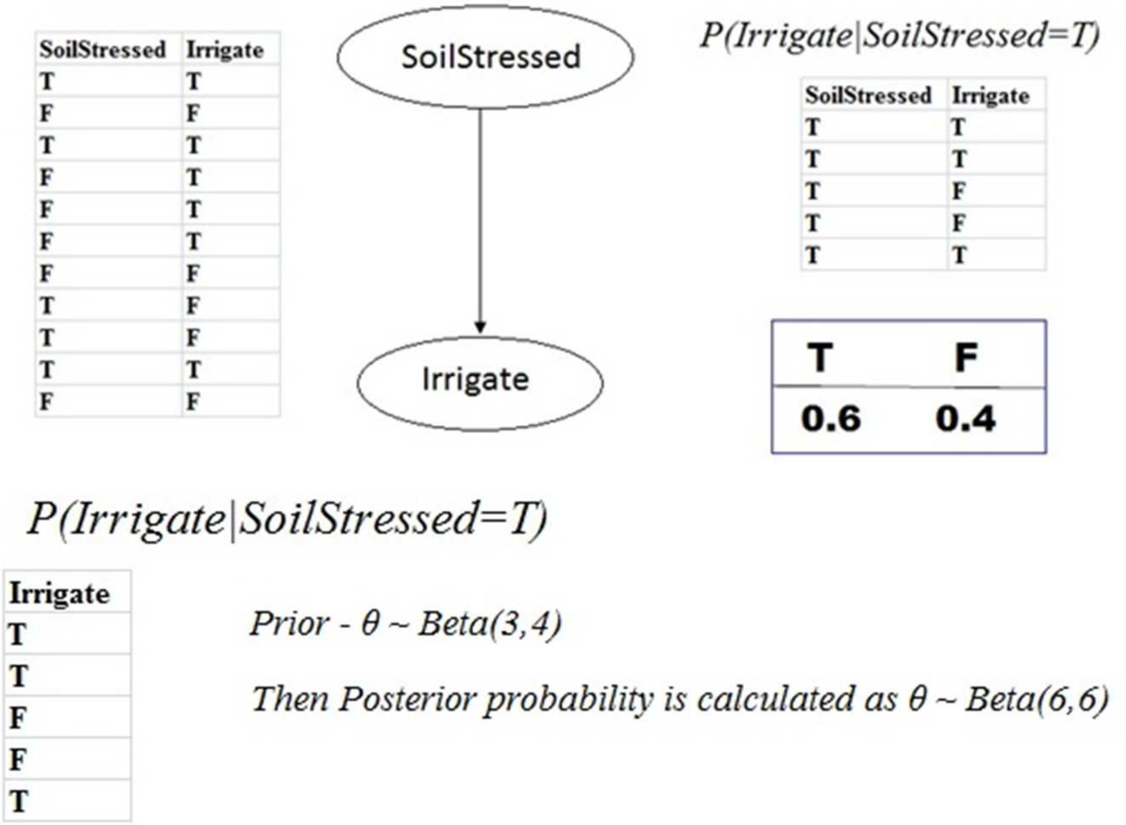
Learning in network models dates back to work done by Chow and Liu (1968). It is used when little is known about the marginal or conditional probabilities of certain nodes or when there is no expert opinion on them, for example, in our case the irrigation decision. By learning, either or both the marginal or conditional probabilities of the nodes can be estimated, given the structure of the network. Or, if we have the observed variables in the system, the network structure (commonly known as the Directed Acyclic Graph, or DAG) itself can be learned (Neapolitan, 2003). The network structure creation can result in different structures, depending on the data selected by the user. Let $P(h)$ be the probability of the hypothesis, h, $P(D)$ the probability of training data, and $P(D|h)$ the probability of D given h. Then Bayes theorem for learning (Mitchell, 1997) and the probability of the hypothesis, h given the training data, D, would be as follows,

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Suppose X is a set of random variables, $X_1, X_2, ..., X_n$ and the training data is a set of $D_1, D_2, ..., D_n$ and we have the model of the distribution of the variables in X with parameters (unknowns), Θ. Then the learning problem is to find the parameters such that the observed data is best explained. To elaborate this, let's consider an example related to this study. Figure 5 shows node "SoilStressed" which denotes the soil moisture status and whether it is stressed. Node "Irrigate" is representative of the farmer's action. The network tries to answer the question: If the soil is stressed, then does the farmer irrigate? If there is no prior knowledge about the process or variable, the network starts from equal probabilities. For example, we want to calculate the probability of "Irrigate" given that the soil is stressed, $P(Irrigate|SoilStressed = T)$, from this network, given observed data. The procedure would be as follows:

1. First pick the cases of random variable "SoilStressed" which are true (T) (refer to Figure 5 and define the distribution of the node "Irrigate" and forget the rest. Two out of five cases represent "No irrigation" (F) condition. Hence the learned maximum likelihood estimate will be represented as a 60.0% probability (let's call it 'p') (or degree of belief) for the next time to irrigate and a 40.0% probability otherwise. The order of the cases does not matter.

2. To find the Bayesian estimate, a beta distribution is assumed (for two likelihood



SoilStressed	Irrigate
T	T
F	F
T	T
F	T
F	T
F	T
F	F
T	F
T	F
T	T
F	F

P(Irrigate|SoilStressed=T)

SoilStressed	Irrigate
T	T
T	T
T	F
T	F
T	T

T	F
0.6	0.4

P(Irrigate|SoilStressed=T)

Irrigate
T
T
F
F
T

Prior - $\theta \sim Beta(3,4)$

Then Posterior probability is calculated as $\theta \sim Beta(6,6)$

Fig. 5: Causal link between nodes '$SoilStressed'$' and '$Irrigate'$', the estimation of the maximum likelihood estimate for the conditional probability of '$Irrigate|SoilStressed = T'$', and lastly, the Bayesian estimate of the posterior probability.

states) for the prior distribution over the parameter, p, which is denoted as $Beta(\alpha, \beta)$ where $\alpha - 1$ is the number of successes (with probability p of success) and $\beta - 1$, the number of failures.

Let's assume it as $Beta(3, 4)$. If it was $Beta(6, 4)$, it would mean placing equal confidence in the initial probability estimates; $Beta(3, 2)$ would mean less confidence; and $Beta(12, 16)$ would mean more confidence. Having high values of α relative to β causes the mode of the distribution to be skewed to the left, while high values of both cause the distribution to be high peaked (because of low value of variance). By adding more data the distribution turns more and more peaked implying higher confidence in our expectation.

3. For this case, the posterior distribution will be given as $Beta(Y+\alpha, n-Y+\beta)$, where n is the total number of trials in an experiment, Y is the number of successes in the experiment, and α and β are the shape parameters assumed for the prior distribution. For this example, n=5, Y=3, so the posterior beta distribution is $Beta(3+3, 5-3+4)$ resulting in $\theta \sim Beta(6,6)$. The posterior distribution is, basically, the prior distribution updated by the new data. It is quite obvious that a different prior would lead to a different posterior distribution (refer to Figure 6). This can be done with any number of variables involved in learning process.

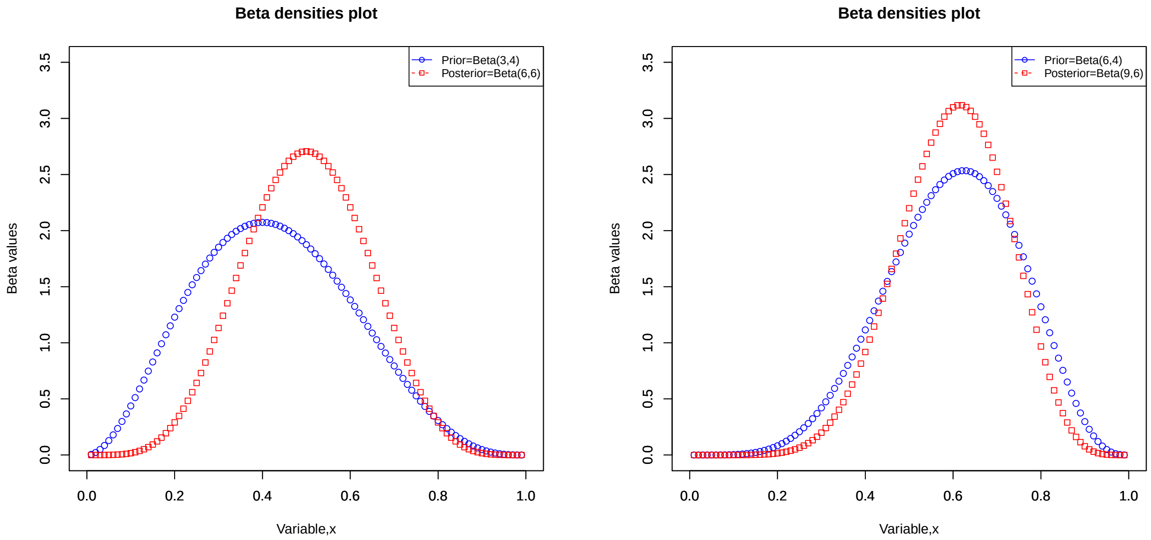


Fig. 6: Beta densities plot for (a) Left - Prior = $Beta(3,4)$ and the resulting Posterior = $Beta(6,6)$, (b) Right - Prior = $Beta(6,4)$ and the resulting Posterior = $Beta(9,6)$. This explains the importance of a good prior and the confidence in initial probability estimates.

No water management literature was found which reported studies of learning capabilities of BBNs. Belief network modeling for this work was done using Netica-J, the Java version of the Netica API (Norsys, http://www.norsys.com/) for batch operations and ease of learning and testing from case files. Netica assumes independent conditional probabilities and the Dirichlet function (continuous probabilities with 0 and 1 limits) for prior probabilities (Spiegelhalter et al., 1993; Castillo et al., 1997). For learning, Netica has provisions to use counting, gradient descent, and expectation maximization algorithms. For this problem, it all the three gave similar results but the last two algorithms took more time to solve the network. Hence, simple counting was used to learn the parameters of the networks.

The BBN developed in this study takes into account those factors which, theoretically, can affect the farmer's irrigation decisions. In spite of including many factors, we may be missing some of the critical ones due to the lack of available data.

2.3 Model Development

2.3.1 Variable Selection

The variables were selected for the BBN to represent the information pertinent to on-farm irrigation decisions. The structure of the model was based on the classical soil moisture balance model and allied literature in irrigation scheduling. To discretize the continuous time series data, reasonable limits for weather variables were used. The model calibration eventually fixed the number of states for various variables. The model components are shown in Figure 7.

Since this model was built to identify the likely factors leading to farmers decisions to irrigate, variables were selected such that they could be measured or, with justification, assumed for such things as real time soil moisture content, weather data available from a local station to which farmers have access, market prices, crop and

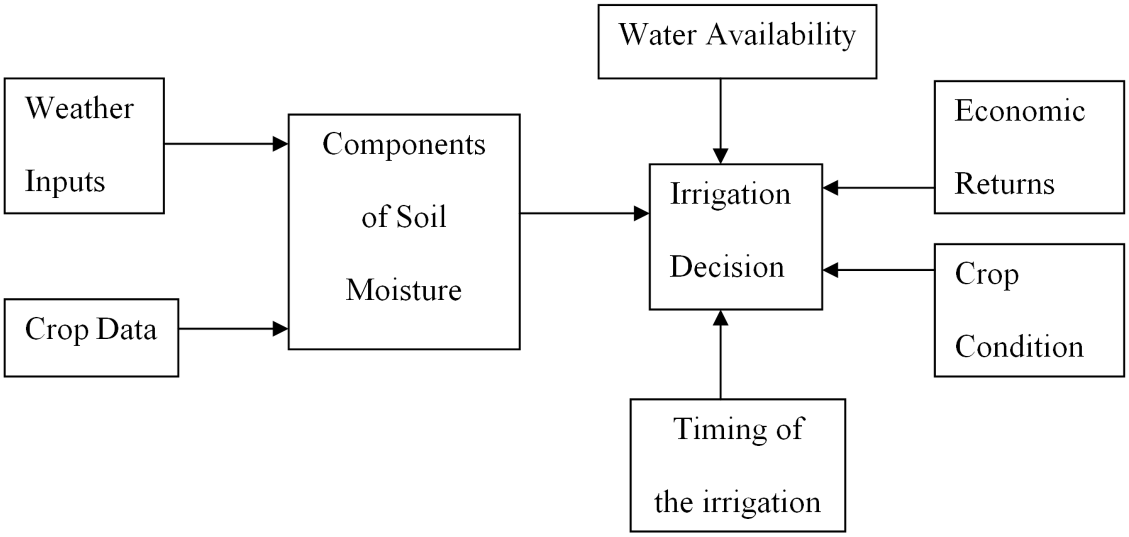


Fig. 7: Components represented by nodes of the Bayesian belief network.

soil condition indicators, etc.

Mathematically, the soil moisture depletion at the end of the day [mm], in root zone depth, r [mm], is given as:

$$D_{r,i} = D_{r,i-1} - P_i - I_i + ET_{a,i} + DP_i \tag{1}$$

where,

$D_{r,i-1}$ is moisture depletion (D) by the end of the previous day [mm],

P_i is the amount of rainfall on day i [mm],

I_i is the depth of irrigation on day i [mm],

$ET_{a,i}$ is the actual crop evapotranspiration on day i [mm],

DP_i is the deep percolation on day i [mm].

Weather inputs used were daily minimum, maximum, and average air temperature, average relative humidity, and average wind speed. Deep percolation (mm) water was estimated by calculating a constant rate of 'loss' of water from the soil from just after irrigation up to three days after irrigation (the approximate time it takes to reach field capacity) and multiplying with total available water(TAW).

$$PercolationAmount = TAW * PercolationRate$$

Irrigation amounts (mm) were calculated as the product of the difference between porosity and the soil moisture content on the day before irrigation, and the application depth (mm).

$$IrrigationAmt = (Porosity - SMC_{i-1}) * ApplicationDepth$$

where, IrrigationAmt is the irrigation amount (mm), and SMC_{i-1} is the Soil moisture content before the day of irrigation.

2.3.2 Nodes and Links of Bayesian Network

With respect to the environment being modeled, the network was divided into various levels such as weather variables, domains affected by weather (e.g., soils, crops), independent factors such as canal flows, and farmer decision to irrigate. If the farmer irrigated, then it meant that there was water available to him. FAO-56 (Allen et al., 1998) was used to define the causal relationships between the variables.

The initial model shown in Figure 8, had 31 nodes and 36 links, the parents (immediate) of the child node 'Irrigate' decision have two states. Other variables had three or more states to consider every possible condition. To simplify the architecture, the network description starts from the child node, 'Irrigate' which was a farmer's decision to irrigate. The Node 'Irrigate' had two states, 'Yes' and 'No'. The contributing factors to this decision were the following irrigation needs from various components of the system:

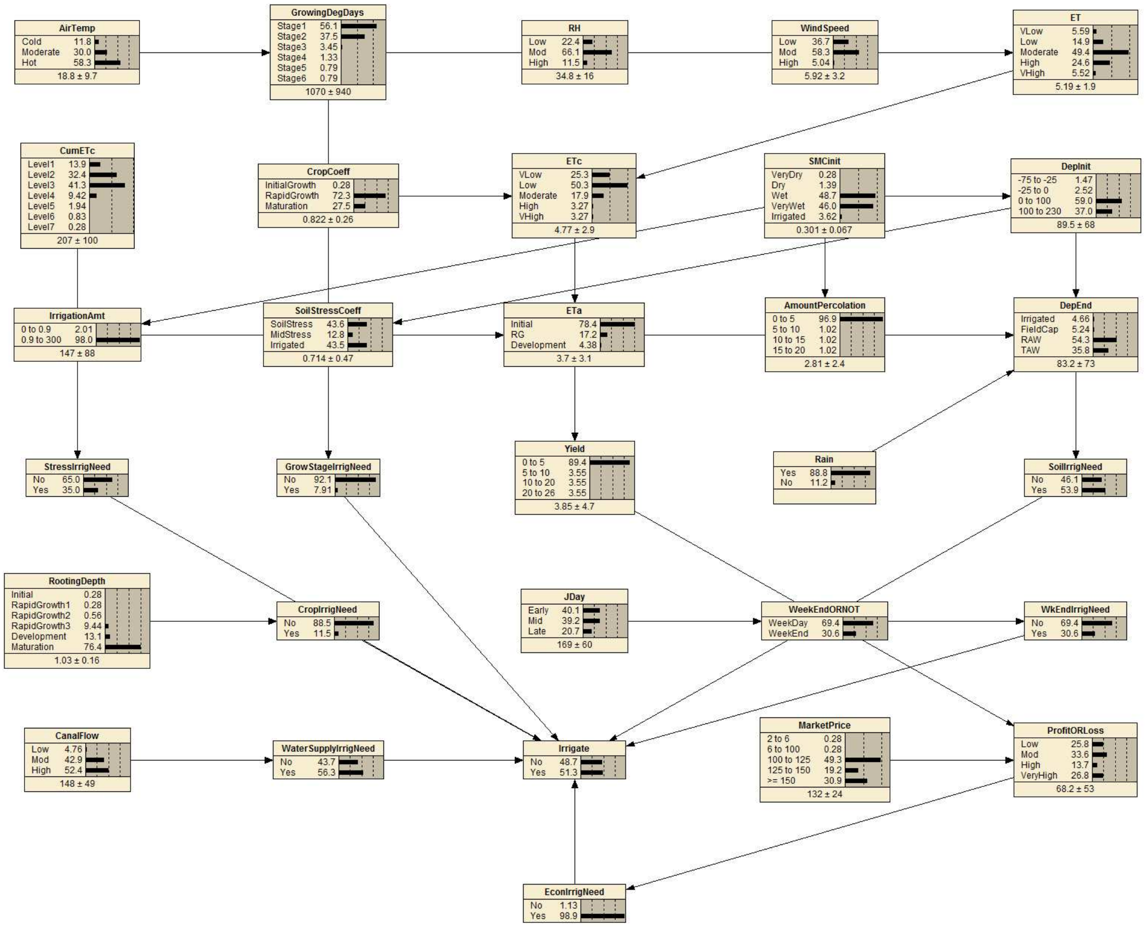


Fig. 8: A snapshot of the built network relationships after learning. The network starts with equal probabilities of the states for all the variables.

1. Node ‘SoilIrrigNeed’ - Soil condition is one of the most important criteria for an irrigation decision. Farmers are very familiar with the texture and feel of dry and wet soils. The soil condition is also reflected in the crop condition. Farmers sometimes irrigate when they see some plants with yellow leaves and presume it is time to irrigate. However, the irrigation principles state that this could be because of water logging. This factor helped to determine whether the soil need was the primary cause of irrigation in every instance the farmer thought of irrigation. If it is probably the main cause, then it would practically end the search for other significant, causal factors. The logic in the node is described

below. The classical Penman-Monteith equation uses relative humidity (Node RH), windspeed (Node WindSpeed) and air temperature (Node AirTemp) with some other variables to calculate Evapotranspiration (Node ET). The other variables used in the calculation have not been used here since they have not been found to contribute to the irrigation decision directly. Crop ET (Node ET$_c$) is obtained by multiplying ET and the crop coefficient, K$_c$ (Node CropCoeff), followed by actual ET (Node ET$_a$) which is a product of ETc and the soil stress coefficient, Ks (Node SoilStressCoeff) given as:

$ET_c = ET_o * K_c$

$ET_a = K_s * ET_c$

Total plant available water (TAW) is defined as the portion of water in the root zone (RD) which can be extracted by the plant. Field capacity (FC) is the upper limit of water held in the soil when the gravitational water has been drained from the soil profile. Wilting point (WP) is the lowest limit of available water which the plant can use.

$TAW = (FC - WP) * RD$

Readily available water is the amount of soil water the plant can extract from the soil profile without suffering any stress:

$RAW = MAD * TAW$

where, MAD is the management allowable depletion and may be different from farmer to farmer and might also be based on the crop. TAW and RAW are hypothetical limits for daily soil moisture depletion. The soil stress coefficient, 'K$_s$' (Node SoilStressCoeff) is 1 until RAW is greater than depletion. As soon as depletion crosses the RAW limit, stress sets in and Ks is computed by the following equation:

$$K_s = \frac{TAW - Depletion}{TAW - RAW}$$

The deep percolation amount (Node AmountPercolation) was estimated by calculating a constant ‘rate’ of loss of water from the soil after irrigation, up to three days after irrigation (the approximate time it takes to reach the field capacity) and multiplying with total available water. TAW is used in this calculation since it is the amount of water held in the soil column.

$PercolationAmount = TAW * PercolationRate$

Irrigation amount (Node IrrigationAmt) was calculated by taking the product of the difference between the soil moisture content on the day before irrigation and porosity, and the application depth (mm).

$IrrigationAmount = (Porosity - SMC_{i-1}) * ApplicationDepth$

where, IrrigationAmt is the irrigation amount, and SMC$_{i-1}$ is the initial soil moisture content (Node SMCinit). Actual rain amounts were used at Node Rain, with states ‘Yes’ or ‘No’. The initial depletion (Node DepInit), D$_{i-1}$ was zero making the field capacity for every soil type, the initial soil moisture content. The depletion at the end of the day (Node DepEnd) is given by the soil moisture balance as follows:

$$D_i = D_{i-1} - P_i - I_i + ET_{a,i} + DP_i$$

2. Node ‘StressIrrigNeed’: It was found during initial data analysis that some of the farmers irrigated according to consumptive use of the crop. This node resembles the stress imposed on the crop due to accumulated evapotranspiration (CumET$_c$). The cumulative ET was reset on the day of subsequent irrigation.

3. Node ‘WkEndIrrigNeed’ - This node was based on the observation that farmers may prefer to irrigate on a weekend because some might have an active job during the weekdays and restrict some farming activities to the weekend. A node for the Julian Day (Node JDay) and another for determining if it is a weekend (Node WeekEndORNOT) were the parents to this node.

4. Node ‘WaterSupplyIrrigNeed’- Some farmers might tend to irrigate when a neighbor irrigates. This node mimes that action of a farmer. It translates into whether the farmer chose to irrigate with the others (on a day of high flow) or took an independent decision (low flow) for irrigation. Canal flow (Node CanalFlow) data was fed into this node.

5. Node ‘GrowStageIrrigNeed’ - Accumulated degree days (Node GrowingDegDays) have been a valuable tool to represent the vulnerability of crop stage to pests. It can also provide an information surrogate for the growth stage reached. This factor is different for different crops. The air temperature (AirTemp) was summed up over the complete growing season. The base temperature was taken as zero (0 degC) for all the crops.

6. Node ‘EconIrrigNeed’ - Crop ET (Node ET$_c$) and Actual ET (Node ET$_a$) feed into the Node Yield according to FAO-33 (Doorenbos and Kassam, 1979). Daily values of expected market price were the inputs to the Node MarketPrice data. K$_y$ is the yield response factor. The product of market price and yield resulted in the values for Node Revenue. The actual yield as weighted by maximum expected yield values were calculated using the following equation:

$$Y_a = Y_{max} * K_y * (1 - \frac{ET_a}{ET_c})$$

This means that the farmer might be irrigating for higher revenues on a certain day because he is losing the quality of his crop.

7. Node 'CropIrrigNeed' - Though the growing stages are reflected through the growing degree days, the Node Rooting Depth accounted for the increasing root depth of the plants, which was assumed to increase with time. This can be important since newly planted crops like alfalfa stop root growth after the development stage and before first cutting, and have already stopped rooting further if they have been developing from previous years.

2.3.3 States of Variables

The number of states of the variables in the network are presented in Table 1. All the nodes feeding into the 'Irrigate' node (CropIrrigNeed, EconIrrigNeed, GrowStageIrrigNeed, SoilIrrigNeed, StressIrrigNeed, WaterSupplyIrrigNeed, and WkEndIrrigNeed) and IrrigationAmt and Rain nodes had two states: 'Yes' and 'No', indicating presence or absence of the factor. Node WeekEndORNOT separates weekdays from weekends. The nodes which reflected time in the growing season, had three states, viz. JDay and CropCoeff, reflecting early, late, and middle season, whereas node SoilStressCoeff, also had three states denoting the wetting and drying phases between two irrigations- Irrigated, Mid-stress (half-way through stress), and Stressed. Nodes representing weather or flow variables, such as AirTemp, CanalFlow, ET$_a$, RH, and WindSpeed, had three states denoting high, medium, and low levels. Nodes such as AmountPercolation, DepEnd, DepInit, Revenue, and Yield, had four states which accommodated different water holding capacities of soil types, or different crop yields according to the area irrigated by the farmer (some farmers had larger fields in comparison to others). ET and ET$_c$, both had five levels in order to have smaller bins to account for day-to-day variations. Similar to the depletion variables, DepEnd and

DepInit, SMCinit had five states to account for different starting values of soil moisture content to account for all soil type and crop type combinations. Finally, nodes GrowingDegDays and RootingDepth had six, and CumET$_c$ had seven states, to have finer discretization during the growing season.

Table 1: Number of states selected for various variables.

Node Name (1)	**Number of states (2)**	**(1)**	**(2)**
AirTemp	3	JDay	3
AmountPercolation	4	MarketPrice	5
CanalFlow	3	Revenue	4
CropCoeff	3	Rain	2
CropIrrigNeed	2	RH	3
CumETc	7	RootingDepth	6
DepEnd	4	SMCinit	5
DepInit	4	SoilIrrigNeed	2
EconIrrigNeed	2	SoilStressCoeff	3
ET	5	StressIrrigNeed	2
ETa	3	WaterSupplyIrrigNeed	2
ETc	5	WeekEndORNOT	2
GrowingDegDays	6	WindSpeed	3
GrowStageIrrigNeed	2	WkEndIrrigNeed	2
IrrigationAmt	2	Yield	4

2.4 Case Study

2.4.1 Canal B, Lower Sevier River Basin, Utah

The study site selected covers 20 square miles near Delta, Utah in the Lower Sevier River Basin. Snowmelt is the major contributor to soil moisture in the early part of the growing season which is usually late spring. Irrigation is the biggest user of the water in this basin. Surface irrigation is the dominant method in the region. Telephonic anecdotal accounts given by the water masters of the canal company who are farmers in the area, were compiled. They explained various reasons for their irrigation decisions, including observing a neighbor irrigating, the plant-soil condition, the amount of water remaining in their water right for the season, and the type of crop and the stage. They told us that the farmers order water but do not necessarily use it to irrigate as soon as they get water. They store it in the ditches itself and use them when needed or also might rent it out. We do not have any means to ascertain these claims, but these facts helped us in modeling the problem better.

2.4.2 Data

Weather data for the study area were obtained from the URL:`http://www.cemp.dri.edu/cgi-bin/cemp_stations.pl?stn=delu`. The station was established by National Climatic Data Center (NCDC), NOAA and has historical weather data since 1965. The station can be located on the NCDC-NOAA website (`http://www.ncdc.noaa.gov/`) using the following metadata:

GHCND ID : USW00023162

COOP ID : 422090

WMO ID : 72479

NCDC ID : 20026236

Precipitation data were not found to be representative of the conditions at the site because localized showers are observed in the area during the irrigation season. Data calculated by Kimberly Penman Reference ET procedures are available on the foregoing website. The calculations were verified for accuracy.

In 2007, the Utah Water Research Laboratory (UWRL), Utah State University (USU) established 44 stations with 88 sensors to record soil moisture content at 1 and 2 ft depths on various farms in Delta, Utah to study agricultural water use. The sensors are maintained by the personnel at UWRL. Soil moisture content measured at these stations was used to determine the day of irrigation and the approximate amount of irrigation. These were obtained from: http://odm.usu.edu/odmmap/default.aspx?NetworkName=Delta. Hourly measurements are available on the website, so daily average values were estimated for the first day of the season, for starting the soil moisture balance calculations. SCADA (Supervisory Control and Data Acquisition) systems maintain the past records of water discharge across Sevier River Basin (Berger et al., 2002) and real-time conditions are accessible on http://www.sevierriver.org/. The site is sponsored by Sevier River Water Users Association (WUA). Daily canal flows in cubic feet per second (cfs) in Canal B were obtained from the URL: http://www.sevierriver.org/rivers/delta/b-canal/.

A soil moisture balance was determined to compute the daily moisture depletion. No specific data for the day of planting for years 2007-2010 were available in the literature, so it was difficult to begin the growing period with a specific planting date. Wright (1982) describes the day of planting for crops grown at Kimberly, Idaho. These dates gave a very general idea of the day of planting, but since this was a field-by-field study, crop and farmer-specific dates were needed to address the difference in the observed irrigation dates. Hence, soil water balance calculations were made starting with planting dates (randomly, some days before the first irrigation), such that the initial depletion was 0 (i.e., starting from field capacity) and the model's day of irrigation matched with the day of first recorded irrigation. This procedure addressed the lack of knowledge of initial conditions and resulted in such assumptions as no stress during the period of crop establishment and soil moisture not being inhibiting initial plant growth. This also accounted for the fact that after snow melt at the site, the soil would not be completely dry. It was assumed that the crops are planted or emerge (in case of alfalfa) as soon as suitable temperatures are reached. The greatest challenge with using the water balance calculations to describe soil moisture through time is that the model indicates irrigation need in the beginning of the season too frequently, which does not agree with how the farmers irrigate. The reason for this is the shallow root depth for crops just planted. The crop demand is also not very high as represented by low crop ET in this part of the season owing to low temperatures. It can be deduced for practical purposes that the farmer does not apply water according to the root growth but considers an "application depth" for early season irrigation which is uniform for all crops. Hence a constant application depth of 1m was assumed. Also, the deeper in the soil column, the less significant are effects of factors such as ET. This method also worked for annual crops such as barley and corn, since at field capacity the farm implements can enter the field easily for ploughing and seeding.

A soils map of the study area indicated three soils types: silty clay loam, silty clay, and loam. The major soil characteristics governing water movement and retention are porosity, field capacity, and wilting point. Usually, porosity defines the saturation limit of a specific type of soil, while field capacity and wilting point put limits on the available plant water, which is important in this case since we are considering crop growth as well as soil water extraction. Standard values for these parameters were considered from Allen et al. (1998).

Representative crop phenology coefficients for alfalfa, barley, and corn were obtained from Wright (1982) and FAO-56 (Allen et al., 1998). The other consideration in this calculation was that we were using it for crop reference ET, hence we had to multiply the values by a factor of 1.2 to consider field crops as opposed to grass reference ET. Literature values for yield response factor, K_y were used (Allen et al., 1998). Since there was no evidence of capillary rise and runoff, they were considered negligible. A daily linear time series was constructed by interpolating the monthly values of market price data available on the USDA website (`http://www.nass.usda.gov/Statistics_by_State/Utah/Publications/Annual_Statistical_Bulletin`).

2.5 Results

2.5.1 Calibration and Testing

Calibration is the process of tuning the model such that its behavior is close to the system being modeled. Instead of using parameters to calibrate the process, the model will learn the process from the data. The model was trained using the data for the days the decision was taken and the one before it, since the number of irrigations were infrequent in any given season. The conditional probability tables (CPT) were populated by learning. The probability distributions for all the nodes are to be found, including the Irrigate decision node. The intermediate nodes were used to reduce the number of variables going into the decision node. A sample of the representative data was run through the network to define the states, to account for all the possible scenarios.

Since the order of the data does not matter, bootstrapping would not result in better results. Hence all the data from different years were mixed and matched for training and testing of the networks. The results of the analysis are presented in Table 2. The networks trained with 2009 data for alfalfa and tested with 2008 data gave the lowest testing accuracy of 81.0%. Also a combination of years 2008-2010 for training and 2007 for testing for alfalfa resulted in testing error of 81.0%. A confusion matrix presents the number of correct and incorrect classifications produced by a classification model, and is a standard output for classification problems. The confusion matrix for the irrigation decision model is shown in Table 3. The cases correctly classified by the model appear on the diagonal of the matrix. The confusion matrix shows that only two irrigation events were missed in the testing phase. The error rate is high because some irrigations were predicted by the model when the farmer did not irrigate. The lack of data for barley and corn crops resulted in low accuracies. Another strange observation was that whenever 2009 data for alfalfa was used for testing, it resulted in lower testing error as compared to training error.

Table 2: Results of calibration and testing of Bayesian belief networks.

Year/s for Training	**Crop**	**Training Accuracy, %**	**Year/s for Testing**	**Testing Accuracy, %**
2009	A	91	2008	81
2008+2009+2010	A	85	2007	81
2009	B	71	2010	61
2007+2010	B	68	2009	56
2008+2009	C	63	2007+2010	61
2009	C	79	2007	59
2007+2008+2009	C	64	2010	59

Table 3: Confusion matrix for Irrigate showing number of events predicted correctly by the model.

No	**Yes**	**Actual**
126	51	**No**
2	175	**Yes**

For our problem, the learning results gave an insight into the irrigation decision-making process and the factors that likely affect it. Table 4 shows the possible reasons for decisions to irrigate across various crop-year combinations. Soil stress was the leading probable rule for irrigation for most years and crops, as can be seen in Table 4. Due to a deep rooting crop, the need to irrigate the crop frequently was eliminated. Deeper roots can utilize subsoil moisture. Hence the farmers might have considered this factor to irrigate barley crops in 2008-2009. For alfalfa in 2008 and 2010, barley in 2007 and 2009, and corn in 2010, farmers might have irrigated similarly to their neighbors. Irrigating on the weekend was found to be one of the dominating reasons for alfalfa in 2010, barley in 2007, 2009, and 2010, and corn in 2008.

Table 4: Factors resulting in highest beliefs to irrigate for different years and crops. (1) denotes 'Beliefs to irrigate' - implying majority of farmers used the rule.

Year /Crop	**(1)**	**Alfalfa**	**(1)**	**Barley**	**(1)**	**Corn**
2007	90	Soil stressed	75	Deep Rooting, Neighbor irrigating, Soil stressed, WeekEnd	73.3	Soil stressed
2008	92.3	Soil stressed, Neighbor irrigating	NA	NA	80	WeekEnd
2009	94.4	Soil stressed	66.6	Deep Rooting, Soil stressed, Neighbor irrigating, WeekEnd	83.3	Soil stressed
2010	90	Neighbor irrigating, Soil stressed, WeekEnd	75	WeekEnd	75	Neighbor irrigating, Soil stressed

2.5.2 Inference

Like all other expert systems, it is essential to see how the BBN probabilistic model performs in drawing conclusions when provided with new inputs or evidence. This is a method of updating the probabilities according to the new observations, known as belief updating. Use of growing season data from a farmer's field can eventually point to the possible reasons of his decision to irrigate, as well as the probability or the belief to irrigate on a given day.

Figure 9 shows the inference results for an unseen data set (for clarity, only a portion of the inference has been shown here). The network does a reasonable job in predicting the probability of irrigation. All the probabilities are greater than 50.0% (at least 3 instances) and rise to 90.0% on three different days. For at least five days, the probabilities to irrigate are around 75.0%. The network does a fairly good job in classifying 7 out of 12 days with “No irrigation”, by assigning them a probability of around 5%, indicating a reasonably low value for days when the farmer did not decide to irrigate. The network predicted some non-irrigation days as irrigation decisions of the farmer, one of which was with 70.0% belief and four of with around 55.0% belief.

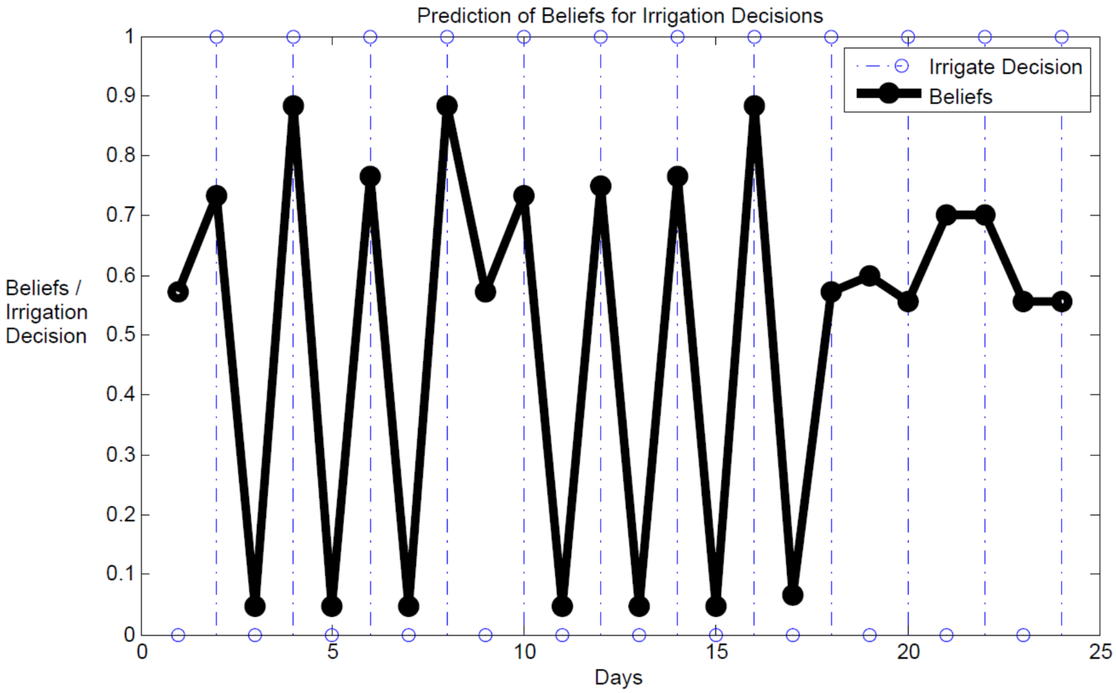


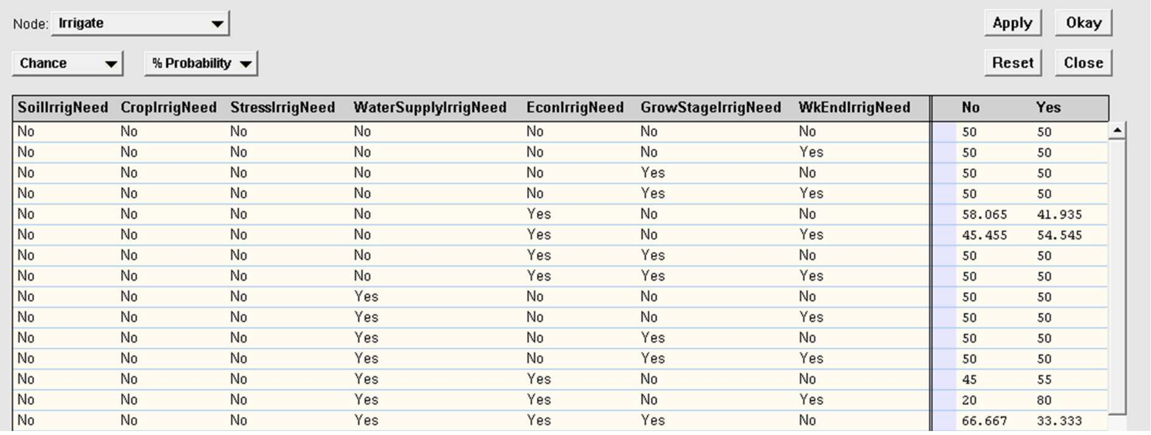
Fig. 9: Beliefs predicted by BBNs for an unseen data set.

2.5.3 Model Sensitivity

For this case if any variable was removed from the network, the model failed to perform well. Hence, all the variables were included during the various tests. None of the nodes were assigned initial probabilities based on expert judgment, hence model sensitivity testing was not needed. The network started out with equal probabilities of the states of the variables. There were some variables which had no variance contributing to the rule, but if they were removed, the network predictions worsened.

2.6 Discussion

The results obtained from the Bayesian belief network for studying and forecasting irrigation behavior provided insights into the irrigation decision process, though the reasons for irrigations of barley and corn were not well captured by the network. To completely understand the process it is important to look at the conditional probability table (CPT) (Figure 10).



SoilIrrigNeed	CropIrrigNeed	StressIrrigNeed	WaterSupplyIrrigNeed	EconIrrigNeed	GrowStageIrrigNeed	WkEndIrrigNeed	No	Yes
No	No	No	No	No	No	No	50	50
No	No	No	No	No	No	Yes	50	50
No	No	No	No	No	Yes	No	50	50
No	No	No	No	No	Yes	Yes	50	50
No	No	No	No	Yes	No	No	58.065	41.935
No	No	No	No	Yes	No	Yes	45.455	54.545
No	No	No	No	Yes	Yes	No	50	50
No	No	No	No	Yes	Yes	Yes	50	50
No	No	No	Yes	No	No	No	50	50
No	No	No	Yes	No	No	Yes	50	50
No	No	No	Yes	No	Yes	No	50	50
No	No	No	Yes	No	Yes	Yes	50	50
No	No	No	Yes	Yes	No	No	45	55
No	No	No	Yes	Yes	No	Yes	20	80
No	No	No	Yes	Yes	Yes	No	66.667	33.333

Fig. 10: A portion of the CPT of "Irrigate" node showing the learnt probabilities to irrigate or not.

As explained before, the CPT is comprised of seven parent nodes. The factor combination learned from the data results in the calculation of the posteriors. If no such combination is found in the data, the probabilities remain unchanged. Again going back to Figure 10, the factor combinations which result in 'Irrigate' probabilities of 41.9% (EconIrrigNeed =Yes) and 54.5% (EconIrrigNeed and WkEndIrrigNeed =Yes) can be compared. The probabilities resulted because many farmers were irrigating on a weekend, so the only possible difference in the two is that the latter has the weekend factor, too. During training, the data did not reflect as many irrigations under those conditions (EconIrrigNeed =Yes) which led to incorrect classification during testing. A counting algorithm goes through the data and puts similar groups together, but maybe there were not a sufficient number of patterns to corroborate certain decisions. The error rate could be high also because the network got more of those factor combinations in the testing phase, while there was no observed irrigations due to those factors during training phase. Also, we did not have as much data for barley and corn as for alfalfa. The other explanation could be that the factor combination was indicating irrigation (since the model had recorded such instances during training) but since the call time in Canal B ranges from 24 hours to 3 days according to operating rules followed by the canal company, the farmer might have taken a decision to irrigate but did not get water in time. This can only be verified if we had access to water order data from the canal company.

During inference, it was noticed that the networks were not able to infer correctly about the irrigation events. On an average a farmer made 3 to 4 (6 at maximum) decisions to irrigate. Inference results were plotted for 12 such events. Evidently, early in the season the network predicted some non-irrigation days as irrigation decisions of the farmer. This could have happened because we did not have the correct information for the date of planting. And as explained earlier, the soil was assumed to be at field capacity on the day of planting. Hence the reasons for the first irrigation might not have been properly recognized in the factors given to the model. The first irrigation is usually given to facilitate germination of seeds or emergence of previous crops like alfalfa.

EconIrrigNeed and StressIrrigNeed were always ‘Yes’ for the irrigation events. As individual variables these might be insignificant, but in combination with other factors, they could have been contributing to the process. If they were removed, the model performance was poor. Since daily values were not readily available, daily time series was constructed for market prices by linear interpolation from the monthly values. The anticipated sales prices increased throughout the entire growing season.

When we linearly interpolated, the prices constructed were rose daily. Economic need was always important in the model because the market prices were always rising throughout the season.

Due to laser leveling the timing between two irrigations was sometimes more than 30 days. Hence, not all of the data from the growing season could be used because it would result in an imbalanced data set with fewer irrigation events in comparison to the number of no-irrigation days.

2.7 Conclusions

Water managers, decision makers and canal operators are always challenged by lack of knowledge about the irrigation water demand that will develop over the short term. This can be partly solved by accounting for farmers irrigation decisions. The decisions and the subsequent water orders can be eventually summed at the command area level to get short-term estimates of water demand.

2.7.1 Bayesian Belief Network as a Tool

BBNs provide a tool to analyze farmer irrigation decision behavior and predict his probable future decisions. They are easy to build, and provide various ways to interpret the results. The only requirement for them is that there should be some information about the relationships between variables. They can also learn the relationships from case scenario data and then simulate future events based on the results of learning. But as with any learning algorithms, these networks have to be trained for each new geographic location. Clearly, irrigation decision making is a multivariate process. The more variables we have, the better model performance we can expect. These models are data-intensive and require a large number of events to improve the prediction accuracy. An important limitation of such models is that they can perform better for immediate decision-making (1-3 days before the decision may be made), but their usefulness for long-term forecasting may be limited. This will depend on the duration between the irrigations. Delays caused in irrigation due to harvesting of alfalfa, for instance, can be useful, if they can inform the model how long it took for the post-harvest process, and incorporate the valid reasons of delay in irrigation, apart from stressing the crop.

2.7.2 Rules Used to Make Irrigation Decisions

Soil stress was found to be one of the most important factors that is apparently used by farmers in Delta to guide irrigation decisions. From the perspective of irrigation principles, we know that soil condition is an important indicator for irrigation. The water in the soil profile is lost by the process of evapotranspiration, which is the immediate reason to irrigate. Soil moisture balance calculates soil moisture for a specific rooting depth. Though we still need supporting evidence, it is unlikely that farmers would track root depth during the growing season. Hence, it is a strange factor to contribute to the decision. Weekend irrigations and irrigating when neighbor irrigate have been some traditionally used triggers for irrigation decisions, and this study found data supporting this. Farmers usually observed neighbors for their irrigation decisions. Most of the farmers in the Delta area have a full-time occupation during the week, which increases the likelihood that they will engage in agricultural activities such as irrigation on the weekend. By using these rules, the prediction accuracy of irrigation decisions was 81.0% for alfalfa and 61.0% for barley and corn.

2.7.3 Behavior

Irrigation decision-making is a complex process involving interaction between a combination of indicators. This study simulates conditions which might have been similar to what the farmers saw when they made a decision to irrigate. Hence soil stress, rooting depth (crop needs), an active profession during the week, and a talk with his neighbor might be some of the possible reasons for irrigation decision behavior in Delta. It is also evident that farmers look at different factors for every irrigation. They clearly look for not one or two but many indicators for irrigation. This work shows that biotic, climatic, and edaphic conditions suffice the requirements of indicators to study and forecast irrigation decisions.

2.8 Recommendations

For future work, it is important to have a reasonable amount of data representing all possible conditions under which the farmers irrigate their crops. It would eventually determine the performance of the network in the testing phase. Also it is necessary to have day of planting, initial conditions, and any other information specific to the early stages of crop growth. Application efficiency and system efficiency are some other variables which could be successfully incorporated in the BBN.

It is difficult but there should be a variable defining the level of satisfaction of the farmer and ways to represent behavior, and its relation to the utility of water for the crop on a certain day. The next step in this process could be managing a farm with several crops. Learning can facilitate information on why the farmer irrigated a certain crop first amongst all of the crops he planted that year. We know now that there can be different relevant factors important for one irrigation but might not be important to the other irrigation. Hence we should model the process, irrigation by irrigation, and give different sets of variables for different stages of crop growth or timings and let the model decide which one is crucial at that given point in the season.

References

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration - guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper no. 56.

Becu, N., Sangkapitux, C., Neef, A., Kitchaicharoen, J., 2006. Participatory simulation sessions to support collective decision: the case of water allocation between a Thai and a Hmong village in northern Thailand. In: March 7-9, 2006, Chiang Mai, Thailand, International Symposium, Towards Sustainable Livelihoods and Ecosystems in Mountainous Regions.

Berger, B., Hansen, R., Hilton, A., 2002. Using the World-Wide-Web as a Support System to Enhance Water Management. In: Workshop Proceedings - Irrigation Advisory Services and Participatory Extension in Irrigation Management - Workshop organised by FAO - ICID, July 2002, Montreal, Canada.

Bontemps, C., Couture, S., 2002. Irrigation water demand for the decision maker. Environ. Dev. Econ. 7 (4), 643–657.

Castillo, E., Gutiérrez, J.M., Hadi, A.S., 1997. Expert Systems and Probabilistic network models. Springer, New York, USA.

Chow, C.K., Liu, C.N., 1968. Approximating discrete probability distributions with dependence trees. IEEE T. Inform. Theory 14 (3), 462–467.

Cohen, Y., 1988. Bayesian estimation of clutch size for scientific and management purposes. J. Wildlife Manage. 52 (4), 787–793.

Crome, F.H.J., Thomas, M.R., Moore, L.A., 1996. A novel Bayesian approach to assessing impacts of rain forest logging. Ecol. Appl. 6 (4), 1104–1123.

Dixon, P., Ellison, A.M., 1996. Introduction: ecological applications of Bayesian inference. Ecol. Appl. 6 (4), 1034–1035.

Doorenbos, J., Kassam, A.H., 1979. Yield response to water. FAO Irrigation and Drainage Paper no. 33.

Ellison, A.M., 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecol. Appl. 6 (4), 1036–1046.

Haas, T.C., 1991. Partial validation of Bayesian belief network advisory systems. AI Applications 5 (4), 59–71.

Haas, T.C., 1992. A Bayes network model of district ranger decision making. AI Applications 6 (3), 72–88.

Jensen, F.V., 2001. Bayesian Networks and Decision Graphs. Springer, New York, USA.

Le Bars, M., Attonaty, J.M., Pinson, S., Ferrand, N., 2005. An agent-based simulation testing the impact of water allocation on farmers' collective behaviors. Simulation 81 (3), 223–235.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill Series in Computer Science, New York, USA.

Neapolitan, R.E., 2003. Learning Bayesian Networks. Pearson Prentice Hall, Upper Saddle River, NJ, USA.

Oliver, R.M., Smith, J.Q., 1990. Influence Diagrams, Belief Nets and Decision Analysis. Wiley, Chichester, UK.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., Cowell, R.G., 1993. Bayesian analysis in expert systems. Stat. Sci. 8 (3), 219–247.

Wolfson, L.J., Kadane, J.B., Small, M.J., 1996. Bayesian environmental decisions: two case studies. Ecol. Appl. 6 (4), 1056–1066.

Wright, J.L., 1982. New evapotranspiration crop coefficients. Proc. ASCE, J. Irr. Drain. Div. - ASCE 108 (IR2), 57–74.

CHAPTER 3
RECURSIVE PARTITIONING TECHNIQUES FOR MODELING IRRIGATION BEHAVIOR

Abstract

Accurate forecasts of short-term irrigation demands can provide information useful for canal operators to manage water diversions and deliveries more efficiently. This can be accomplished by analyzing the actions of the farmers who make water use decisions. Readily available data on biophysical conditions in farmers' fields and the irrigation delivery system during the growing season can be utilized to anticipate irrigation water orders in the absence of any predictive socio-economic information that could be used to provide clues into future irrigation decisions. Decision classification and the common factors, forming a basis for division of farmers into groups, can be then used to make predictions of future decisions to irrigate. In this paper, we have implemented three tree algorithms, i.e., classification and regression trees (CART), random forest (RF), and conditional inference trees (Ctree), to analyze farmers' irrigation decisions. These tools were then used to forecast future decisions. During the training process, the models inferred connections between input variables and the decision output. These variables were a time series of the biophysical conditions during the days prior to irrigation. Data from the Canal B region of the Lower Sevier River Basin, near the town of Delta, Utah were used. The main crops in the region are alfalfa, barley and corn. While irrigation practices for alfalfa are dependent on the timing of cuts, for barley and corn the critical crop growth stages are often used as indicators of farmer decisions to irrigate. Though all the models performed well in forecasting farmer decisions to irrigate, the best prediction accuracies by crop type

were: 99.3% for alfalfa using all the three models; 98.7% for barley, using the CART model; and 97.6% for corn, with Ctree approximately. Crop water use, which is the amount of water lost through evapotranspiration, was the prime factor across all the crops to prompt irrigation, which complies with irrigation principles. The analyses showed that the tree algorithms used here are able to handle large as well as small data sets, they can achieve high classification accuracy, and they offer potential tools to forecast future farmer actions. This can be subsequently useful in making short-term demand forecasts.

3.1 Introduction

In this data-rich world, there is a lack of pertinent information about certain phenomena that are either hard to model or lack a complete physically-based cause-effect description of the problem. This presents challenges in the use of conventional approaches such as deterministic models to predict future conditions. Such a problem exists in understanding and predicting a farmers decision to irrigate. Substantial scientific theory and large quantities of data are available to analyze the irrigation problem and forecast short-term irrigation demand, but the problem of accurately anticipating short-term water demand of an individual irrigator still remains. This is due to a limited understanding of the irrigation practices that are followed by different farmers and how farmer preferences influence decisions about the timing of irrigation.

The Canal B command area, the site selected for this study, is equipped with technologies to monitor reservoir releases and canal diversions, and it has dependable forecasts of evapotranspiration (ET). The entire Canal B command area is approximately 30 square miles, with around 20 square miles falling under irrigation in an average year. Some of the irrigated fields have real-time soil moisture measurements to study agricultural water use in the area. In spite of these developments, day-today irrigation demands are difficult to forecast. Information about such demands can be vital to help irrigation system operators achieve greater efficiency in water deliveries. In an on-demand irrigation delivery system, farmers make the basic water use decisions. Hence it is essential to consider their decision-making mechanisms in forecasting short-term irrigation demand.

Irrigation behavior has rarely been a topic of research. Each farmer has personal goals to achieve in a season, ranging anywhere from profit maximization, to crop quality, to being environmentally conscious about saving water. Because of different opportunity costs, a farmer whose primary profession is agriculture will make different choices from the one who considers agriculture as a secondary occupation. These characteristics make it more difficult to forecast behavior. The few studies that have dealt with irrigation behavior have been inconclusive in understanding the factors that contribute to decisions regarding if and when to irrigate. To find out the scope of studies done on farmers behavior previously, we are presenting some of the notable ones in the field.

Becu et al. (2006) used a multi-agent system for a study of water sharing between two villages located at the upstream and downstream ends of a watershed. The objective was to evaluate various options to allot water to the villages and provide feasible solutions to different water users for dealing with water scarcity. The solutions were found by analyzing the impact of different land use and water management options on the water deficit. Since it involved water use decisions, a farmers behavior was considered in terms of what crops are planted, when they are harvested, and how they are irrigated during the season. Farmers were grouped into various classes on the basis of different cropping patterns identified in the region. This study simulated irrigation decisions taken by the farmers on the basis of the crops they were growing. Bontemps and Couture (2002) developed a sequential decision model to study water use behavior under conditions when the farmers paid a negligible amount to obtain water and there was no charge for supplying it. Le Bars et al. (2005) developed a multi-agent systems paradigm to simulate farmer-agents and their decisions over a number of years, under conditions where water supply was limited. The water manager controlled, the amount of water given to a farmer by using allocation rules that were based on the amount of water requested by farmers at the beginning of the season. The farmer-agents each owned a farm with several plots and could decide their own cropping plan. Weather variables were generated in the simulation at random.

From the limited literature on farmer irrigation decision behavior, it is clear that few studies have been conducted to analyze decisions already made or to forecast future irrigation decision under simulated conditions. Models that could provide such forecasts could be potentially useful for improving irrigation system operations. The study reported here is a first attempt at analyzing farmers' decisions using "decision" trees. We use data about the biophysical conditions during the growing season to isolate information available to the farmer about differences on the days leading up to the time of irrigation. We also look into the possibility of using those differences to forecast farmer decision-making.

3.2 Theory

A wide range of machine learning techniques is available today to address modeling problems, where missing information is an issue. These show promise for the analysis of problems involving the forecasting of decision behavior under conditions where it is not possible to quantify all of the process-specific factors that affect the decision. Figure 11 shows a tree structure. The nodes are the variables related to the process in form of root, intermediate or terminal nodes (which do not have any child nodes). As we descend in the tree the importance of the variable to the process

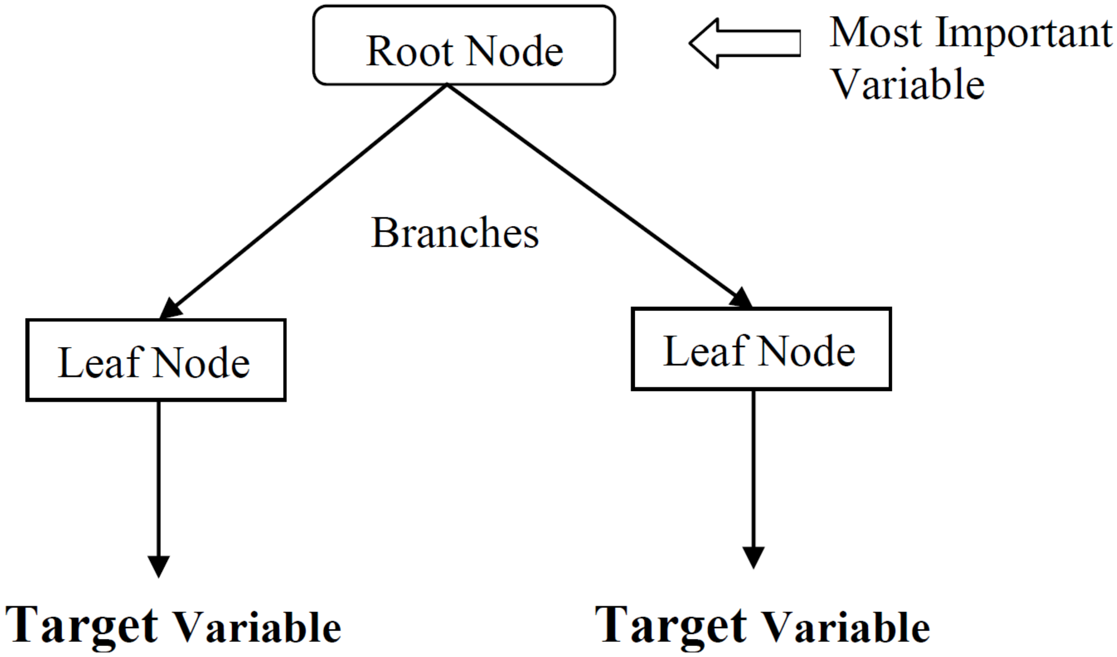


Fig. 11: A tree structure showing root/parent nodes, branched into leaf nodes which can be intermediate nodes or terminal nodes and the variable of interest or the target variable.

decreases. The variable at the root node is the most important. The effect of all the variables leading to the terminal node is collective.

Trees are used to understand systems that have little a priori information about how and which variables are related. Classification trees have been used by Kastellec (2010) to analyze judicial decisions and laws. Random forests have been used to model ecology applications (Cutler et al., 2007). Das et al. (2009) used conditional inference trees to assess crash severity in road accidents and found the factors involved. These are some applications of trees to real problems.

Decision trees have been a powerful tool for classification and forecasting. The features and capabilities of the trees are described ahead (Hill and Lewicki, 2007). They can give insights into nonparametric, nonlinear relationships between a large number of continuous and/or categorical predictor inputs, and output variables, which may be continuous or categorical. When the output variable is continuous it is a regression analysis and when categorical then a classification problem. They divide a heterogeneous group of features into small homogeneous groups with respect to the output variable. A binary tree, formed by two child nodes split from each parent (root) node, is one such structure. The split is best when it separates the data into groups with two different predominant classes. "Levels" in a tree are referred to as the depth of the tree, and "size" is the number of nodes in the tree. The measure often used to estimate the split is known as "purity." The best split is defined as the one which increases the purity of sub-groups by a considerable amount and creates nodes of similar size (not very small ones). Dense structures can often be simplified by pruning. The tree models make no prior assumptions about the data. No unit conversions are required. Raw data can be used as it is. The variables at the root of the tree are deemed the most important.

3.2.1 Classification and Regression Trees

Breiman et al. (1984) popularized the classification and regression tree, commonly known as C&RT or CART algorithm. Binary recursive partitioning forms the basis of CART analysis (Breiman et al., 1984). For the analysis, binary partitioning is repeated, hence the term recursive is used. Each parent node results in two child nodes and these nodes are further split into other child nodes. The dataset itself is partitioned into sections to form homogeneous groups with similar features. As an example, we assume that the categorical variable at node t has two responses: 'Yes' and 'No'. There are basically four steps in CART analysis Breiman et al. (1984):

i. Tree building : All the data are at first placed at the root node. During learning, the first variable in the sample is split at all the possible values in the data. For each split there are two resulting nodes, a 'Yes' and a 'No' response. All the cases with corresponding responses, ‘Yes’ and ‘No’, are classified accordingly. A node is assigned a class, even though it may or may not be split further into child nodes. After this a goodness-of-split criterion is applied to each split to assess the reduction in impurity (or heterogeneity). In CART, one of the measures of impurity is the Gini Index given as:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t)$$

if the misclassifications are equally costly, where, the $p(i|t)$ and $p(j|t)$ are the probabilities of category i and j, respectively, at the node t. Then the best split on the variable is the one for which the reduction in impurity is the highest. The above is done for all the remaining variables. In the next step, CART ranks the “best” splits on each variable according to the Gini Index. The best split is the one which most reduces the Gini Index (Breiman et al., 1984). We repeat the above steps for all the non-terminal child nodes of the tree.

ii. Stopping criterion: At this point a large tree has been produced which over-fits the information contained within the learning data set. New splits are stopped when they result in very little or no improvement in the predictions. Resubstitution error rate of the classifier is one accuracy estimate used at this point. It is the proportion of cases misclassified on the same sample that was used during learning. It is given as (Hill and Lewicki, 2007):

$$R(d) = \frac{1}{N} \sum_{i=1}^{N} X(d(x_n) \neq j_n)$$

where, $d(x)$ is the classifier and X is the indicator function such that,

$X = 1$, if $X(d(x_n) \neq j_n)$ is true and

X=0, if $X(d(x_n) \neq j_n)$ is false.

This indicator is usually biased and underestimates the true error rate.

iii. Tree pruning: CART reduces the splits using 10-fold cross-validation (for details refer to Hill and Lewicki (2007); Breiman et al. (1984)), which results in the creation of a sequence of simpler tree structures by removing the unimportant nodes. For example, if 10 splits result in 90% accuracy during prediction, but 11 splits result in 91% accuracy, then 10 splits is preferred. This also gives more accurate ('honest') estimates of the (true) prediction error in comparison to the resubstitution error. For a sequence of trees, these estimates of error are plotted against tree size, and the size with the minimum error is selected.

iv. Optimal tree selection: The tree which does not over fit the information in the learning dataset is selected from a sequence of pruned trees by evaluating the resulting cross-validation error. Breiman et al. (1984) suggested using the 1-SE rule where in the optimal tree is the smallest tree such that its estimated error rate is within one standard error of the minimum. The test sample estimate (error) is then evaluated as the mean squared error between the predicted and the observed data.

CART analysis also produces a variable importance table. This provides a list of all the explanatory variables used and not used in the tree-building process with a score linked to each variable. This score is based on the improvement each variable makes as a surrogate to the primary (the one which shows up in the tree structure) splitting variable. The variable with the highest sum of improvements is scored 100, and all other variables are scored lower, descending towards zero. This helps identify the variables whose significance is disguised by other variables in the tree building process.

3.2.2 Random Forests

The random forest (RF) algorithm uses many decision trees to perform ensemble classification (Breiman, 2001). Random forest algorithms typically have good accuracy, do not over-fit, run efficiently on large data sets with complex interactions, and are robust (Strobl et al., 2008). In RFs, individual classification decisions from a large number of random classifiers (trees) are grouped. This is done when the tree “votes” (provides a classification) for that class. The classification with the largest number of votes across all the trees is chosen by the forest (which is comprised of N classification trees). This ensures more accurate predictions than a single tree classifier. Each tree in RF is built using bootstrapped samples, equivalent to approximately 63% of the observations from the original dataset (Cutler et al., 2007), with replacements, leaving about one-third of the cases which are termed as *oob* (out-of-bag) data. Both inputs and variable selection are used randomly at each split in the tree. If there are N explanatory variables, then a number $n < N$ is selected such that n variables are randomly selected out of N and the best split on these n variables is used for node splitting. The number of variables n, remains constant during forest growth. The trees are not pruned. When a large number of classification trees have been grown, class membership of new data is predicted for the *oob* cases. Though the *oob* cases are from the original data set, they do not occur in a bootstrap sample. The predictions for these cases provide unbiased (cross-validated) estimates of the prediction accuracy of the model (Cutler et al., 2007). All the new cases are sent down the trees starting from the root. Every tree in the forest gives its classification for those cases at their terminal nodes. For example, if the classes are “Yes” and “No” , then number of trees having “Yes” classification/votes are counted, and the percent of “Yes” votes of the total votes is the predicted probability. This gives a combined predicted classification and is referred in the literature as “majority voting”. Error rates are estimated using the *oob* predictions and are averaged over all the cases in the data set. Each tree in the forest can be associated with a misclassification rate for the *oob* cases. For variable importance, the values of 'n' predictor variables in the *oob* data are randomly permuted and put down the tree to get new predictions. The measure of importance of the variable is the difference between the misclassification rate for the original and the permuted *oob* data, divided by the standard error (Cutler et al., 2007). Details can be found in (Breiman, 2001).

3.2.3 Conditional Inference Trees

Conditional inference tree (Ctree) models regress relationships between predictor variables and target variables by recursively partitioning data in a conditional inference framework (Hothorn et al., 2006). The trees are built using the following steps: The global null hypothesis of independence between the input and output variables is tested. The model terminates if the hypothesis is not rejected. In the case when the hypothesis is rejected, the algorithm selects that input variable which is strongly associated with the target variable using a p-value resulting from a test for partial null hypothesis. A binary split is performed on this input variable. Testing and splitting is repeated for all covariates, recursively. A certain stopping criterion based on hypothesis tests is adopted (e.g. $p < 0.05$). This usually avoids any over-fitting or biased variable selection (Hothorn et al., 2006). By using the Gini index, the chances of finding a good split increases if the variable is continuous or has numerous categories. CART is found to have a bias in variable selection for continuous variables. Conditional trees use a $chi-square$ significance test for variable selection, as opposed to CART which selects the variable that maximizes an information measure like the Gini index. In spite of these advantages the model is still new and experimental.

3.3 Case Study and Methods

3.3.1 Study Site and Data Available

The data used in this study are from the Canal B region of the Lower Sevier River Basin, near the town of Delta in south-central Utah This area covers approximately 20 square miles of irrigated farm land. Alfalfa, barley and corn are the main crops grown in the area. Irrigation consumes a large amount of water in this basin. The station was established by National Climatic Data Center (NCDC), NOAA and maintains historical weather data since 1965. The station metadata are as follows and can be located on the NCDC-NOAA website (http://www.ncdc.noaa.gov/) using them:

GHCND ID : USW00023162

COOP ID : 422090

WMO ID : 72479

NCDC ID : 20026236

Weather data for Delta was obtained from the following website: `http://www.cemp.dri.edu/cgi-bin/cemp_stations.pl?stn=delu`. Data estimated using Kimberly Penman Reference ET rules are available on this website.

Table 5 presents the variables used to build trees and predict the irrigation decision. Variables 1, 4, 5, 6, and 21 are weather variables. Data were also available for canal flow rates (Variable 13). SCADA (Supervisory Control and Data Acquisition) systems maintain the past records of water discharge across Sevier River Basin (Berger et al., 2002) and real-time conditions are accessible on http://www.sevierriver.org/. The site is sponsored by Sevier River Water Users Association (WUA).

Table 5: Predictor variables, the represented factors as seen by the farmer, and the target variable used for trees analysis. The variables have been compiled for this study considering the plant-soil-water components which the farmer is thought to evaluate before making an irrigation decision.

S. No.	**Variable Name**	**Represented Factor**	**Continuous OR Categorical - (No. of classes)**
1	AirTemp	Average Air Temperature	Continuous
2	GrowingDegDays	Growing Degree days accumulated till a given day and reset on the day of irrigation	Continuous
3	GrowStageIrrigNeed	Sensitivity of Growth Stage to water stress as indicated by growing degree days	Categorical -(2)
4	WindSpeed	Wind speed	Continuous
5	RH	Relative Humidity	Continuous
6	ET	Potential Evapotranspiration (ET)	Continuous
7	ET$_c$	Crop Evapotranspiration	Continuous
8	CropCoeff	Crop-specific coefficient	Continuous
9	SoilStressCoeff	Soil Stress Coefficient	Continuous
10	ET$_a$	Actual Evapotranspiration	Continuous
11	CumET$_c$	Cumulative Crop ET	Continuous
12	StressIrrigNeed	Consumptive use as indicated by CumET$_c$	Categorical -(2)

Continued on next page

Table 5 – continued from previous page

S. No.	Variable Name	Represented Factor	Continuous OR Categorical-(No. of classes)
13	CanalFlow	Canal Flow rates	Continuous
14	WaterSupplyIrrigNeed	If the farmer irrigated when his neighbors irrigate as indicated by CanalFlow	Categorical -(2)
15	JDay	Julian Day in the season	Continuous
16	WeekEndORNOT	Saturday/Sunday	Continuous
17	WkEndIrrigNeed	If the farmer irrigated on a weekend as indicated by WeekEndORNOT	Categorical -(2)
18	RootingDepth	Rooting depth of the plant	Continuous
19	CropIrrigNeed	Plant need indicator, deeper the root, frequent are the needs for water as indicated by RootingDepth	Categorical -(2)
20	SMCinit	Soil moisture content at the start of the day	Continuous
21	Rain	Precipitation amount	Continuous
22	AmountPercolation	Amount of irrigation water percolated	Continuous
23	IrrigationAmt	Estimated amount of irrigation from the soil moisture probes	Continuous

Continued on next page

Table 5 – continued from previous page

S. No.	Variable Name	Represented Factor	Continuous OR Categorical-(No. of classes)
24	DepInit	Depletion at the Start of the day	Continuous
25	DepEnd	Depletion at the End of the day	Continuous
26	SoilIrrigNeed	If the soil is dry or not (also indicated in plant condition) as indicated by SoilStressCoeff	Categorical -(2)
27	Year	Year, indicating a dry, moderate or wet year	Categorical -(4)
28	Yield	Yields estimated using ET_a and ET_c	Continuous
29	MarketPrice	Price of the crop	Continuous
30	Revenue	Revenue for the farmer	Continuous
31	EconIrrigNeed	Economic need to irrigate the crop as indicated by Revenue	Categorical -(2)
32	ID	Different farmers	Categorical -(39)
33	SoilType	Type of soil	Categorical -(2)

Continued on next page

Table 5 – continued from previous page

S. No.	Variable Name	Represented Factor	Continuous OR Categorical-(No. of classes)
34	Irrigate	The irrigation decision	Categorical -(2), (0-1 for regression and Yes-No for classification)

Daily canal flows in cubic feet per second (cfs) in Canal B were obtained from the URL: `http://www.sevierriver.org/rivers/delta/b-canal/`. Three types of soil are found at Delta: silty clay loam, silty clay, and loam. Farmer identification numbers convey information to the model that the data are from a different subject. In 2007 Utah Water Research Laboratory (UWRL), Utah State University (USU) established 44 stations with 88 sensors to record soil moisture content at 1 and 2 ft depths on various farms in Canal B irrigation command area to study agricultural water use. The sensors are maintained by the personnel at UWRL. Soil moisture content measured at these stations was used to determine the day of irrigation and the approximate amount of irrigation. These were obtained from: `http://odm.usu.edu/odmmap/default.aspx?NetworkName=Delta`. Hourly measurements are available on the website, so daily average values were estimated for the first day of the season, for starting the soil moisture balance calculations. Soil moisture contents were corrected using a mass balance constraint on soil moisture. Variables 7-10, 18, and 20-25 are required for the soil moisture balance computation. A daily time series was created for market prices of alfalfa, corn, and barley using the monthly data available for the USDA website (`http://www.nass.usda.gov/Statistics_by_State/Utah/Publications/Annual_Statistical_Bulletin/`) for Millard County, Utah. Single value for monthly prices are posted on the website. These prices for previous years (2007-2010) were compiled from the site. These single monthly values were linearly interpolated between months to obtain daily values. Approximate planting dates were established by initiating the soil moisture calculations from a random day such that soil moisture matched the day of first irrigation, which was known from the soil moisture probe data. We assumed that the initial depletion was zero and began the computations from field capacity.

Phenology coefficients (K_c) for all the crops were derived from Wright (1982) and FAO-56 (Allen et al., 1998) and were found to be representative of the conditions in Delta (personal communication with Dr. Robert Hill, Irrigation & Water Resources Specialist and Extension Expert, USU). K_c estimates in Wright (1982) were obtained from experimental studies near Kimberly, Idaho. Since we were using the values of K_c for crop reference ET, we had to multiply the values with a factor of 1.2 to model a field crop, instead of grass reference ET.

All the other variables were either derived from the primary data or categorized to simplify their representation in the model. If both the data and the derived variable behave the same, then the derived ones can be removed.

3.3.2 Models, Specifications and Performance Evaluation

All of the models used in this study were implemented using R-statistical software (R Development Core Team, 2007). For all the classification problems it is necessary that the target variable be categorical (e.g. “Yes” and “No”), which can be done using the “factor” function. For all the models, the data were randomly partitioned into training and testing sets. For all the data sets, one-fourth of the data were used for testing. The input to all these algorithms for our case was the decision to irrigate ("Yes") or not ("No"). During training, the model was tuned according to the irrigation decision. During testing, this target variable was forecasted. The outputs of all the models were the confusion matrix and the error rate. We used the accuracy rate (calculated as the difference between 100 and the error rate) and the confusion matrix to evaluate model performance.

To do a CART analysis, we used the "rpart" (Therneau et al., 2012) package in R. It is powerful and easy to use, and is based on the same algorithm as Breiman et al. (1984). During the model fitting process, the "rpart" function was applied to the training data with the dependent variable being the irrigation decision and method = "class". The "predict" function uses the fitted tree and predicts the classification for the test data set. The "table" function can be used to obtain a confusion matrix, and the accuracy rate can also be obtained. The accuracy rate is the sum of the diagonal elements of the confusion matrix, divided by the sum of all the elements. The "printcp" function can be used to print the complexity parameter (cp) table (Table 6 shows a sample output) for the fitted tree. This way we can find the optimal pruning of the tree based on 'cp'.

Table 6: Cost-complexity parameter table to find optimal pruning.

S. No.	**Cost-complexity parameter (cp)**	**Number of splits (nsplit)**	**Relative error (rel error)**	**Cross validation error (xerror)**	**Cross validation standard deviation (xstd)**	**Sum of xerror & xstd**
1	0.709	0	1	1.069	0.05	1.119
2	0.03	**1**	0.291	0.291	0.035	**0.326**
3	0.02	5	0.172	0.345	0.038	0.382
4	0.015	7	0.133	0.33	0.037	0.367
5	0.01	8	0.118	0.31	0.036	0.346

It can be seen in Table 6 that ‘nsplit’ (number of splits) denotes the size of the tree, and (‘nsplit’ + 1) is the number of nodes in a tree. Scaled errors are presented such that the error at the first node is 1. Using the 1-SE rule to find the best number of splits, the smallest “xerror” is added to the corresponding “xstd” as shown in the last column. The number of splits resulting in the smallest error is the “best split”, e.g. in this case, the optimal number of splits is 1. For this case the pruned tree will have 2 nodes. If in case the sum of “xerror” and “xstd” are all equal, the “best split” will be the tree with the fewest number of splits other than zero, which would leave only the root node. To assess the performance of a predictive model, we have also used cross-validation. We cross-validated using bagging (Breiman, 1996), derived from “**b**ootstrap **agg**regat**ing**. Bagging is an ensemble method which decreases the variance of the original individual models by using a bootstrap of the training set to build every new model and then taking the average of the predictions from those models. For running 10-fold cross-validation with bagging we used the “adabag” (Alfaro et al., 2012) package in R. For cross-validation, two-thirds of the data were used for training and the remaining for testing. The function “bagging.cv” requires the target variable as input from the training data set and the number of iterations, “mfinal” variable (default =100). The most important variable is at the root node.

We used the “randomForest” (Liaw and Wiener, 2002) package of R for model development. The “randomForest” function needs the following as parameters for tuning the forests

- “mtry” is the number of variables randomly sampled as possible choice at each split,

- “ntree” is the number of trees to be populated,

- “importance=TRUE” calculates the variable importance and can be retrieved using the “varImpPlot” function, which plots “MeanDecreaseAccuracy” and “MeanDecreaseGini”. The plot presents the variables in the descending order of importance. We used “MeanDecreaseGini” to assess the important variables. This is related to the fact that the stopping criterion of splitting in a tree is when a new split does not reduce the Gini index any further (Breiman et al., 1984). This means that the variable is not important for tree building. “rfcv” with “cv.fold =10”, was used to evaluate the 10-fold cross-validated prediction performance of the models. This was done by sequentially reducing the number of predictors (as ranked in variable importance) using a nested cross-validation procedure. We also used out-of-bag error estimates also to evaluate the models.

For the analysis of conditional inference trees, the “party” (Hothorn et al., 2012) package in R library was used. “ctree_control()” sets the parameters for the tree. We ran with the default settings, which include mtry=0, implying the variables were not selected randomly at each split. These settings were chosen to avoid different trees at each run. This feature is available in the “randomForest” package but we did not find a similar parameter in the “rpart” package. This makes it different from the classification trees analysis. The input to the “ctree” function was the irrigation decision and parameters were supplied through “ctree_control()”. The “plot” command was used to produce the resulting trees. The mean “y” and the number of cases “n” ending up at the terminal nodes are also displayed. The predictions were obtained using the “predict” function and the accuracy rate was calculated.

3.4 Results and Discussion

Figure 12 presents a plot of some of the weather variables to illustrate some existing groups based on the irrigation decision. No obvious classes can be found in

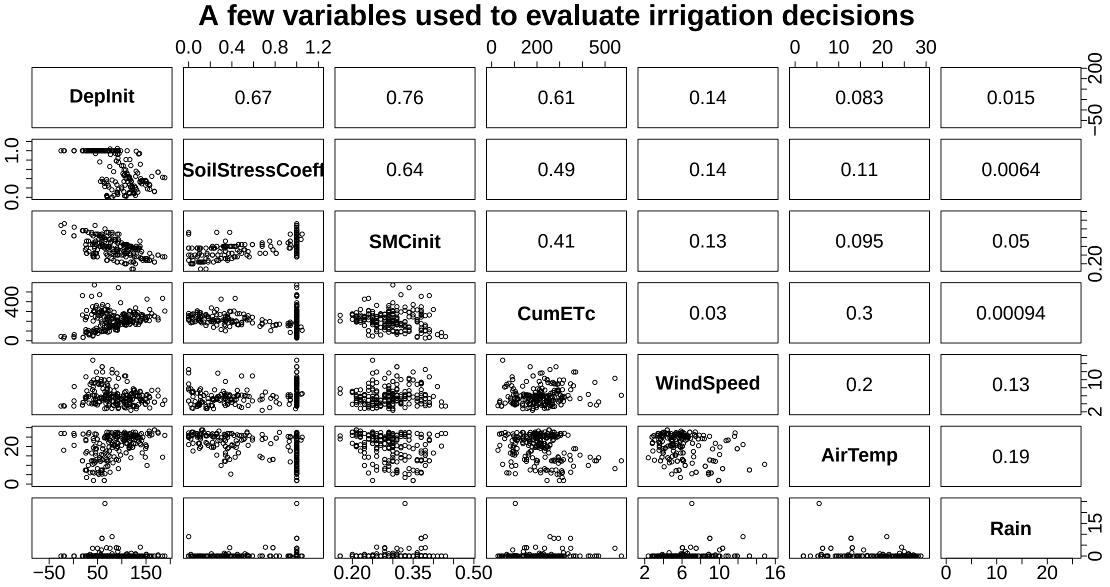


Fig. 12: Pairs plot of some weather variables used in the tree analysis, with the intention of finding groups of similar features.

the plot. Obvious classes can be distinctly found in the plots if there are two or more clusters in the data representing groups. If apparent groups existed, we would not use such algorithms. Usually in practical problems it is difficult to find classes to discern groups based on the target variable. We have to then seek the help of specialized techniques like recursive partitioning, in our case, for exploring the groups, if any.

Since the day of irrigation can be anywhere from 24 hrs to 3 days from the day of order (call-time), we decided to examine the importance of the variables presented in Table 5 for:

- all the days in the season (referred to as “All days” in the results)
- four days before the day of irrigation and the day itself (referred to as “4-days” in the results)
- the day of irrigation and a day before (referred to as “1-day” in the results).

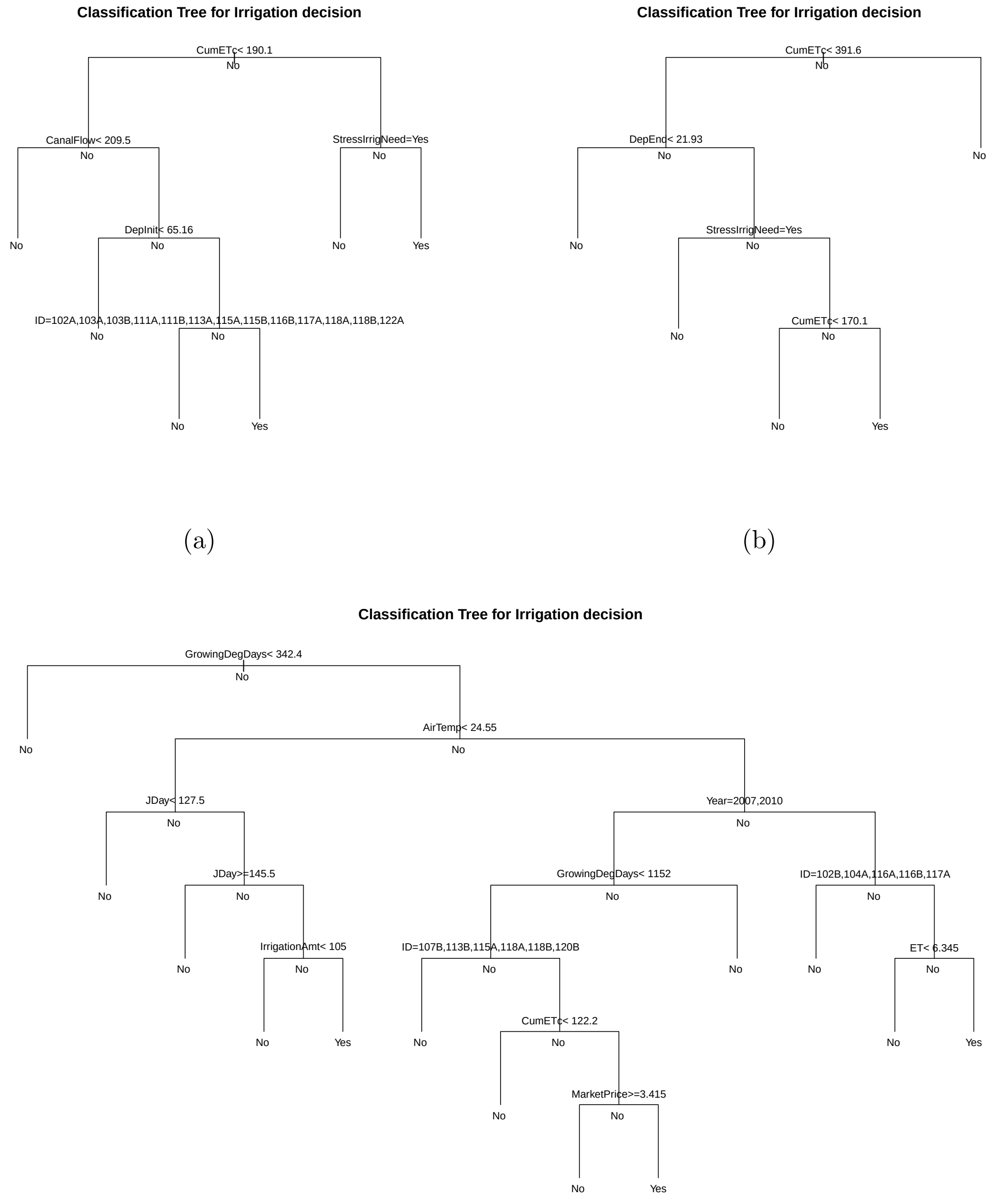
The prediction accuracy for the tree analysis is presented in Table 7. The best results are displayed in bold. It can be seen that the predictions of the models, which were given the description of the whole season (i.e., the All-days model), performed better than the 1-day or 4-day models. Given that there is a possibility of some missing information, all three algorithms work exceptionally well to predict future decisions. All three algorithms had close to 99.0% accuracy for alfalfa irrigation decisions. CART had accuracy estimates of 98.7 and 96.9% for barley and corn. RFs predicted the decisions for barley and corn with an accuracy of 97.8 and 96.2%, respectively. Ctree had accuracy measures of 98.0 for barley and 97.6% for corn.

Table 7: Accuracy estimates on test data for CART, RF and Ctree models. *Resub*-Resubstitution accuracy estimate, Xval- 10-fold Cross-validation accuracy estimate. (a) 1-day, (b) 4-day, (c) All days models.

Crop	**Alfalfa**			**Barley**			**Corn**		
Model	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**
CART-Resub	83.5	91.1	99.2	48.5	83.1	**98.7**	66	84.9	96.8
CART-Xval	83.3	94.5	99.4	53.4	81.2	98	46	82.1	97.1
RF	85.6	97.1	**99.3**	78.8	81.9	97.9	51.1	84	96.2
Ctree	80. 6	93.1	99.3	57.6	91.6	98.1	44.7	80.7	**97.6**

3.4.1 CART

A 10-fold cross validation was done on all the datasets and the cross-validation accuracy was reasonably close to the resubstitution accuracy, except for the corn 1-day model, as shown in Table 7. Cross-validation is performed to help determine if the classifier is being over fitted. For our case, the performance is promising. Figure 13 shows trees built for the three crops in the study. Clearly there are different strategies for the three crops, starting with the same variable. The most important variable for all three crops was cumulative crop ET, or the consumptive use.



(a)

(b)

(c)

Fig. 13: CART structures for 13(a) - alfalfa, 13(b) - barley, and 13(c) - corn irrigation decisions.

According to many standard agricultural handbooks, crop growth stage and its sensitivity gauged by accumulated heat units (or growing degree days, “GDD”), and the consumptive use that this implies, is almost always considered by farmers when deciding about irrigation. StressIrrigNeed is the indicator for critical CumET$_c$, so it was expected that it would be next in importance in explaining irrigation decisions. At each terminal node, there is a group based on a variable or its level which makes it different from the others, according to CART interpretation. In other words, there is a class at every node where we see either a “Yes” or a “No” response. Since we are interested in the reasons behind irrigation decisions, we are only evaluating “Yes” responses.

Figure 13(a) shows that the alfalfa growers in the first group were irrigating when their neighbors irrigated (Canal Flow $>$ 209.5cfs, which was a high flow rate), at medium soil moisture depletion ($>$ 65.16 mm), and when farmers other than the ones shown in Figure 13(a) were irrigating according to these common rules. Since we wanted to avoid bias in selecting the training and testing sets, we used bootstrapping to sample the data sets. It also becomes important to note that these data sets were a mix of information from any of the four years (2007-2010). The second class evidently used the CumET$_c$ (StressIrrigNeed = Yes) measure to time the irrigations. CumET$_c$ is the crop ET accumulated between irrigations and is the same as depletion. As a general rule, alfalfa is primarily irrigated either just before or after it is cut. Although if it is irrigated after the cut, then the drying of the bales on the fields might further delay it. Proper drying can add value to alfalfa hay, hence this can be an important factor. All the resulting principles conform to recommended irrigation practices, since depletion would not be used generally as one of the indicators to trigger irrigation for alfalfa. The possibility that farmers are choosing to irrigate when their neighbors are irrigating is also refuted since the fields involved have different crop planting dates. We can say this because different planting date would mean that farmers might not irrigate at the same time, if they are using crop growth stages or soil moisture conditions to irrigate. Where one farmer's crop might be moisture stressed, at the same time due to delay in planting another farmer's crop might not be under stress. This hints at different maturity and cut timings and might even point at a different crop quality.

Barley (Figure 13(b)) irrigation strategies were very straight forward, with the low depletion of 22 mm and consumptive use (>170 mm) being the only indicators CART could discriminate. This suggests that the farmers who were growing barley were not taking risks with respect to the irrigations because this would mean loss in yield to them, even though they could wait longer to irrigate. We did not have enough data to make strong conclusions but these may be probable reasons for the observed timing of irrigation.

For corn (Figure 13(c)), CART presented a huge tree with many variables. It clearly showed three classes (i.e., three terminal nodes with "Yes" as the irrigate decision). For group one, the day in the growing season (JDays between 127 and 145 were crucial) and an irrigation amount more than 105 mm appeared to be the driving factors. The day in the season is indicative of a certain critical growth stage for corn. The irrigation amount may seem a strange choice for grouping farmers, but it implies that farmers who replenished the soil moisture to this level would irrigate similarly. This corresponds indirectly to consumptive use. For group two, in years 2007 and 2010, the farmers other than those shown irrigated while ET was constraining on the crop. This meant that temperatures could have been high for long periods when the farmers decided to irrigate. In the third class, the group irrigated when consumptive use was more than 122 mm and the predicted market price was higher than before. The consumptive use for corn is always a driving force for irrigation. If there is high moisture stress, the amount of carbohydrates available for kernel development in corn is inhibited, which can affect the yield. The implication is that corn growers were keeping the stress levels in control by irrigating at moderate levels of moisture depletion.

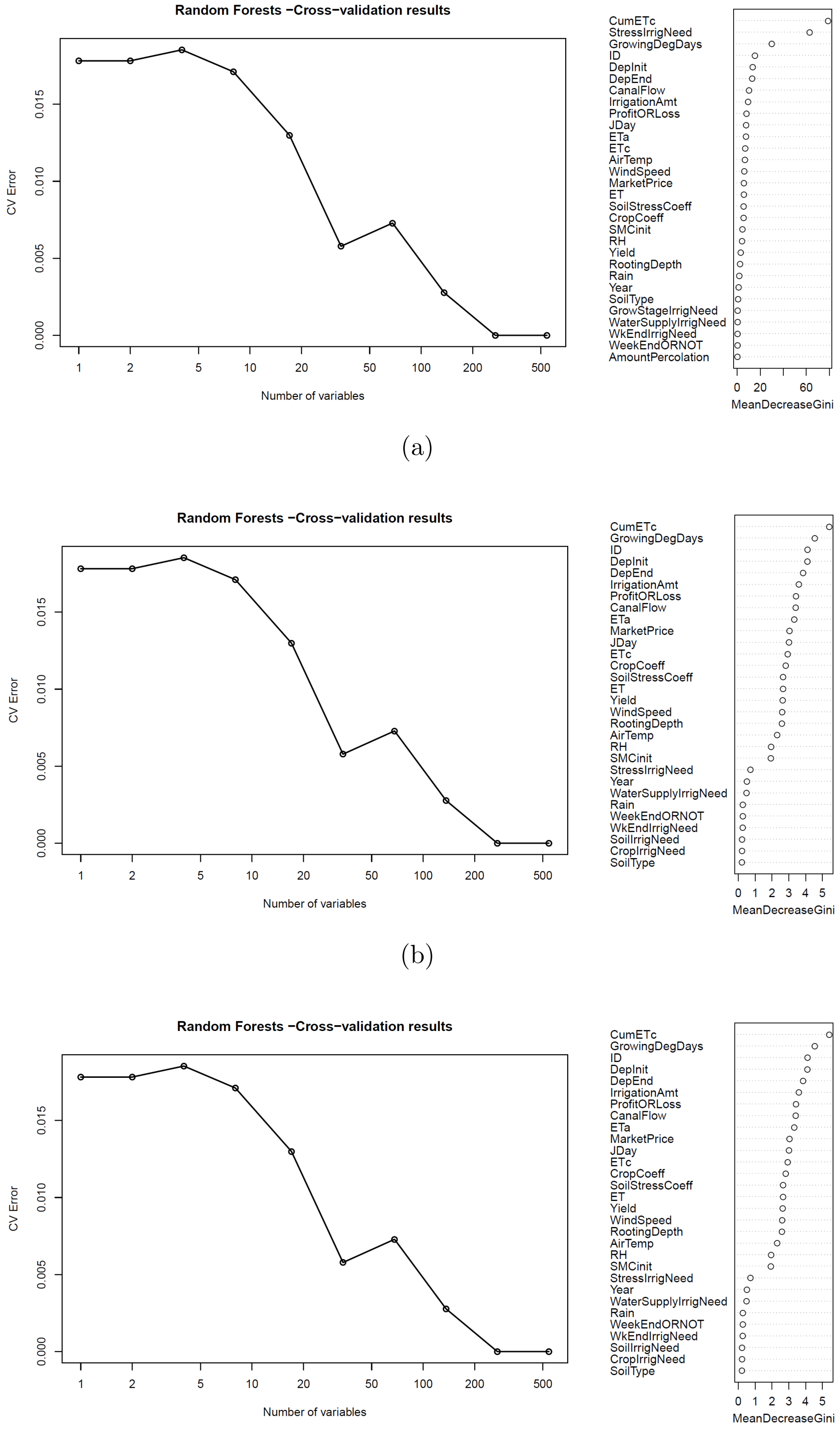
The only marginal costs to the farmer of irrigating in the Canal B area are associated with the cost of hiring labor for irrigation and/or ditch and gate maintenance. This results in marginal input costs that are very small in comparison to other costs of production.

CART pruned trees performed the same as the full grown versions, but the advantage was the smaller number of variables for interpretation. We have opted for the full grown versions, however, since it gives us an in-depth analysis of the factors leading to irrigation. Though the pruned tree narrowed the choices of variables it does not refute the fact that farmers consider multiple factors in the thought-process of scheduling irrigation.

3.4.2 Random Forests

Random forests are a modern tool for classification. Since they have several trees, they have generally been shown to perform exceptionally well in grouping predictor variables according to target decisions. A 10-fold cross-validation (CV) of the All-days data is shown in Figure 14.

The CV error evaluates the effect of adding input variables to the classifier in the order of importance. For alfalfa, Figure 14(a), the error remained the same at 1.8% for the addition of 1 to 5 variables, but started dropping sharply as the number of variables increased. For barley, Figure 14(b), the error was unstable throughout. With variable addition from 1 to 5 it stayed around 2.2%, but with 20 variables it dropped to around 1%. For variable additions between 20 to 50 it rose to 1.5% and



(a)

(b)

(c)

Fig. 14: Random forest 10-fold cross-validation performance and variable importance plots using Gini Index for 14(a) - alfalfa, 14(b) - barley, and 14(c) - corn.

then dropped down again. For corn, Figure 14(c) , the error was inconsistent, with variables 1 to 10 maintaining a constant error of around 3%. Further addition of variables dropped the error to 0.5%, but it went up close to 2% between variables 50 to 100. The instability in the errors might be because alfalfa had large amounts of data for learning, while barley and corn did not. In spite of the lack of data, RFs performed better in predicting a decision to irrigate for the test cases across crops, as shown in Table 2. The data sets with limited information (1-day and 4-day models) were not found to perform well. The proportion of times that the predicted class is not the same as the observed class averaged over all cases is the oob error estimate. The out-of-bag error estimates were 0.59%, 1.89% and 2.66% for alfalfa, barley and corn crops, respectively. These indicated that the model built to forecast alfalfa irrigation practices was more reliable than the ones for barley and corn. The driving factors for irrigation in alfalfa, barley and corn as found in CART were confirmed in random forests and are distinct from the Gini Index. Consumptive use and growing degree days were found to be important factors. This means that temperature-related factors were found to be important by RFs.

3.4.3 Conditional Inference Trees

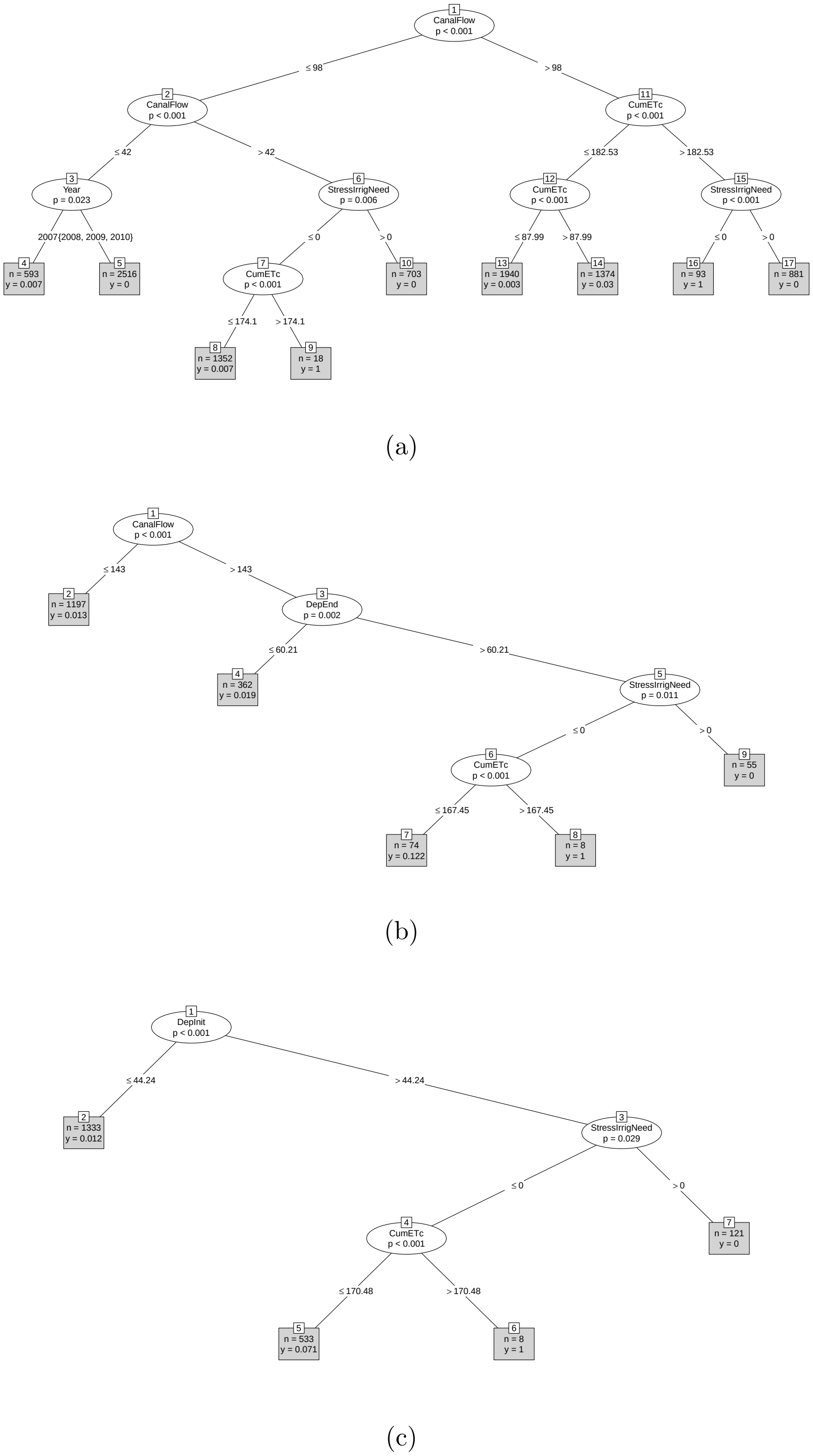
Conditional inference trees perform regression. Since we had continuous covariates, we attempted to analyze them using Ctrees. With the exception of a few cases where data were limited, they performed well (refer to Table 7). The tree structures are presented in Figure 15. They gave insights into some factors which were ignored by other algorithms (Table 8). Ctree showed that alfalfa (Figure 15(a)) and barley (Figure 15(b)) growers might be irrigating with their neighbors (as measured by high canal flow levels), but we do not have any related information to confirm this suggestion. Additionally for alfalfa (Figure 15(a)), year and CumET$_c$ helped farmers to decide the irrigation timing. "Year" variable denotes the year when the measurements were recorded. While CumET$_c$ due to high temperature is justified, the year factor would seem something strange. However, it is pertinent to alfalfa since it is a perennial crop and will be typically cultivated for a period of 3 to 5 years. The first year irrigation practices will be different for alfalfa since the crop will be germinating and growing, as opposed to the other years where it will emerge and the crop root will already be developed. For barley (Table 8), the farmers also used depletion levels, besides CumET$_c$, to decide irrigation timing. Corn planters (Figure 15(c)) used consumptive use (CumET$_c$ and depletion) to make irrigation decisions.

Table 8: Important variables for irrigating different crops, according to various models.

Model/Crop	**Alfalfa**	**Barley**	**Corn**
CART	CumET$_c$-Canal Flow	CumET$_c$-Depletion	GDD-AirTemp-Day-Year
RF	CumET$_c$-GDD	CumET$_c$-GDD	CumET$_c$-GDD
Ctree	Canal Flow-Year-CumET$_c$	CanalFlow-CumET$_c$-	Depletion-CumET$_c$

3.5 Conclusions

Irrigation system managers would benefit from information about short-term irrigation demand. This study applied different classification and regression trees to infer how farmers, the water users, make irrigation decisions. This information can be used to predict future actions and forecast short-term water demands, relying on readily measurable biophysical data alone as input. The results from this study show that biophysical variables can be used as indicators of irrigation behavior, and have a potential to be used as predictors for future irrigation decisions.



(a)

(b)

(c)

Fig. 15: Conditional inference trees for for 15(a) - alfalfa, 15(b) - barley, and 15(c) - corn irrigation decisions.

The tree algorithms provide analysis of the factors leading to decisions and present a possible forecasting tool. While modeling such problems, one should consider that, while RF and CART are classification algorithms, Ctree is a regression type solver. CART and Ctree present simplified trees, while RF has no means of representing the forest built by it. In terms of modeling different problems, it is important to tune the models and find the best-fit parameters to improve accuracy estimates. It was found that all the models had high classification accuracy to predict irrigation decisions when larger data sets were used. Smaller data sets supplied incomplete information to the models, resulting in poor classification rates.

Table 8 summarizes the probable important factors exhibited in the tree structures and variable importance measures. The predictors which are most useful in forecasting irrigation decisions are consumptive use, growing degree days or cumulative temperatures, and irrigating when a neighbor irrigates. The variable Year is specific to a perennial crop like alfalfa. Since ET is dependent on temperature, temperature and canal diversion measurements can be used to forecast farmers' future actions. The other important aspect in getting accurate forecasts is the amount of information given to the model. Information for the full growing season should be provided, which means that the models will not be able to handle missing information for this problem. This feature is similar to a farmer managing his farm who monitors day-to-day crop and soil conditions and makes decision accordingly. If he skips a few days in observing these conditions, he will not be able to make appropriate decisions due to the gap in information. We conclude that the most important factor for irrigation behavior appears to be crop need, followed by farmers' observations of their neighbors' actions. These findings are promising and can be used to make estimates of short-term demand forecasts.

References

Alfaro, E., Gámez, M., García, N., 2012. Adabag: Applies AdaBoost.M1, AdaBoost-SAMME and Bagging, R package version 3.1-52. URL: `http://cran.r-project.org/web/packages/adabag/index.html`.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration - guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper no. 56.

Becu, N., Sangkapitux, C., Neef, A., Kitchaicharoen, J., 2006. Participatory simulation sessions to support collective decision: the case of water allocation between a Thai and a Hmong village in northern Thailand. In: March 7-9, 2006, Chiang Mai, Thailand, International Symposium, Towards Sustainable Livelihoods and Ecosystems in Mountainous Regions.

Berger, B., Hansen, R., Hilton, A., 2002. Using the World-Wide-Web as a Support System to Enhance Water Management. In: Workshop Proceedings - Irrigation Advisory Services and Participatory Extension in Irrigation Management - Workshop organised by FAO - ICID, July 2002, Montreal, Canada.

Bontemps, C., Couture, S., 2002. Irrigation water demand for the decision maker. Environ. Dev. Econ. 7 (4), 643–657.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140. URL: `http://www.springerlink.com/index/10.1007/BF00058655`.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth International Group, Belmont, CA, USA.

Cutler, R.D., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. J. Safety Res. 40 (4), 317–327.

Hill, T., Lewicki, P., 2007. STATISTICS: Methods and Applications. StatSoft Inc., Tulsa, OK, USA.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2012. party: A Laboratory for Recursive Part(y)itioning. R package version 3.1-52. URL: `http://cran.r-project.org/web/packages/party/index.html`.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. J. Comput. Graph. Stat. 15 (3), 651–674.

Kastellec, J.P., 2010. The statistical analysis of judicial decisions and legal rules with classification trees. J. Empirical Legal Studies 7 (2), 202–230.

Le Bars, M., Attonaty, J.M., Pinson, S., Ferrand, N., 2005. An agent-based simulation testing the impact of water allocation on farmers' collective behaviors. Simulation 81 (3), 223–235.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22. URL: `http://cran.r-project.org/doc/Rnews/`.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing, R package version 3.1-52.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9 (1), 307.

Therneau, T.M., Atkinson, B., Ripley, B.D., 2012. rpart: recursive partitioning, R package version 3.1-52. URL: `http://cran.r-project.org/web/packages/rpart/index.html`.

Wright, J.L., 1982. New evapotranspiration crop coefficients. Proc. ASCE, J. Irr. Drain. Div. - ASCE 108 (IR2), 57–74.

CHAPTER 4
EXPLORING IRRIGATION DECISION BEHAVIOR USING HIDDEN MARKOV MODELS

Abstract

In an on-demand system, canal operators deliver water to the fields after receiving a water order from a farmer. These water orders are a result of a farmer's decision to irrigate. Irrigation decisions made by farmers are often ignored while modeling such problems, owing to their high variability and unpredictable nature. By analyzing farmers actions along with measurements of biophysical conditions, it might be possible to forecast short-term water demands for an irrigated area. Short-term demand estimates can eventually be a useful piece of information for the operators of the canal and reservoir system that serves irrigated lands. A hidden Markov model (HMM) was built to analyze irrigation decision behavior of farmers and make forecasts of their future decisions. The model inputs were each of the selected factors from cumulative evapotranspiration, depletion, soil stress coefficient, and canal flows. The irrigation decision series were the hidden states for the model. The paper evaluates data from the Canal B command area of the Lower Sevier River, near Delta, Utah. The main crops of the region are alfalfa, barley, and corn. A portion of the data was used to test the model capability to predict future irrigation events. It was found that the farmers cannot be classified into distinct classes based on their irrigation decisions, but vary in their behavior from irrigation to irrigation across all years and crop types. The variables selected as inputs to the model can be used to infer possible causal or indicative factors that influence a farmer's decision to irrigate. HMMs can predict an impending irrigation event when maximum likelihood (ML) estimates of model

parameters are known based on historical evidence. The study shows that HMM is a capable tool to study irrigation decision behavior which is not a memory-less process.

4.1 Introduction

Short-term irrigation demand forecasts could be useful information for the canal operators. Predictive information about farmer irrigation decisions could help in efficient management of the water diversions to the farmer fields.

Practically, little is known about the mechanics involved in irrigation decisions. Farmers decisions are varied, and presumably depend on factors such as weather, market prices, water remaining in their irrigation right for the season, the crop stress indicators, etc., but there is minimal understanding of how farmers use information about these factors to make irrigation decisions.

Irrigation behavior is further complicated by decision-making at the farmer level because every farmer is different in his approach towards growing crops. Some farmers may aim for good crop quality, while others might use water sparingly by using soil moisture measurements, yet others may irrigate as soon as they see some stress in the plants. Due to these characteristics there can be several farmer irrigation-decision paradigms, making the forecasting problem even more complex. Hence it is important to deal with these problems with suitable tools.

Since farmer irrigation decision behavior is probably not dependent on one factor, we can roughly define it as a multivariate process. There are some variables in the system which are indirectly affected by a farmers behavior. For example, the soil moisture content depletes because of evapotranspiration (ET), but it is the farmer who replenishes the soil reservoir. This makes a case for studying the variables in isolation to discover that important variable that can best predict a farmers behavior. This problem can be thought of as randomly observed data being dependent on some unobserved or hidden random states.

Markov models are used to model the states through which a system has passed to produce measured observations. Even simpler are the HMMs, which have been used to study speech recognition (Rabiner, 1989), and weather states (Zucchini and Guttorp, 1991; Hughes and Guttorp, 1994; Hughes et al., 1999). HMMs are first-order Markov models. From the behavioral perspective, there is some literature documenting about how HMMs can be applied to human subjects. Jeong et al. (2008) in their psychological study found patterns in students' learning activities while interacting with a computer, which can be a characteristic of their behavior.

Farmers' behavior has been simulated in some studies. Becu et al. (2006) built a multi-agent system to study water sharing between two villages located at the extreme ends of a watershed. The basis on which the farmers make irrigation decisions was studied. This included crop growing practices and harvesting and irrigation strategies used. Farmers groups were identified by studying their cropping patterns. Once the crop grown by the farmer was decided, the irrigation decisions were simulated accordingly. This study provided solutions to the villages to avoid water scarcity. Le Bars et al. (2005) also modeled a multi-agent system to simulate agent-farmers who made irrigation decisions under conditions of limited water supplies. Farmers had a water quota, from which they ordered in the beginning of the irrigation season depending on the crops grown and farm size. There was a water manager agent who managed the water using allocation rules. Random climate variables were assumed. Bontemps and Couture (2002) created a sequential decision model to simulate farmer decision making when they paid a minimal amount for ordering water. The water was supplied for free. These few studies do not analyze a farmer's behavior but try to recreate his actions under different scenarios. This work is a first attempt to study farmers' irrigation decision behavior using HMMs. We have tried to study biophysical variables that represent behavior to deduce information about those variables which are important from the decision-making perspective. We have used data from years 2007-2010 and have also tested the model to infer future decisions.

4.2 Hidden Markov Models and the Viterbi Algorithm

The very well-known first-order hidden Markov model (HMM) is specifically a simple probability model and is represented graphically as shown in Figure 16. If a simple system that is evolving over discrete time steps is described by observed variables X_t, which are related to an unobserved hidden state, S_t, then such a system follows a hidden Markov process (Rabiner, 1989). The parameters defining such a process are referred to as a hidden Markov model (HMM). Our problem, a "decoding" type problem of HMMs, is one of finding the most probable sequence of hidden states. The Viterbi algorithm is used for this purpose.

The Viterbi algorithm, initially given by Forney (1973), assumes an initial HMM for an observation sequence, and determines one single, "the most likely sequence" of underlying hidden states that might have generated the sequence. A HMM, represented as M= (A, B, π), is specified by the following probabilities (Rabiner, 1989):

1. A vector of initial state probabilities, $\pi = \pi_i$.

2. A matrix of transition probabilities, A=a_{ij}, where, a_{ij}=P($s_i|s_j$) and P($s_i|s_j$) is the conditional distribution of the present state s_i, given the previous state, s_j.

3. A matrix of emission/observation probabilities, B= $b_i(v_m)$, where, $b_i(v_m)$ = P($v_m|s_i$) and P($v_m|s_i$) is the conditional distribution of v_m given the hidden state, s_i.

The observation sequence, O= $o_1o_2......o_k$ is given. We have to find the state sequence, Q=$q_1......q_k$, which maximizes P($Q|o_1o_2......o_k$). The maximum probability is:

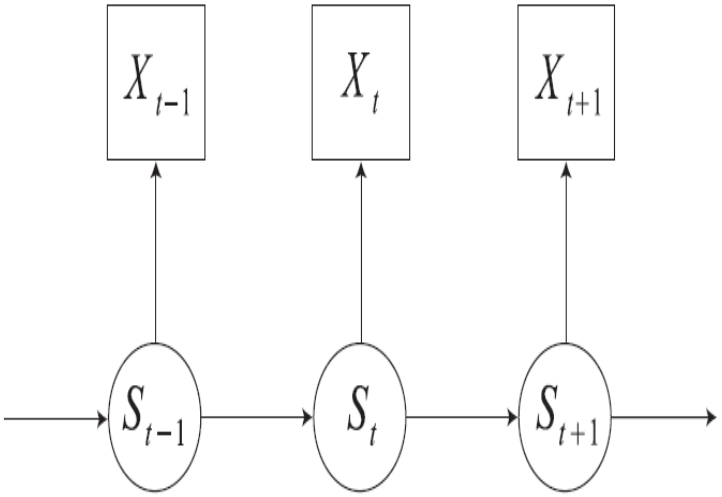


Fig. 16: Graphical representation of a hidden Markov process where X is the observed variable and S is the unobserved hidden state.

$\delta_k(i) = max(P(q_1......q_{k-1}, q_k = s_i, o_1o_2......o_k))$

and produces observation sequence $o_1o_2o_3......o_k$ while walking through any hidden state sequence $q_1......q_{k-1}$ and getting into $q_k = s_i$. In other words, if the best path ending in the present state, $q_k = s_j$ passes through the previous state, $q_{k-1} = s_i$, then it should coincide with best path ending in the previous state $q_{k-1} = s_i$.

The procedure for finding the best state sequence is as follows (Rabiner, 1989):

1. Initialization:

$\delta_1(i) = max(P(q_1 = s_i, o_1)) = \pi_i b_i(o_1), 1 \leq i \leq N$

To retrieve the state sequence we need to keep a track of the argument which maximised $\delta_k(i)$ for each 'k' and 'i'. We do this by using array $\psi_k(i)$.

$\psi_1(i) = 0$

2. Forward Recursion:

$$\delta_k(j) = max(P(q_1......q_{k-1}, q_k = s_j, o_1 o_2......o_k)) =$$

$$max[a_{ij} b_j(o_k) max P(q_1......q_{k-1} = s_i, o_1 o_2......o_{k-1})]$$

$$= \max_{1 \leq i \leq N} [a_{ij} b_j(o_k) \delta_{k-1}(i)], 1 \leq j \leq N, 2 \leq k \leq K$$

$$\psi_t(i) = \arg \max_{1 \leq i \leq N} [\delta_{k-1}(i) a_{ij}], 1 \leq j \leq N, 2 \leq k \leq K$$

3. Termination: choose best path ending at time K

$$P^* = \max_{1 \leq i \leq N} [\delta_K(i)]$$

$$q_K^* = \arg \max_{1 \leq i \leq N} [\delta_K(i)]$$

4. State sequence or path backtracking:

$$q_k^* = \psi_{t+1}(q_{k+1}^*), k = K-1, K-2, ..., 1$$

Commonly, the observed variable of interest is known, and the hidden state distribution is unobserved, but is approximated using distributions. The model parameters are first estimated with the hidden states the system went through for all the time steps for which the variable was observed. The model is then tested using new cases from the system and hidden states are predicted at every step.

In the case of our problem, the observed variables will be the factors the farmer might have considered while deciding whether to irrigate, and the hidden states will be indicative of his decision on a given day. The decision has been treated as unobserved so as to check if the model is able to extract a similar decision path from the observed variable. Instead of assuming distributions to simulate the irrigation decision, we used the actual observed decision to estimate the HMM parameters. We will try to find how the model can classify the variables into two different states and if those states coincide with a given farmer's irrigation pattern. From the context of farmer irrigation decision behavior, this approach is new. This study tries to find those variables which are the informational basis of his decisions.

4.3 Case Study

The study area selected for this work is the Canal B command area, a 20 square mile region in south-central Utah near the community of Delta, in the Lower Sevier River Basin. Irrigation is the largest user of the water in this basin, and surface irrigation is the prevalent method in the region. Weather data for the area are available on: http://www.cemp.dri.edu/cgi-bin/cemp_stations.pl?stn=delu. This station at Delta was established by National Climatic Data Center (NCDC), NOAA and has historical weather data since 1965. The station metadata are as follows and can be located on the NCDC-NOAA website (http://www.ncdc.noaa.gov/) using:

GHCND ID : USW00023162

COOP ID : 422090

WMO ID : 72479

NCDC ID : 20026236

Kimberly Penman Reference ET procedures have been used to estimate evapotranspiration rates, which are also available on this website. These ET calculations were verified for accuracy. Soil moisture probe data were used the determine the day of irrigation. These data are available at `http://odm.usu.edu/odmmap/default.aspx?NetworkName=Delta`. Forty-four stations with 88 sensors were established by the Utah Water Research Laboratory (UWRL), Utah State University (USU) in 2007 to record soil moisture at 1 and 2 ft depths on various farms in Delta, Utah to study agricultural water use. The sensors are maintained by the personnel at the UWRL.

Daily soil moisture depletion was computed. Most of the farmers irrigated up to saturation, hence depletion values on those days were negative (Note: soil moisture content at field capacity translates to a depletion of 0). For this the soil stress was also computed. The soil stress coefficient (Ks) indicates whether the soil is water stressed or not, where 1 indicates no stress, and 0 indicates high stress because of consumption of all of the plant available water. The planting dates for alfalfa, barley, and corn were estimated by running the soil moisture calculations from random dates before the first irrigation such that the day of first irrigation was matched. Calculations were started from soil water content at field capacity. Silty clay loam, silty clay, and loam were the three predominant soil classes in the region. Standard values for porosity, field capacity, and wilting point were used from Allen et al. (1998). Crop coefficients for alfalfa, barley, and corn were estimated from Wright (1982) and FAO-56 (Allen et al., 1998). Crop ET values (CumET$_c$) were accumulated between the irrigations. SCADA (Supervisory Control and Data Acquisition) systems maintain the past records of water discharge across Sevier River Basin (Berger et al., 2002). The site is sponsored by Sevier River Water Users Association (WUA). Real-time daily canal flows in cubic feet per second (cfs) were obtained for Canal B from: `http://www.sevierriver.org/rivers/delta/b-canal/`.

4.4 Model Development

For model development we used the "hsmm" package (Bulla et al., 2010) developed in R (R Development Core Team, 2007). The plant-soil-irrigation phenomenon is not memory-less; irrigations are related to antecedent conditions that occur over several days prior to the day of irrigation. Hence we have used a hidden semi-Markov model to include the time dependency of events up to the day of irrigation. Equal initial, transition, and output probabilities were assumed. Default observation distribution, a Gaussian distribution (od="norm"), and the run length, a logarithmic distribution (rd="log"), were used. The observation sequence was one of the following variables: K_s, Depletion, CumET$_c$, or Canal Flow. CumET$_c$ was assumed to have two states (high or low), and all the others had four states (low, medium, high, critical). The limits for all the variables used to create various states are presented in Table 9. State numbering has been used to indicate a different state, although it has no significance. The state ordering can be different in presentation. It should just indicate levels, such as high, medium, and low, if there are three states. CumET$_c$ was discretized in two ways for alfalfa because the days to irrigation became finer with a different interval for the variable. In spite of proper initialization, not all of the variables worked for different data sets. The path was a series of hidden states: Irrigate (1)/No irrigation (0) decisions. The data were analyzed crop-wise for:

a. all days from the growing season from various fields referred to as all-days HMM.

b. a series of four days before irrigation and the day of irrigation from all fields. To explain this, if the day of irrigation was day "t", then observations and path for days $t-4, t-3, t-2, t-1$, were collected for all irrigation events on all fields. These are addressed as four-days HMM.

The model parameters were learned and then applied to the test data. During initial model set up we tried to train the model with half, two-thirds and three-fourths of the data. It was found that only when training with three-fourth data we see some changes in the transition matrix, that showed successful training. Therefore we chose to use one-fourth of the data for testing.

Table 9: Limits assumed to discretize various observed variables. These limits were developed after various trials with different discretization of variables.

Crop	Factor	State 1	State 2	State 3	State 4
All	Ks (unitless, 0-1)	< 0.3	< 0.4	< 0.5	< 1
	Canal Flow (cfs)	<75	< 150	< 200	< 250
Alfalfa	CumET$_c$1 (mm)	< 100	< 575	-NA-	-NA-
	CumET$_c$2 (mm)	< 250	< 575	-NA-	-NA-
	Depletion (mm)	< 0	< 75	< 150	< 195
Barley	CumET$_c$(mm)	< 405	< 100	-NA-	-NA-
	Depletion (mm)	< 0	< 75	< 125	< 255
Corn	CumET$_c$ (mm)	< 100	< 334	-NA-	-NA-
	Depletion (mm)	< 0	< 75	< 150	< 185

4.5 Results and Discussion

Four days before the day of irrigation analysis (referred to as a four-days HMM) was done with the objective of discovering general patterns in the irrigation decision behavior. The all days model (referred to as the all-days HMM) was used to verify what we found with the four-days model. The results for all those variables have been presented, which were interesting and reflected behavior. Since the HMMs deal with states, it is up to the user to interpret the states. This can be done if something is known about the process. The prediction results of the unobserved states (irrigation decision) from the given observations have been presented below.

The vertical axis of the plots presented here represents the chosen input states for the factor, while the colored states on a time series represent the model output states. We have chosen to use input states on the vertical axis to show the changes made after training. The observed variables have been presented as a scaled factor value to represent the state to which they belong. Hence, the model outputs have been plotted on real-scaled time series of the variable to show the variable limits modeled by HMMs.

4.5.1 Alfalfa

For alfalfa, it was found that all the farmers irrigating below a value 0.275 for K_c (state 2), as shown in Figure 17, were grouped into one class. This group accepted much risk by irrigating up to the maximum limit of plant available water by stressing the crop. The other group even irrigated when there was hardly any soil stress. Referring to Table 9, it is very clear that the model ignored states 2 and 3, since it did not find instances of irrigation in them.

Figure 18 presents results for the depletion factor for alfalfa, the first group (state 3) was irrigating when the depletion (134 mm) reached close to total available water (150 mm, refer to Table 9). These farmers stressed the crop. Farmers in state 2 irrigated when the depletion (63.2 mm) was half way between field capacity and wilting point. Figure 19(a) and 19(b) shows some farmers behaving the same way as the four-days HMM interprets them. These two instances of farmers did not take risk and irrigated as soon as the readily available water was depleted, without risking the crop quality.

We could not get meaningful results for CumET_c factor for four day HMMs, but the all-days model worked as shown in Figure 20(a). It can be seen that the farmer behavior is different from irrigation to irrigation. State 1 is up to 100 mm accumulated crop ET (CumET_c1 in Table 9). The third irrigation was earlier in comparison to the fourth irrigation, hence lesser accumulated ET. Figure 20(a) shows that HMM does a good job in predicting impending irrigations. Figure 20(b) shows another example of farmer behavior, where the crop was irrigated seven times in the season. The first,

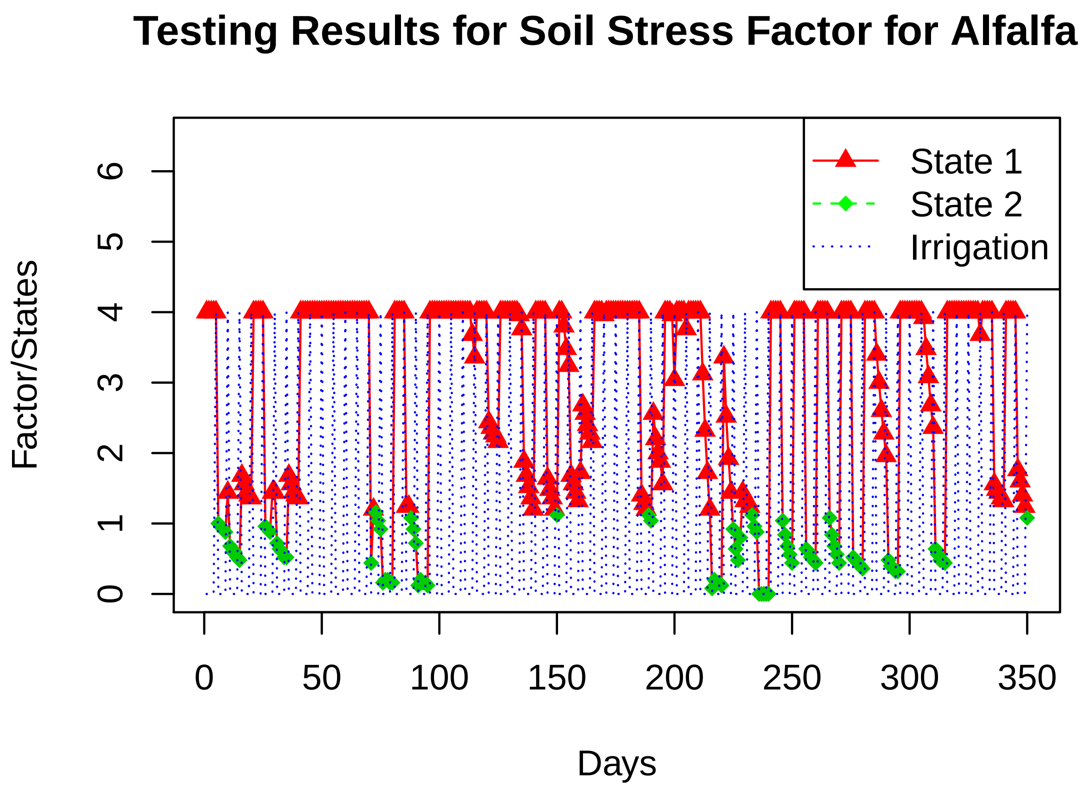


Fig. 17: Four-days HMM results for Alfalfa with Soil stress factor as an indicator for irrigation behavior.

fifth, sixth and last irrigations were probably given keeping CumET$_c$ in mind, but there was probably another factor for other irrigations.

Figure 21(a) shows another instance of a farmer, who used crop indicators to irrigate. Figure 21(b) gives us a reason to study other factors, since crop stress might not be the only reason for irrigation. Figure 22 shows a different discretization of states for CumET$_c$ (CumET$_c$2 in Table 9). State 2 is starts at 260 mm of cumulated crop ET, which might be a critical stage w.r.t crop demand.

The use of a better discretization enabled the model to predict the irrigation close to the actual event. This means that the thresholds for various farmers could be different.

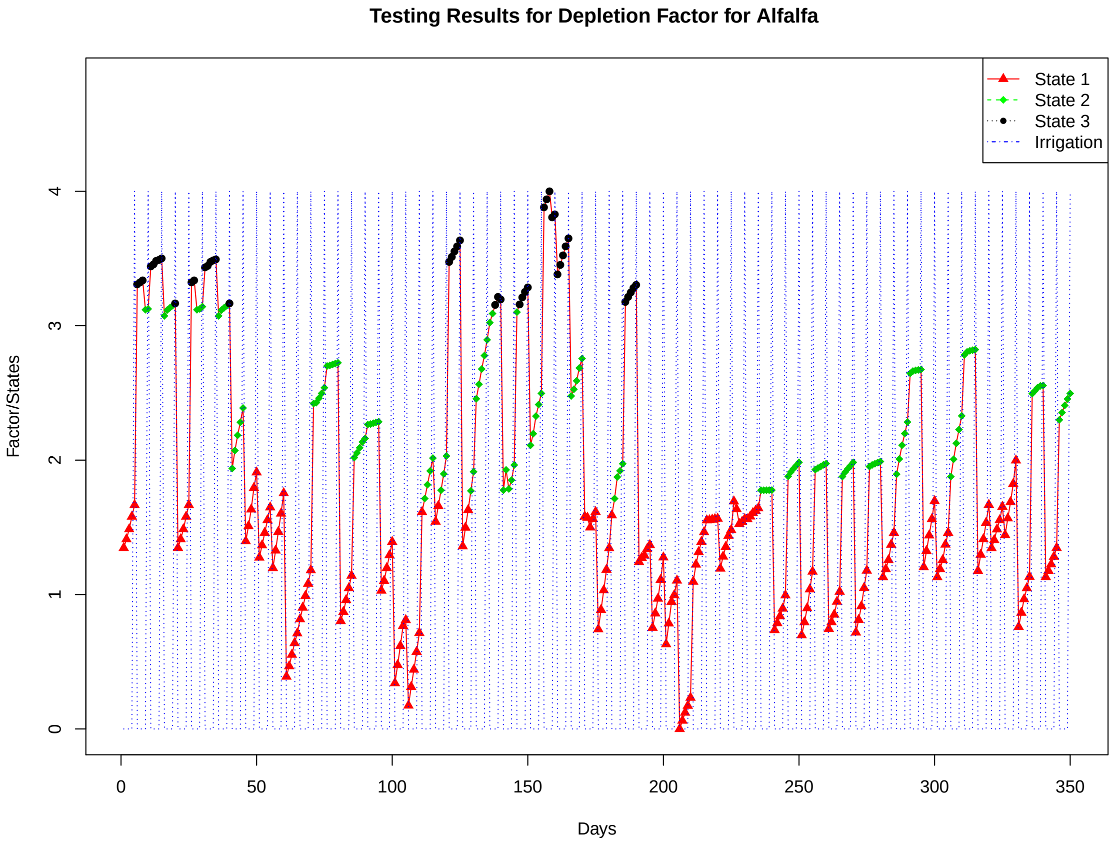
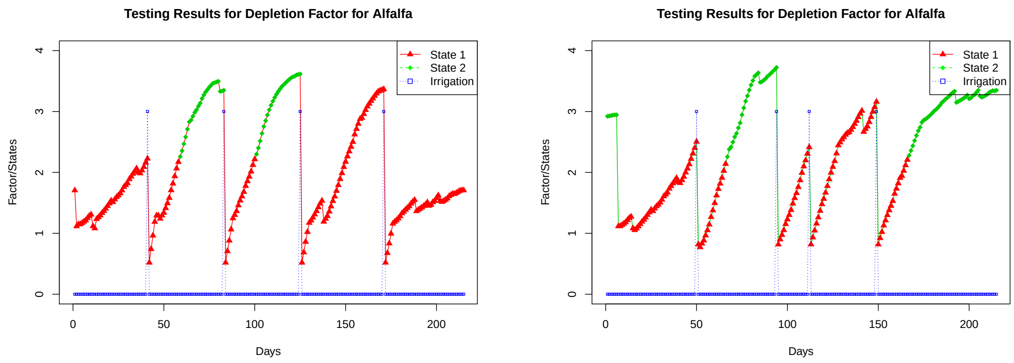


Fig. 18: Four-days HMM results for Alfalfa with depletion factor as an indicator for irrigation behavior.

4.5.2 Barley

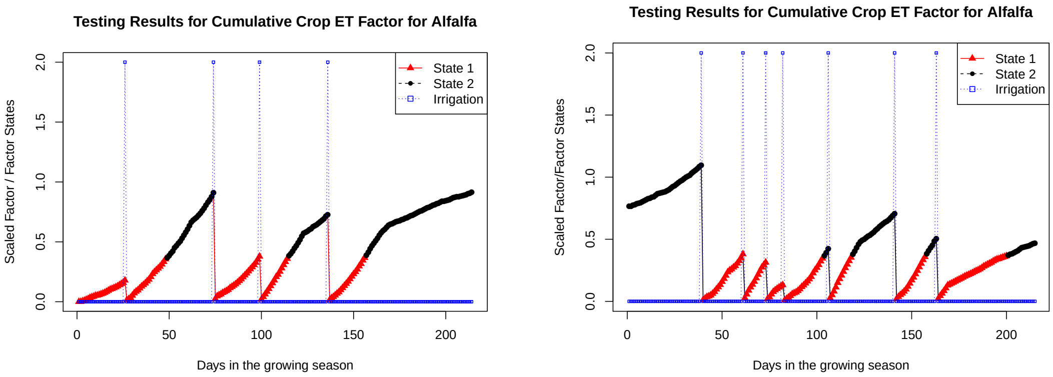
For barley, K_c categorized irrigations into two classes (Figure 23(a)). The critical value for K_c was 0.225 for which three irrigations were found. Probably the farmer did not have water for the crop or it was the last irrigation. The other irrigations were done when there was hardly any stress. Figure 23(b) shows the critical $CumET_c$ was below 94 mm categorized as State 2. It was found that the farmers irrigated at this level during the first or second irrigation.

Figure 24 shows results for canal flow factor to discriminate farmers. Only four irrigations were done when canal flow was in State 3 (above 227 cfs), the high flow. This means that the farmers during these irrigations irrigated with their neighbors.



(a) (b)

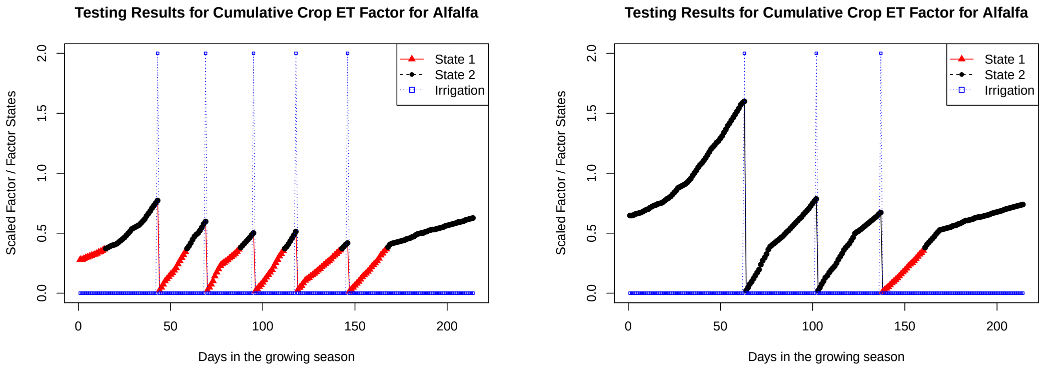
Fig. 19: 19(a) and 19(b) Farmer-wise all-days HMM results for Alfalfa with depletion factor as an indicator for irrigation behavior.



(a) (b)

Fig. 20: 20(a) and 20(b)Farmer-wise all-days HMM results for Alfalfa with CumET$_c$ factor as an indicator for irrigation behavior.

When canal flow was in State 2 (above 150 cfs), the farmers did not irrigate. This state usually occurred after the irrigation and was of no significance. Hence, corn growers might not worry about irrigating with their neighbors.



(a)

(b)

Fig. 21: 21(a) and 21(b) Farmer-wise all-days HMM results for Alfalfa with CumET$_c$ factor as an indicator for irrigation behavior.

4.5.3 Corn

K$_c$ for corn was quite an evident factor, influencing farmers behavior. Since the farmers have to maintain humidity levels for corn heading stage, corn growers in Delta do not take any risk as is evident from Figure 25. The K$_c$ level for state 2 is 0.125 and below. Again the assumed lowest limit for this factor was 0.3 (Table 9). This could have happened in the early part of the season.

Cumulative Crop ET was also a determinant for farmers behavior. Above 97 mm was categorized by the HMM as State 2 (Figure 26(a)). Figure 26(b) shows the canal flow factor for forecasting farmers' behavior. It classifies behavior into three groups. The farmers do not consider when their neighbors are irrigating since the crop need for corn appears to be the prime driving force. Depletion (Figure 27) seems to be an important factor. The farmers are grouped into three classes. For this particular instance, the thresholds for state 1 were 0-179 mm, state 2 were 179-234 mm and state 3 was above 234 mm. As shown in Figure 28(a), Figure 28(b), there were farmers who did not wait as long as to stress the crop and hence, there were just two states found, in some cases.

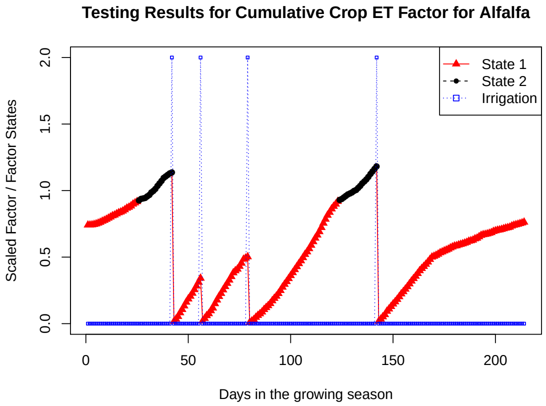
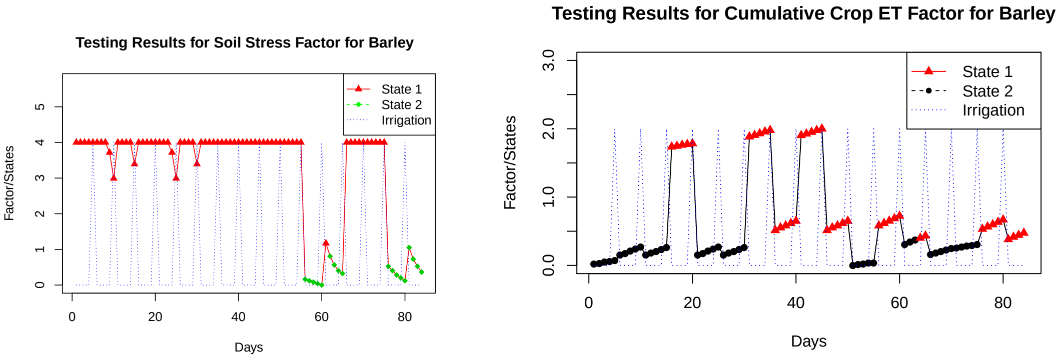


Fig. 22: Four-days HMM results for Alfalfa with CumET$_c$ factor as an indicator for irrigation behavior.

4.5.4 Importance of the ML-Fitted Transition Probabilities

We mentioned in Section 4, Model Development, that four states were fitted with the model for K$_c$ and Depletion factors. But, HMMs narrowed the states to either two or three. Hence, it is vital to know how and why three or four states were narrowed to two states sequence. This can be answered if we look at the ML-fitted transition probabilities for these variables.

Table 10 and Table 11 have been fitted using data representing alfalfa irrigation practices, using the four-day HMM. There were four input states for soil stress. But the transition matrix in Table 10 shows that there will be two output states. This might happen because the equally probable, or close to equally probable, states are negated and the model pick only one of them to calculate path sequence. If we look at the first row of the matrix, transition from state 1 to 2 and 4 is equally probable (0.342), and so is the transition from state 1 to 1 and 3 (0.149 and 0.168, respectively). The latter are close in the probabilities. Similarly, row 3 has equal probabilities of 0.377 and near-equal ones of 0.117 and 0.130. The second and fourth rows clearly have three almost equal, 0.286, 0.259, and 0.286, and one unique, 0.168, probabilities for transition from one state to another.



(a)

(b)

Fig. 23: 23(a) Four-days HMM results for Barley with soil stress factor as an indicator for irrigation behavior, 23(b) Four-days HMM results for Barley with CumET$_c$ as an indicator for irrigation behavior.

Similarly for Depletion, four input states assumed, but the model found only three states. Referring to Table 11, we can again find the same thing, such as in row one, the transition from state 1 to 1, and 1 to 3 are unique probabilities (0.309 and 0.0960, respectively), but in going from 1 to 2 and 4 the probabilities are of a similar measure. Hence during testing we see only 3 states.

Table 10: Fitted transition probability matrix for Soil Stress Factor.

	[,1]	**[,2]**	**[,3]**	**[,4]**
[1,]	0.149	0.342	0.168	0.342
[2,]	0.168	0.286	0.259	0.286
[3,]	0.117	0.377	0.13	0.377
[4,]	0.168	0.286	0.259	0.286

Table 11: Fitted transition probability matrix for Depletion Factor.

	[,1]	**[,2]**	**[,3]**	**[,4]**
[1,]	0.31	0.297	0.097	0.297
[2,]	0.039	0.391	0.179	0.391
[3,]	0.325	0.248	0.179	0.248
[4,]	0.039	0.391	0.179	0.391

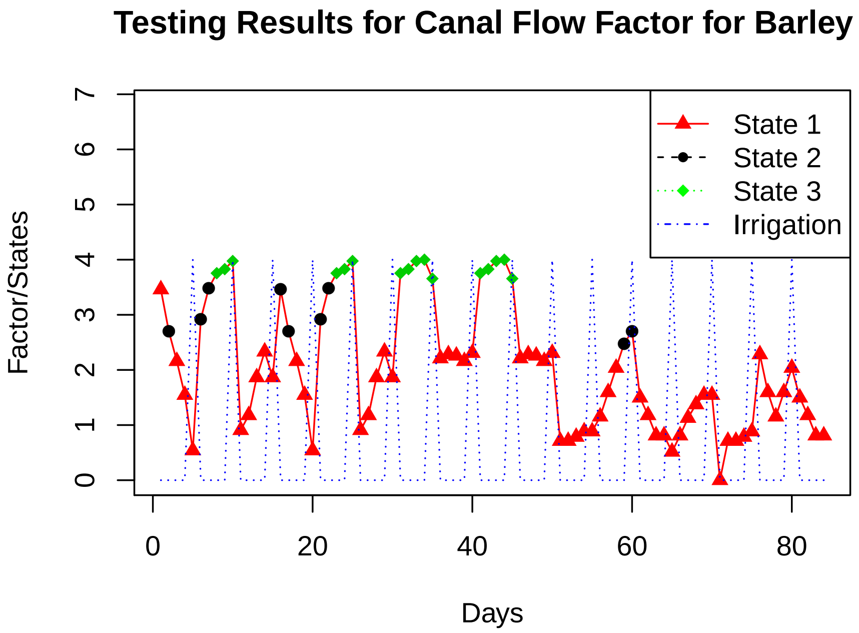


Fig. 24: Four-days HMM results for Barley with Canal Flow factor as an indicator for irrigation behavior.

Table 12: The factors and their levels that might have been used by majority of the farmers at Canal B, Delta as a criteria for making irrigation decisions. Note: These results are based on four-days HMM unless otherwise explicitly mentioned as all-days HMMs.

Factor \ Crop	**Alfalfa**	**Barley**	**Corn**
Soil moisture stress	Low	Low	Low
Depletion	Low and medium	Not known	Medium (all-days)
CumET$_c$	High (all-days)	Low and high	High
Canal Flow	Not known	Low	Low

4.6 Conclusions

Forecasts of short-term irrigation demands can provide important information to

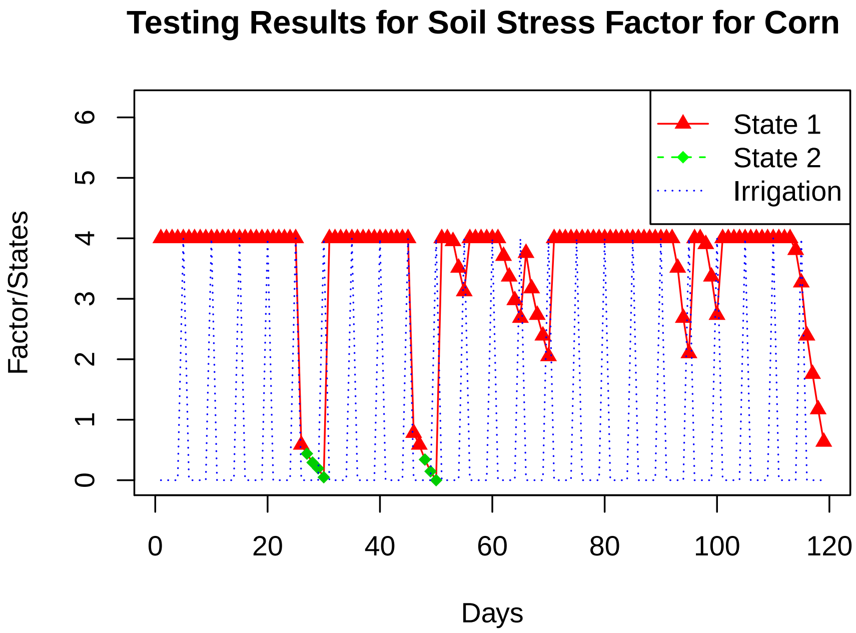
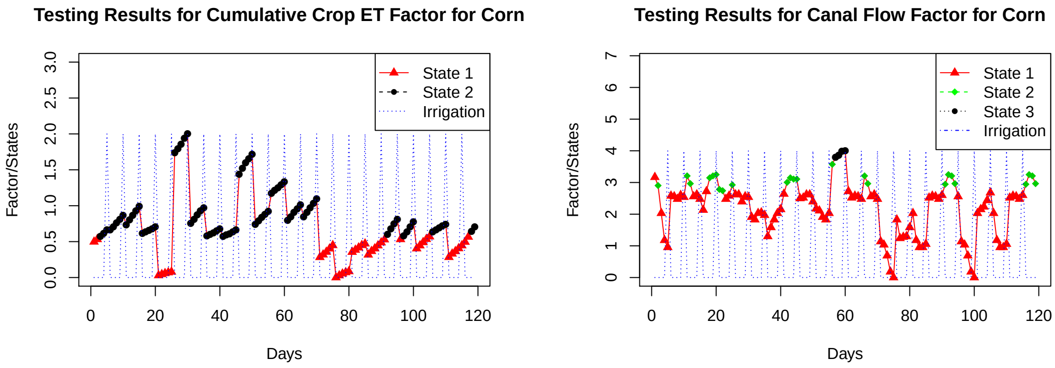


Fig. 25: Four-days HMM results for Corn with Soil stress factor as an indicator for irrigation behavior.

irrigation water managers. These demands can be estimated by analyzing irrigation decisions. This study does not examine the management aspect of demand but deals with the short-term forecasting of it. In Canal B, the upper bound on demand is the system capacity. This constraint is well understood by the water masters of the canal company. The research presented here applied HMMs to study the variables which can be potentially used to forecast farmers irrigation decisions. These variables can be helpful to provide answers about why farmers irrigate on certain days and not others. Sometimes the cumulative effect of the variables can be seen to contribute to the decision, but since we are analyzing the effect of singular variables we limit the analysis to that. This study fractionated the variables and studied the variables alone to explore their individual effect on the decision.



(a)

(b)

Fig. 26: 26(a) Four-days HMM results for Corn with CumET$_c$ factor as an indicator for irrigation behavior, 26(b) Four-days HMM results for Corn with Canal flow factor as an indicator for irrigation behavior.

Irrigation behavior was treated as a Markov process, where the soil-plant-farmer-weather interaction is not memory-less. Though we could not get exact predictions of irrigation events, the model did a good job to capture impending irrigations. Also Rabiner (1989) clearly mentions that "only for degenerate models a "correct" state sequence can be found". In other words it finds only the optimal state sequence and not the exact sequence. HMMs can easily do away with non-occurring or infrequent states. They are not just driven by the discretization given by the user for a variable, but can change discretization for a variable if they find a better path. Given the state of the observed variable, they can infer the (hidden) state of the irrigation decision and adequately describe the process. Their drawbacks are that they cannot predict the exact day of irrigation, and do not indicate the important observations (or their states) with respect to predicting hidden states. Barring these shortcomings, all in all HMMs show a great potential to provide insights into this complex process.

It can be seen from the analysis that instead of farmer classes, there are groups that arise on an irrigation-to-irrigation basis. This indicates that farmers adapt their

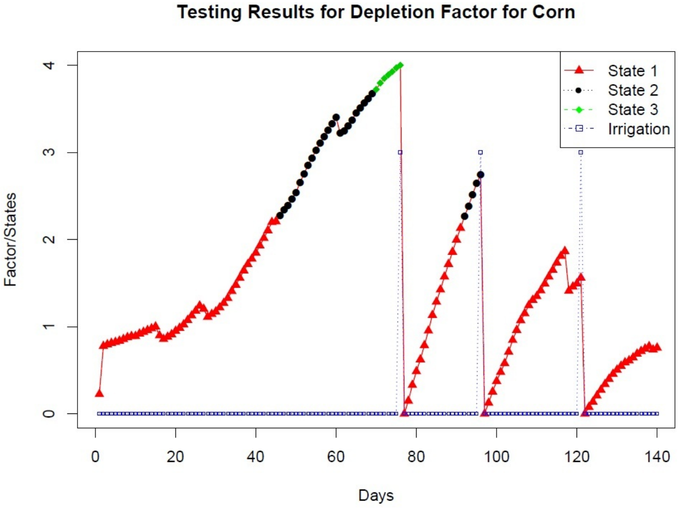
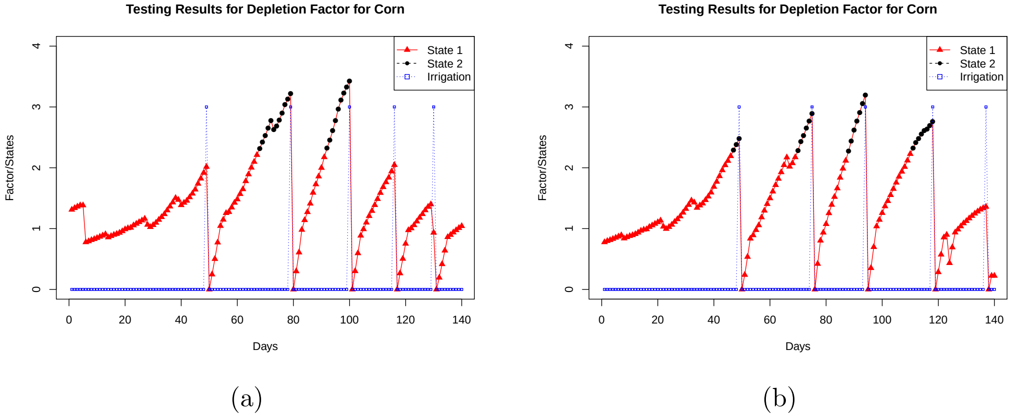


Fig. 27: Farmer-wise all-days HMM results for Corn with depletion factor as an indicator for irrigation behavior.



(a)

(b)

Fig. 28: 28(a) and 28(b) Farmer-wise all-days HMM results for Corn with depletion factor as an indicator for irrigation behavior.

irrigation practices to every irrigation, according to ambient conditions. All the variables had some information about why the farmer irrigated or did not irrigate.

4.7 Behavior

Farmers are usually different from each other in their understanding crop processes, in keeping with irrigation traditions, and in financial status. It is understandable that they are flexible and cater to plant and weather conditions and vary in their decisions on an irrigation-by-irrigation basis. This is an important finding where irrigation decision behavior is concerned. Some farmers could also be deciding irrigations by keeping rough estimates of levels of variables like crop water use. Table 12 shows the criteria that might have been used by the farmers at the site for deciding irrigation timings. As can be seen, the farmers were not taking risk in terms of stressing their crops. Depletion levels also indicated that the farmers believed in using the water as soon as they saw some evident stress in the crop. Cumulative crop ET was mostly high when decision was made, implying that the farmers might be keeping track of the water used by the crop by some means. The canal flow or the irrigations made by observing their neighbors was not important. For alfalfa, the results were inconclusive for canal flow as an indicator of irrigation because the irrigation of the crop is timed according to the cuts. Overall it can be easily concluded that farmers might prefer readily available factors like crop condition and crop water use during the season, than factors for which information is difficult to obtain or that has an uncertain future importance, such as market prices. Biophysical indicators can be successfully used to represent the crop-water-soil conditions observed by the farmer during the growing season and can be used to forecast irrigations. The irrigations of all the farmers summed at the command area level can give us estimates of short-term irrigation demands. This information can help the canal and reservoir operators in managing the resources efficiently.

References

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration - guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper no. 56.

Becu, N., Sangkapitux, C., Neef, A., Kitchaicharoen, J., 2006. Participatory simulation sessions to support collective decision: the case of water allocation between a Thai and a Hmong village in northern Thailand. In: March 7-9, 2006, Chiang Mai, Thailand, International Symposium, Towards Sustainable Livelihoods and Ecosystems in Mountainous Regions.

Berger, B., Hansen, R., Hilton, A., 2002. Using the World-Wide-Web as a Support System to Enhance Water Management. In: Workshop Proceedings - Irrigation Advisory Services and Participatory Extension in Irrigation Management - Workshop organised by FAO - ICID, July 2002, Montreal, Canada.

Bontemps, C., Couture, S., 2002. Irrigation water demand for the decision maker. Environ. Dev. Econ. 7 (4), 643–657.

Bulla, J., Bulla, I., Nenadić, O., 2010. hsmm - An R package for analyzing hidden semi-Markov models. Comput. Stat. Data An. 54 (3), 611–619.

Forney, G.D., 1973. The Viterbi algorithm. Proc. IEEE 61 (3), 268–278.

Hughes, J.P., Guttorp, P., 1994. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resour. Res. 30 (5), 1535–1546.

Hughes, J.P., Guttorp, P., Charles, S.P., 1999. A non-homogeneous hidden Markov model for precipitation occurrence. J. Roy. Stat. Soc. C-App. 48 (1), 15–30.

Jeong, H., Gupta, A., Roscoe, R., Wagster, J., Biswas, G., Schwartz, D.L., 2008. Using Hidden Markov Models to Characterize Student Behaviors in Learning-by-Teaching Environments. In: 9th International Conference on Intelligent Tutoring Systems, B.P. Woolf, E. Aimeur, R. Nkambou and S. Lajoie (Eds.), Montreal, Canada, pp. 614-625.

Le Bars, M., Attonaty, J.M., Pinson, S., Ferrand, N., 2005. An agent-based simulation testing the impact of water allocation on farmers' collective behaviors. Simulation 81 (3), 223–235.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing, R package version 3.1-52.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Wright, J.L., 1982. New evapotranspiration crop coefficients. Proc. ASCE, J. Irr. Drain. Div. - ASCE 108 (IR2), 57–74.

Zucchini, W., Guttorp, P., 1991. A Hidden Markov Model for Space-Time Precipitation. Water Resour. Res. 27 (8), 1917–1923.

CHAPTER 5
SUMMARY AND CONCLUSIONS

The agricultural soil-water-plant environment contains interactive processes such that farmers decide to allocate each resource to achieve optimal quality of produce. The mechanisms by which farmers make decisions to allocate scarce inputs to agricultural production, especially irrigation water, have not been extensively researched from the standpoint of the informational content of biophysical data that could be used to forecast short-term irrigation decisions. This makes strong ground for this work.

The approaches used in this study to analyze and forecast irrigation decisions are (1) learning Bayesian belief networks, (2) decision trees, and (3) hidden Markov models. These algorithms have been used in different applications, but generally have not been verified using data and most certainly have not been used to test their capabilities to forecast future decisions. The advantage of using these approaches is that they can work with limited as well as missing data and can make use of various information sources, such as expert judgment and categorical inputs. They are beneficial for applications where the relationship between observed variables and the target variable is indirect and usually unknown.

Data from a 20-square-mile region of irrigated agricultural land served by the Canal B irrigation system, near Delta, Utah, in the Lower Sevier River Basin, located in South-Central Utah, have been used to evaluate the capabilities of the modeling approaches. The data set, from 2007-2010, is comprised of weather variables, canal flow rates, market prices for alfalfa, barley, and corn, and soil moisture probe data. The days on which irrigation occurred have been extracted from the probes as well.

Chapter 2 presents a Bayesian belief network (BBN), which has been developed using as inputs the factors thought to be pertinent to the irrigation decision. The data quality is important to get reasonable forecasts of the day of irrigation. Classical soil moisture balance computations have been used to create the network relationships. Information available for the day when the farmers took the irrigation decision, together with that from a day before it, were collected for all irrigation events and used to train the model. The model learns from the data, which are either raw data or computed variables. The accuracy of the networks show that they are capable of handling data sets indicating the process of decision-making. They perform fairly in the testing phase and can benefit from extensive information about the process.

Chapter 3 discusses two tree algorithms, for classifying irrigation decisions and a third, a regression approach. There were no obvious clusters in data for this application. The models were first tried out with a few very important variables, but the models failed to find groups, hence we thought of adding derived variables which were a part of the process. The overall results showed that the models can handle parts of the irrigation data as well as the data from the full growing season with a very few irrigation events. But if the data for the whole season is available, they can do a good job of predicting future decisions.

Chapter 4 describes a hidden Markov model, commonly used to study behavior. It models problems where a variable is observed or quantified, but the state sequence the system went through to generate those observations is unknown. The analysis recovers this sequence of states from the observed data. For the irrigation behavior, the problem is to find that variable which is best representative of the irrigation decision sequence. This would translate in to the fact that the farmer might have used that information to take irrigation decision. The models showed capabilities of predicting the possible days of irrigation and could be used as a possible tool to forecast decisions.

The first objective of the study was to infer the variables which farmers take into consideration when making irrigation decisions. The BBN results showed a predominant factor, soil stress, as being the key variable for deciding to irrigate for many farmers across various crops. For alfalfa, the soil stress “on the fields” was caused probably because the crop is cut and left to dry out. This process can be long unless the farmers have access to equipment to speed up drying. The soil moisture probes would record soil moisture though there might have been no crop on the ground. For all the other crops, the farmers were taking a risk by utilizing soil water to its limit. The fact that they realized within the right time that their crop was going to become stressed shows good management skills. Irrigating when their neighbors irrigate and irrigating over the weekends were the next most important reasons according to BBNs. Before deciding to irrigate, some farmers might consult their neighbors or friends who have the same crops and similar amounts of water for a similar growing season conditions. Most farmers have an active profession keeping them occupied during the work week, so they might opt to irrigate on a weekend instead. Some irrigations occurred when the crop was not stressed, which showed that some farmers were averse to taking risks with their crops. According to tree-classifiers, critical accumulated crop ET was possibly one of the prime reasons for irrigation, with different limits for the three crops. The cumulative ET or a variant of it, the accumulated heat units, are often used by farmers as a measure to decide a stage sensitive to water stress or pest attack. Since trees were given continuous variables, they could pick out the critical limits when the farmers decided to irrigate. Irrigating simultaneously with a neighbor, as one of the reasons was supported by trees. Soil stress indicators, depletion and cumulative ET_c, were found to be at low levels when a farmer decided to irrigate. This indicates that crop indicators are not the only reasons for the farmers to irrigate. HMMs made a revelation by showing that farmers have different reasons with each subsequent irrigation and are quite flexible with various growth stages of the crop. For all three crops, the soil stress coefficient was found to be between low to medium when the farmers irrigated. Only a few farmers reached the lowest limit of soil water, the total available water, before irrigating. Cumulative crop ET was probably one of the factors considered by the farmers in making decision. Clearly, according to HMMs, farmers did not often care about irrigating with their neighbors.

The second objective was to classify irrigation decisions and discern the various types of farmers. HMMs showed that the farmer decisions were generally variable. For example, with the soil stress coefficient, depletion, and cumulative crop ET as indicators, farmers generally opted to stay close to fulfilling crop ET requirements and did not follow irrigation decisions that would impart stress to their crops. This means most of them do not take risks in their irrigation decisions. However, since some of the irrigations were done under severe stress, some farmers might accept risk in order to save water for a critical stage later in the season, or possibly were out of water. Maintaining soil moisture that that crops felt no water stress was found to be consistent with irrigations in the early part of the season, when the crop was in the germination or emergence stages, and less so with later growth stages. Only few of the farmers were classified as following their neighbors, which meant that most farmers had their own strategies to irrigate. However, trees found that the farmers could be classified by the irrigation factors they consider, with a maximum of three classes found for corn, and two each for alfalfa and barley. These classes were formed on the basis of crop water use and its derived variables, which makes it an important factor.

The third objective was to distinguish decisions reflecting the influence of a single-factor versus multiple factors. The BBNs and tree, classified farmers based on many variables, but HMMs classified types of irrigators on the basis of a single factor. HMMs showed that the factors considered might not have been the reason for certain irrigations. For example, when the soil moisture did not show stress, the farmers irrigated, meaning there were other factors influencing the decision to irrigate which were not obvious.

The fourth objective involved identification of patterns in decisions from irrigation-to-irrigation, crop-to-crop, and year-to-year. The analyses using HMMs showed that there were various factors for the same farmer over irrigations that may have lead to irrigation decision. There was no particular difference found by BBNs for irrigations given to various crops or in different years. For trees and HMMs, the year-wise data was too limited to determine a classification. The crop-wise analysis has been discussed previously.

All the algorithms were tested with new data, and their ability to identify clusters in the data based on what they learned in the training phase was presented in each analysis. The model built by learning from previous data was used to make predictions. The trees can be successfully used to predict future irrigation events, while BBNs were satisfactory in their prediction capabilities. As with HMMs, they can forecast the period of imminent irrigation quite well.

5.1 Research Challenges

This research encountered challenges that are worth examining. As has been mentioned, as many variables as possible that are thought to be contributing to the decision process should be used for learning. This improves the capability of the models to better understand the dynamics of the irrigation system.

It was found that it is good to have a variable represented in various ways. The range of a variable is important for the evolutionary algorithms used in this study. We used two such variables to convey information about the same factor, soil moisture, in the models: the soil stress coefficient and depletion. The soil stress coefficient is between 0 and 1, and depletion is between, saturation (negative depletion) or field capacity (zero depletion), and total available water. Total available water has a larger range for the whole growing season.

In terms of representation, RFs cannot present the results in the form of a tree structure, while conditional inference trees result in huge trees if the input variables are continuous. For HMMs it is better to map observation sequence "states" with the output state sequence to get a better understanding of the effect of a particular variable. This would result in a manageable transition matrix as opposed to a continuous observation sequence.

In BBNs, though the state of a variable can be deduced by back tracking, it doesn't clearly show it. It just presents the probabilities of the variable to be in the states for a given decision. For HMMs it is better to assume higher number of states, than less. The study presented in this work has examples of variables discretized into two as well as three states. While HMMs work with discretized variables, BBNs convert continuous variables to discrete by applying user-supplied intervals/bins on them, whereas trees can handle continuous variables appropriately. Trees explicitly classify the decision based on the limits of a variable, which is somewhat implicit in BBNs.

5.2 Final Conclusions

The main goal of this work was to extract those biophysical variables which contribute to irrigation decision-making. To achieve this aim we proposed three machine learning techniques which can take the available information and infer the connections of the information with irrigation decision. This way we can discern the factors which have a high influence on the irrigation decision. The models can answer questions regarding farmer irrigation decisions, such as, “Why does the farmer irrigate on a certain day”, “Can the decisions be grouped such that farmer types can be detected”, “Do the farmers look at many factors or have a single factor to rely on during the course of crop growth”, “Are there any similarities in irrigations for the same crop or year”, “If the same farmer is making decisions for the last few years, can we find a tool which can avail this information and make forecasts about his future actions” and “Can we find general irrigation rules at the command area level”. The presented models have not been exploited before as is evident from the limited documented research in the field.

To incorporate the available information correctly, specific types of models were used in this study to learn past farmer actions and infer about the future decisions. All the models were able to handle continuous as well as categorical inputs. In reality, the measured variables are not related on a one-to-one basis with the decision. To overcome this problem, we picked a practical indicator which the farmer looks at and assumed the variable that reflected the indicator. For eg. the farmer feels the soil to find out the moisture level. For this practical indicator we used soil stress coefficient, which denotes the level of soil moisture content.

Learning Bayesian belief networks, as the name suggests learnt the input-output relations between the factors and the target variables to build a framework capable of making inference on future events. The model suggested soil moisture condition, rooting depth, irrigating on weekends and when one’s neighbor irrigates as some of the reasons for irrigation by the farmers. The prediction accuracies for future events obtained from these models was good for alfalfa, and fair for barley and corn, due to limited information on the latter two. Trees framework found that crop water use, and irrigating when a neighbor irrigates, may be the basis of irrigation decision-making. These models needed complete information of the growing season to correctly forecast future decisions and had excellent prediction accuracies. The third analysis used hidden Markov models, which have the capability of modeling non-memory less systems. The results showed that farmers adapt to crop needs on an irrigation-to irrigation basis and hence it may be hard to classify them into different groups. Instead of estimating the approximate day of irrigation, these models predicted impending irrigation event based on the input factor levels.

On the whole, crop water use and irrigating when one's neighbor irrigates were found to be probable reasons used for irrigation at the study site. All other factors can still be debated. There was a different noticeable strategy for each irrigation as it came. The farmers were found to be sensitive to crop needs, quite flexible with their irrigation strategies, and risk-averse. With respect to irrigations and growth stages, the farmers were classified in to groups. One of those groups could take only minimal risk. The other groups aimed for good crop quality and as a result irrigated as soon as there was evident stress. At the end, it can be concluded that the farmers irrigation behavior can be studied by using biophysical conditions during the growing season to adequately analyze and forecast irrigation decisions.

Evolutionary algorithms provide a new avenue to model such problems and to make optimal use of the available information. These models also help in minimizing the uncertainty in the irrigation decisions. Though exact day of irrigation is difficult to estimate, the models used in the study find the time window where the farmer might make a decision. The decisions of farmers who decide to irrigate on a given day can be summed up to estimate, irrigation water use and eventually the expected amount of the water to be diverted.

Efficient command area water management is a long-standing concern for canal operators, especially those who handle on-demand irrigation water orders every day. This indicates that practical means of accurately forecasting farmer irrigation decisions could be useful in estimating short-term irrigation demands. This study introduced models which can be successfully used to study farmer's irrigation behavior and predict their subsequent actions. Lastly, the biophysical variables used in this study, or the variables for which they are surrogates, appear to be used in some way by the farmers to make an irrigation decision, and hence were found to be effective as predictors of future irrigation decision behavior.

5.3 Future Work

Using data meant specifically to model behavior should be a priority. The studies can be improved by getting data on planting dates and information on water orders. Water orders are crucial. The time between ordering water and applying it on the field is also crucial information. The possibility of renting water also complicates the process. The farm size and the crop rotation used by farmers for various years could be useful information. This would give insights into how farmers prioritize irrigating different crops in the growing season. From the modeling perspective, an approach needs to be designed which can refine the variables at each stage of analysis or present different variables at different irrigations to the models. It would be valuable to collect information from the farmers, in the form of surveys, as well as verify it by having monitoring devices like soil probes, or a flow meter at the field level. This would give reliable information on irrigation behavior.

APPENDIX

Table A.1: Soil Characteristics used for various soil types, fields (identifier used for the fields), years (2007-2010) and crops. Day of planting (DOP) in Julian day for starting the calculations for the growing season.

Crop	**Soil**	**Field ID**	**Year**	**DOP**	**Porosity**	**FC**	**WP**
A	SICL	101A	2007	78	0.42	0.38	0.203
A	SICL	101A	2009	78	0.412	0.37	0.203
B	SICL	101A	2010	73	0.36	0.33	0.25
A	SICL	101B	2007	78	0.42	0.38	0.203
A	SICL	101B	2009	78	0.412	0.37	0.203
B	SICL	101B	2010	73	0.36	0.33	0.25
A	SIC	102A	2007	95	0.48	0.42	0.17
C	SIC	102B	2007	91	0.5	0.46	0.23
C	SIC	102B	2009	100	0.45	0.42	0.23
B	SIC	102B	2010	76	0.423	0.4	0.25
A	SICL	103A	2007	97	0.33	0.33	0.2
A	SICL	103A	2008	81	0.33	0.297	0.203
A	SICL	103B	2007	96	0.34	0.31	0.203
A	SICL	103B	2008	82	0.33	0.297	0.203
C	SIC	104A	2008	86	0.4	0.36	0.2
B	SIC	104A	2010	146	0.35	0.32	0.25
A	SIC	105A	2008	72	0.48	0.42	0.29
A	SIC	105B	2007	92	0.482	0.42	0.28
A	SIC	105B	2008	72	0.482	0.42	0.28
A	SIC	105B	2009	90	0.482	0.42	0.28
A	SIC	105B	2010	94	0.482	0.42	0.28

Continued on next page

Table A.1 – continued from previous page

Crop	Soil	Field ID	Year	DOP	Porosity	FC	WP
A	SIC	106B	2008	90	0.482	0.42	0.28
A	SIC	106B	2009	90	0.482	0.42	0.28
A	SIC	106B	2010	94	0.482	0.42	0.28
A	SICL	107A	2007	79	0.45	0.43	0.203
A	SICL	107A	2008	67	0.412	0.37	0.24
A	SICL	107A	2009	81	0.412	0.37	0.203
A	SICL	107A	2010	96	0.412	0.37	0.203
C	SICL	107B	2007	98	0.42	0.38	0.2
A	SICL	107B	2008	100	0.412	0.37	0.203
A	SICL	108A	2009	102	0.45	0.41	0.29
A	SICL	108A	2010	97	0.45	0.41	0.29
A	SICL	108B	2010	97	0.45	0.41	0.29
B	SIC	109A	2007	78	0.36	0.33	0.25
A	SIC	109A	2008	68	0.47	0.4	0.28
A	SIC	109A	2009	92	0.482	0.42	0.28
A	SIC	109A	2010	105	0.43	0.415	0.28
B	SIC	109B	2007	78	0.36	0.33	0.25
A	SIC	109B	2008	67	0.482	0.42	0.28
A	SIC	109B	2009	90	0.482	0.42	0.28
A	SIC	109B	2010	105	0.43	0.415	0.28
B	SICL	110A	2008	75	0.34	0.31	0.16
B	SICL	110A	2009	70	0.31	0.29	0.22
A	SICL	110A	2010	105	0.31	0.29	0.22

Continued on next page

Table A.1 – continued from previous page

Crop	Soil	Field ID	Year	DOP	Porosity	FC	WP
B	SICL	110B	2009	69	0.31	0.29	0.22
A	SICL	111A	2007	87	0.412	0.38	0.203
A	SICL	111A	2008	89	0.37	0.33	0.2
A	SICL	111A	2009	85	0.31	0.31	0.2
B	SICL	111A	2010	67	0.32	0.305	0.2
A	SICL	111B	2007	97	0.45	0.43	0.203
A	SICL	111B	2008	79	0.45	0.4	0.2
B	SICL	111B	2010	67	0.32	0.305	0.2
A	SICL	112B	2007	86	0.412	0.38	0.203
C	SICL	112B	2008	95	0.45	0.42	0.2
B	SICL	113A	2009	71	0.31	0.29	0.22
A	SICL	113A	2010	106	0.33	0.305	0.22
C	SICL	113B	2007	87	0.38	0.34	0.2
C	SICL	114A	2007	90	0.42	0.38	0.2
A	SICL	114B	2009	103	0.412	0.38	0.25
A	L	115A	2007	71	0.39	0.36	0.29
A	L	115A	2008	69	0.39	0.36	0.29
C	L	115A	2009	96	0.37	0.35	0.25
C	L	115A	2010	90	0.32	0.3	0.2
A	L	115B	2008	82	0.47	0.45	0.37
C	L	115B	2009	90	0.42	0.39	0.26
C	L	115B	2010	98	0.38	0.365	0.26
C	SICL	116A	2007	109	0.37	0.33	0.2

Continued on next page

Table A.1 – continued from previous page

Crop	Soil	Field ID	Year	DOP	Porosity	FC	WP
C	SICL	116A	2008	118	0.35	0.33	0.28
B	SICL	116A	2009	86	0.33	0.31	0.25
A	SICL	116A	2010	105	0.33	0.32	0.25
C	SICL	116B	2007	112	0.32	0.3	0.2
C	SICL	116B	2008	109	0.31	0.29	0.2
B	SICL	116B	2009	69	0.3	0.29	0.16
A	SICL	116B	2010	97	0.3	0.29	0.16
C	SICL	117A	2008	90	0.45	0.42	0.2
A	SICL	117A	2010	84	0.45	0.43	0.2
B	SICL	117B	2009	82	0.36	0.33	0.25
C	SICL	118A	2007	84	0.42	0.39	0.2
A	SICL	118A	2008	86	0.412	0.37	0.24
A	SICL	118A	2009	91	0.412	0.38	0.25
B	SICL	118A	2010	85	0.412	0.38	0.25
C	SICL	118B	2007	84	0.42	0.39	0.2
A	SICL	118B	2008	86	0.412	0.37	0.24
A	SICL	118B	2009	91	0.412	0.38	0.25
A	SICL	118B	2010	92	0.43	0.41	0.25
A	SIC	119B	2009	77	0.482	0.44	0.28
A	SIC	119B	2010	70	0.482	0.47	0.31
C	SIC	120B	2010	117	0.42	0.4	0.26
B	SICL	121A	2007	95	0.36	0.31	0.16
B	SICL	121B	2007	95	0.36	0.31	0.16

Continued on next page

Table A.1 – continued from previous page

Crop	Soil	Field ID	Year	DOP	Porosity	FC	WP
A	SICL	121B	2009	93	0.37	0.35	0.24
A	L	122A	2007	75	0.39	0.37	0.29
A	L	122A	2009	91	0.39	0.344	0.25
A	L	122A	2010	82	0.44	0.42	0.25
A	L	122B	2009	74	0.39	0.34	0.24
A	L	122B	2010	94	0.36	0.35	0.24