Ethanol Plant Predictive Regression Models: The Importance of Plant Data Analytics

By

Cassandra Schneider

B.S., Kansas State University, 2012

A THESIS

Submitted in partial fulfillment of the requirements

for the degree

MASTER OF AGRIBUSINESS

Department of Agricultural Economics

College of Agriculture

KANSAS STATE UNIVERSITY

Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. Jason Bergtold

ABSTRACT

The modern use of data analytics is not new to production processes, however the substantial reliance of it in the ethanol industry has been increasing over recent years. Being able to pull larger amounts of data is important to monitor an ethanol plant's KPI's (key performance indicators). Taking data analysis to the next level of being able to run regression models and predictive type examination to accurately determine ethanol yield is the succeeding phase currently facing the ethanol business on the plant level. Using simple regressions and an industry survey to deliver data can help both ethanol plant personnel and vendor data scientists to work together to be able to use all the information on hundreds of ethanol plant variables that are gathered daily to provide predictive guidance. Over the years, the ethanol industry has also become a crucial business for imports and exports in countries, such as the United States, Canada, and Brazil. However, the ethanol industry does come with many risks and challenges and many of them are beyond an ethanol plant's control. For this reason, the purpose of this thesis is to examine the importance and impact of data analytics in ethanol production, and to determine the value that predictive modeling of ethanol yield can have for an ethanol plant. An ethanol industry-aimed survey was developed, conducted, and data summarized. Dependent variable and independent variables for regression analyses were analyzed to see the trends for 2010 from an Excel extract provided by Plant ABC. A linear regression model of ethanol plant data was used in this thesis to be able to examine contemporaneous dependence of fifteen different variables with the dependent variable, ethanol yield. Regression modeling was also used to determine the factors that are statistically significant in predicting ethanol yield using other

types of models with alternative functional forms, including the semi-log, double-log, and quadratic. Ethanol yield linear regression was estimated and showed that the independent variables of ratio milo, Drop pH, Drop DP4+, Drop Glucose, Drop Lactic Acid, and Drop Acetic Acid had *p*-values under 1% and have significant correlations to ethanol yield. The quadratic model yielded the lowest RMSE indicating the best predicted model out of the four models estimated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to take the time to express my sincere gratitude to my major professor, Dr. Jason Bergtold, for his continuous support and guidance through the completion of my thesis. I would also like to thank the rest of my thesis committee professors: Dr. Edward Perry and Dr. Daniel O'Brien. Also, many thanks goes out to my employers for their acceptance of higher education and knowing that work would be competing with school work as well. Kansas Ethanol, LLC and CEO, Michael Chisam, guided and helped mentor my path to finding a successful career. Novozymes North American, Inc. hired me knowing I was only halfway through a Master's program and allowed me the time, effort, and assistance needed to finish strong.

I would like to thank and congratulate my fellow talented and diverse MAB 2019 cohort. I would also like to recognize the MAB program staff of Dr. Allen Featherstone, Mary Bowen, and Deborah Kohl for everything they do daily as well as their guidance that kept me accountable throughout the whole program and thesis process.

I reserve my most sincere thanks to my family. My husband, Joe, and his unwavering support, was what ultimately got me through the program. He always had the encouragement I needed when the work, family, and school balance was too much to handle. To my children, Clayton, Adleigh, and Paisleigh, who did not consent to the time that was sacrificed from them. This will hopefully be appreciated and become an example for them someday. Thank you for your endless patience, as together we share this success.

CHAPTER I: INTRODUCTION

As the ethanol industry continues to offer new opportunities for the United States and countries all around the world, it has become an engine for agricultural growth, bringing many opportunities for employment, and economic development, especially in rural locations. With today's environmental issues, it also provides a cleaner burning and potentially lower costing fuel alternative. Over the years, the ethanol industry has also become a crucial business for imports and exports in countries, such as the United States, Canada, and Brazil. However, the industry does come with many risks and challenges and many of them are beyond an ethanol plant's control. For this reason, the purpose of this thesis is to examine the importance and impact of data analytics in ethanol production and the value predictive modeling of ethanol yield can have for an ethanol plant.

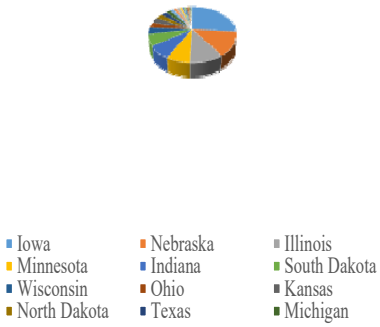
The ethanol industry provides jobs and income in numerous states in the United States. Most ethanol producing states in 2017 were located in the Midwest (Table 1.1). In 2017, the United States alone produced 15,856 million gallons of ethanol, producing an income of 24,087 million dollars, and provided around 358,780 jobs. Iowa, Nebraska, and Illinois are the top three ethanol producing states, with a 50.5% share of production (Figure 1.1).

Table 1.1: Contribution of the Ethanol Industry to Individual State Economies, 2017

Plants	Production (Mil Gal)	Production Share	GDP (Mil $)	Employment (Jobs)	Income (Mil $)
Iowa	4177	26.3%	$ 3,868.00	39018	$ 3,010.00
Nebraska	2176	13.7%	$ 2,015.00	21685	$ 1,664.00
Illinois	1659	10.5%	$ 1,536.00	16070	$ 1,316.00
Minnesota	1204	7.6%	$ 1,115.00	11923	$ 1,010.00
Indiana	1173	7.4%	$ 1,086.00	12127	$ 990.00
South Dakota	1060	6.7%	$ 982.00	11051	$ 913.00
Wisconsin	583	3.7%	$ 540.00	6505	$ 593.00
Ohio	548	3.5%	$ 508.00	6172	$ 569.00
Kansas	491	3.1%	$ 455.00	5629	$ 531.00
North Dakota	465	2.9%	$ 431.00	5381	$ 513.00
Texas	385	2.4%	$ 357.00	4619	$ 459.00
Michigan	354	2.2%	$ 328.00	4323	$ 439.00
Missouri	261	1.6%	$ 242.00	3437	$ 376.00
Tennessee	225	1.4%	$ 208.00	3001	$ 352.00
California	218	1.4%	$ 202.00	2937	$ 347.00
New York	150	0.9%	$ 139.00	2317	$ 301.00
Colorado	127	0.8%	$ 118.00	2107	$ 286.00
Georgia	120	0.8%	$ 111.00	2044	$ 281.00
Pennsylvania	110	0.7%	$ 102.00	1953	$ 274.00
Oregon	98	0.6%	$ 91.00	1843	$ 266.00
Virgina	64	0.4%	$ 59.00	1533	$ 243.00
Idaho	60	0.4%	$ 56.00	1497	$ 241.00
Mississippi	54	0.3%	$ 50.00	1442	$ 237.00
Arizona	50	0.3%	$ 46.00	1406	$ 234.00
Kentucky	36	0.2%	$ 33.00	1278	$ 225.00
Florida	8	0.1%	$ 7.00	1023	$ 206.00
All Other			$ 29,771.00	186,459	$ 8,211.00
TOTAL U.S.	15,856	100.0%	$ 44,456.00	358780	$ 24,087.00

Source: (Urbanchuk 2017)

Figure 1.1: State Production Share, 2017



Source: (Urbanchuk 2017)

Plant operations operate in a risky environment as all inputs and outputs from ethanol production, as well as other coproducts are dependent on markets. Markets can be erratic and it is important for ethanol plants to be as efficient as possible to help manage their risks. Consistency and improving what happens in the production process is something plants and employees do have an influence on. Ethanol plants need to focus on the types of things they do have control over, such as process controls, maintenance schedules, lab testing, ingredient dosing, and other plant optimization tools, to do the best they can and be as efficient as possible.

1.1 Thesis Objectives

Innovation is constantly occurring in the industry, so even the few parts of the ethanol process that can be controlled, change over time and need to be consistently monitored and evaluated for efficiency. Most production type industries use data as a way

to observe and make changes for better plant results. For many years, ethanol plants have relied on vendors to occasionally analyze their data, and consequently, this does not provide the analysis needed to make day to day changes. The objective of this study are to:

1. Survey the industry and understand how their data is being handled.
2. Identify if plant changes are being made using data analytics and by whom.
3. Formulate simple regressions to determine variable effects on ethanol yield.
4. Develop predictive type regressions to examine marginal changes that might be made to increase ethanol yield in the future.

1.2 Thesis Organization

Contents following throughout this thesis include a literature review in chapter two used to review past literature and documentation that might be related to production industries, the ethanol industry, or similar studies. Data information that was collected for this thesis includes a ten question survey (full survey in appendix) that was aimed at ethanol plant management at the plant level. Data also included secondary data pulled from daily production data that is collected at all or most ethanol plants. Some data was pulled using technology software, while the rest was hand entered, which may result in inputting errors. Explanation of all variables and the theory will be presented in chapter three. Chapter four examines methods, including linear regression analysis to examine factors impacting ethanol yield. Survey data also is used to aid in defining the independent variables used in the regressions conducted. Simple regression models evaluated as prediction models were estimated to assess their use in predicting ethanol yield. Results are presented in chapter five. A concluding summary and explanation of future research will complete the thesis in chapter six.

CHAPTER II: LITERATURE REVIEW

The literature review examines the relationship of different aspects of the research in the relevant literature. Literature on the ethanol industry related to the process and regression data analysis is sparse. Looking at similar production type processes though helps to understand concepts related to the ethanol industry and associated usefulness of data analytics.

2.1 Putting Data to Work

An article produced by Ethanol Producer Magazine was published around the same time that vendors started actively introducing JMP (pronounced jump) software to producers from the SAS Institute (Jessen 2014). A significant reason why ethanol plants historically didn't do a lot of data analysis in-house was because of time constraints. The article describes how vendors and plant employees can easily use the software and that data analysis progressively gets easier and faster as it continues to be useful. Jessen (2014) discusses the price of the software in 2015 as: "An annual license for JMP is $1,540 in 2015, JMP Pro, which has additional capabilities, has an annual license cost of $14,900, according to SAS representative" (Jessen 2014, p.1). Lastly, the article depicts a few real life experiences at an ethanol plant where they were able to troubleshoot a rising sulfate issue that ended up being affected by longer fermentation times using JMP to quickly model the problem and produce graphs that depicted the negative trend in sulfate levels (Coward-Kelly 2011). JMP can also be used to optimize dosing strategies based off data analytics.

2.2 Observational Data-Driven Modeling and Optimization

A paper by Sadati, Chinnam and Nezhad (2017) isn't directly related to the ethanol industry, as it references biopharmaceutical production, but it has the same concepts as to

why data-driven decisions are important. “In the context of production, data-driven approaches can exploit observational data to model, control and improve process performance” (Sadati, Chinnam and Nezhad 2017, p.456). This quote, accurately expresses the main objective of this paper. The authors make a good point that even though technology continues to make data capturing much easier than before, there is still a need to conduct data analysis, and to be able to achieve improved yields in whatever industry production takes place in. The authors also present various modeling and higher level equations that can be used to optimize variables.

2.3 Data-Driven Modeling and Monitoring for Plant-wide Industrial Processes

Similar to the previous article, Ge (2017) recognizes the need and responsiveness that data modeling has in many industry processes. The author discusses the difficulties with data collection in industry processes, because of the hefty amount of data points that are being collected daily, creating a lot of data, which requires processing and sorting. As an industry, ethanol plant employees and data scientists are able to collect and extract more data than ever, but being able to organize it all and explain it all accurately is a struggle. Ge (2017) reviews the challenges facing different industries, such as data monitoring. He continues on to discuss potential future issues that data modeling might bring to light, such as the large volumes of data being able to be generated with today’s technology and the increasing complexity of production plant processes.

CHAPTER III: THEORY

A profoundly overlooked part of every ethanol plant is the data that is being generated all over the facility. There are data points being pulled and stored from the grain entering the facility to the ethanol and co-products being produced on the backend. This makes it difficult to maintain efficiency when dealing with such numerous and meticulous processes and a vast array of data points. It is essential to be able to optimize using these data points to become as efficient as possible through the power of data analysis.

3.1 Data Summary

Due to the excessiveness of data being produced every day, there have been many different technology platforms and companies that are able to pull data, store data, and conduct data analysis. Plant management and staff normally have a very busy daily schedule, making time for data analysis a lower priority. Thus, many vendors in the industry have made this a priority as a customer service product, but more and more customers are realizing the value of analyzing data in-house. However, with so many variables effecting overall ethanol yield, it makes it difficult to evaluate. Ethanol plant employees are finding simple and quick ways to show simple control charts, but there is added value by use of other techniques, such as regressions to examine the ethanol production process and predictive methods to help predict how, say a fifty to seventy-hour fermenter might yield before it is complete. There are many plants, vendors, and companies working to find a tool or way to do these types of analyses, as it can be valuable for the future of the ethanol industry.

3.2 Theoretical Model

Experience suggests that the relationship of the dependent variable ethanol yield percentage is a function of a number of independent variables, as follows:

Ethanol Yield (%) = f (AGE, MILO, BACKSET, SLURRY, LIQ, pH, BRIX, TEMP, DP4+, DP3, MALTOSE, GLUCOSE, LACTIC, GLYCEROL, ACETIC)

Table 3.1 represents the independent variables and their expected signs for the coefficients.

Table 3.1: Theoretical Ethanol Yield Model Coefficient

Coefficient	Independent Variable	Expected Sign
β_{AGE}	Fermentation Age	Postive
β_{MILO}	Milo Grind Ratio	Postive
$\beta_{BACKSET}$	Backset	Negative
β_{SLURRY}	Slurry Solids	Postive
β_{LIQ}	Liquefaction Solids	Postive
β_{pH}	Drop pH	Negative
β_{BRIX}	Drop Brix	Negative
β_{TEMP}	Drop Temperature	Negative
β_{DP4+}	Drop DP4+	Negative
β_{DP3}	Drop DP3	Negative
$\beta_{MALTOSE}$	Drop Maltose	Negative
$\beta_{GLUCOSE}$	Drop Glucose	Negative
β_{LACTIC}	Drop Lactic Acid	Negative
$\beta_{GLYCEROL}$	Drop Glycerol	Negative
β_{ACETIC}	Drop Acetic Acid	Negative

Ethanol plants may chose fermentation times (age) based on a variety of reasons, but overall it seems the longer the better in terms of ethanol yield. Most plants typically run 100% corn as a feedstock, so milo (via the milo grind ratio) may have a positive impact on ethanol yield beyond cost effectiveness. Backset may sometimes result in process contamination that might result in an infection (a bacterial infestation in the process), so it

is assumed it would generally have a negative effect on ethanol yield. Unless margins are tight, both slurry and liquefaction solids are ran at an optimized rate to help with ethanol yield, so in most cases these factors should be positively related to ethanol yield. Fermentation variables of drop pH, drop brix, and drop temperature may fluctuate quite often so could be negative. Fermentation variables of drop DP4+, drop DP3, drop maltose, and drop glucose are all sugars. The more percentage of these compounds left, means not all sugar was converted to ethanol, which would have a negative effect on ethanol yield. Fermentation variables of drop lactic acid and drop acetic acids are organic in nature and are likely detrimental to ethanol production. Drop glycerol is ideal to be as low as possible, but given it is formed during the fermentation process, it would be assumed to have a negative relationship with ethanol yield.

3.3 Variables

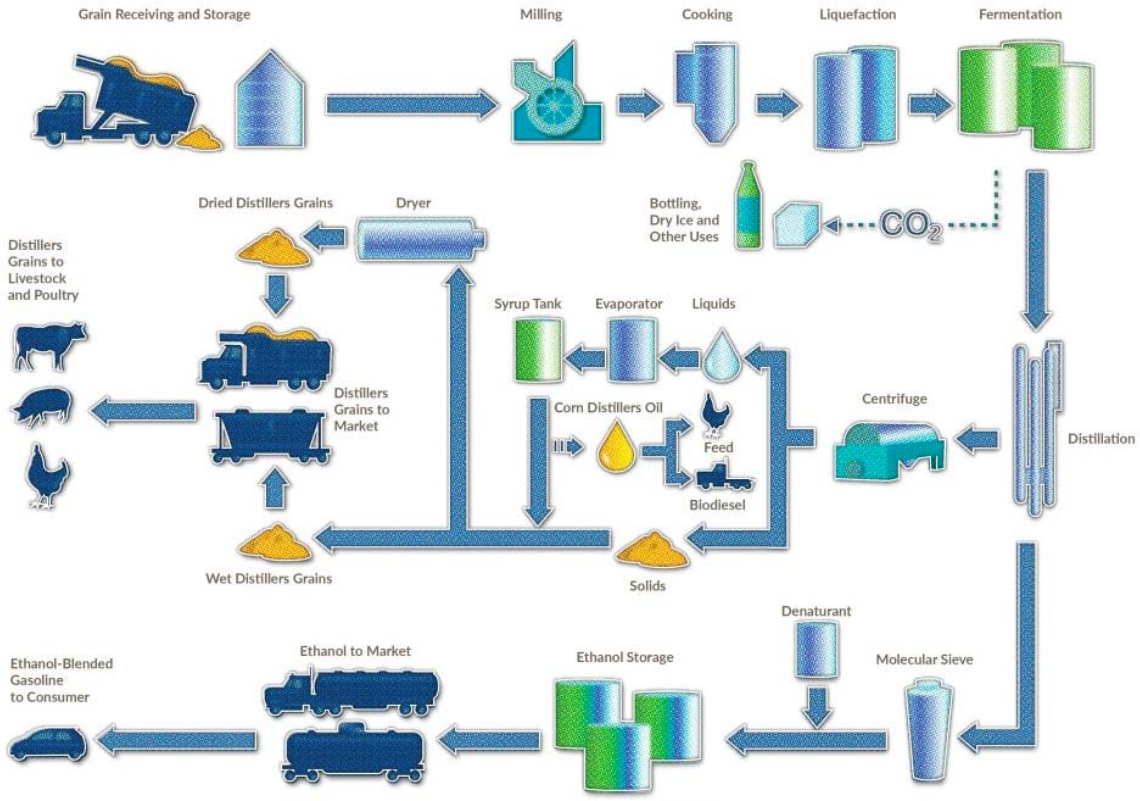
Figure 3.1 below, shows a basic diagram of a modern day dry mill ethanol plant facility. The following basic eight production steps used, will be referenced back to where the dependent and independent variables fit into the ethanol production process:

1. Grain Receiving and Storage
2. Milling (hammer mills)
3. Cooking/Slurry
4. Liquefaction
5. Fermentation
6. Distillation
7. Centrifuges
8. Ethanol Storage

Grain comes into the ethanol facility normally via truck and rail. The grain is stored until it is ready to use. The grain is milled through the hammer mill milling process to create ground grain. The ground grain is then mixed with water to make a mash. In the slurry (cooking) phase of the process, alpha amylase enzymes are added to break down the grain starch and the mash then goes through liquefaction. Liquefaction allows for residence time to continue to allow enzymes to convert and lower the viscosity of the mash. Mash then is pumped to fermentation tanks where it will stay for two to three days. After the fermentation process, the mash is sent through the distillation process to separate out the 200 proof ethanol and the rest of the leftover mash. The ethanol is then sent through a molecular sieve to remove additional water (Renewable Fuels Association n.d.), (Figure 3.1). The final product ethanol is either sent to ethanol storage or combined with other possible chemicals such as a denaturant and/or corrosion inhibitor to make the ethanol not consumable and to help reduce corrosion in storage. Denaturant may also be blended into the ethanol stream while loading a truck or rail car by use of an ethanol blending skid. The leftover mash is sent through large centrifuges to separate off the stillage from the solids. The solids are then used as the wetcake for wet distiller's grains (WDG), typically for cattle feed end users. Wetcake may also be dried in what is normally a barrel dryer to produce distillers dried grains with solubles (DDGS), which also might be used for cattle feed or other uses such as poultry feed or extruded co-products. There are also additional technologies out in the industry that bolt on to the ethanol production process, such as the ability to capture CO2 off of the fermentation process, which is purified in another additional procedure to later become CO2 products. Oil production may also be added on

as a technology to create corn and/or milo oil used for feed products or to be used at biodiesel refineries.

Figure 3.1: Dry Mill Ethanol Process Diagram



Source: (Renewable Fuels Association n.d.)

3.3.1 Dependent Variable - Fermentation Ethanol Yield

The dependent variable in this analysis is ethanol yield at the end of the fermentation process or what is often referred to as a fermentation drop ethanol yield percentage. No matter what the ingredients used are, new technologies installed, co-products produced and sold, or the way that a plant is ran, ethanol yield is the main goal for all plants. It is very important to produce efficiently and be cost effective. Ethanol yield can be measured in multiple ways. Much of the following data will come from data retrieved using HPLC "(High-performance liquid chromatography; formerly referred to as high-

pressure liquid chromatography). HPLC is a technique in analytical chemistry used to separate, identify, and quantify each component in a mixture. It relies on pumps to pass a pressurized liquid solvent containing the sample mixture through a column filled with a solid adsorbent material" (Contributors 2019, p.1).

3.3.2 Independent Variables

The independent variables were chosen based off of the survey and internal interests from customers. (Table 3.1,) These variables may interact with each other, having an impact on regression results based on where the independent variables come into the process. Fermentation Age, Ratio Milo, Backset, Slurry Solids and Liquefaction Solids are all variables that are at the front side of the process. Drop pH, Drop Brix, Drop Temperature, Drop DP4+, Drop DP3, Drop Maltose, Drop Glucose, Drop Lactic Acid, Drop Glycerol, and Drop Acetic Acid are all variables that are captured at the end of the fermentation process.

Fermentation Age (hours) – Fermentation age is a reference to the total hours that the fermentation process takes to complete before the batch goes through the distillation process. This time may be effected by various plant related changes such as throughput rates, cleaning times as well as plant mechanical or maintenance items either normal or out of the ordinary. Even though every plant has different operational procedures resulting in longer or shorter fermentation times, normally the range spans between 50-70 hours.

Grind Ratio Milo (%) – Assumingly, most ethanol plants in the United States usually grind 100% corn as their feedstock for ethanol fermentation. Many of the plants in Kansas choose to grind sorghum, or milo, as well for the benefits of being readily grown and available logistically, with purchase costs typically being cheaper when compared to corn prices. Unfortunately, milo can be dirty, sandy, corrosive, and rough on the

maintenance side of the plant. Even with the additional costs that may occur due to running milo through the plant, most find the financial benefits to outweigh the negative effects or costs it might have. Given market dynamics, a corn/milo plant might make grind ratio adjustments as needed or warranted. These plants have the option of running whatever ratio of corn to milo they want, given what might make economic sense at the time. Grain procurement costs along with grain availability comes into play in most of these scenarios. Sometimes during fall crop harvest, it might be logical to run 100% corn. While other times throughout the winter, when it is not driving season and ethanol margins are lower, it's possible it would be more profitable to run a higher milo blend. From experience, corn does seem to typically have a higher yielding pattern in general. However, the choice between corn and milo use always comes down to the profitability. Grind ratio would be adjusted at step 2 milling in Figure 3.1. Grind ratio data was analyzed by inputting when each change was made throughout the 2010 calendar year for the results data.

Backset (%) – Backset is also referred to as Thin Stillage or Centrate and occurs at step 7 centrifuges from Figure 3.1. The definition of backset is sometimes hard to understand. Essentially, backset is the whole stillage leftover from the fermentation process, which is sent to multiple large centrifuges, on the backend of the ethanol process. The whole stillage is separated into a liquid, which is the backset, and solids, known as wetcake and can be used either as wet distiller's grains or dried in a dryer to become Distillers Dried Grains with Solubles (DDGS) for cattle feed or other uses. The percentage of the backset/thin stillage stream is recycled back to the front of the process and can vary by plant.

Slurry Solids (%) – Slurry is also called "mash" and is produced in step 3 in Figure 3.1. Most ethanol plants have only two slurry tanks. The mash made up of slurry solids is mainly ground grain, flour from the hammer mill process, and backset recycled back around. Alpha amylase is also added to slurry to help keep viscosity low and start converting the grain starch to sugar for the yeast to consume in fermentation. Generally, at ethanol plants, slurry solids running anywhere from 31-35% solids in mash depending on the plant.

Liquefaction Solids (%) – The mash that leaves slurry moves onto two or three liquefaction tanks in step 4 of Figure 3.1, depending on the plant. Liquefaction mash normally stays in this stage long enough to let alpha amylase enzymes break down the starch in the flour into sugars for the yeast to consume. In the field, liquefaction solids constitute anywhere from 31-35% solids in mash depending on the plant. At times liquefaction solids may be slightly lower as a percent of mass than slurry solids for a specific plant and its production processes.

Fermentation Drop pH – "Fermentation Drop" refers to the last sample that was taken at the end of fermentation time before the fermenter is pumped to the distillation side of the process. Drop pH is used to confirm that there was not an excursion of pH throughout the process. Actual physical pH levels are very important in the cooking phase when enzymes are working to convert starch. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Brix (%) – Brix is a measure of sugar present in the fermentation drop sample. Traditionally measures are done using a refractometer. Now

there are digital reader options to make it much more consistent and accurate. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Temperature (°F) – Fermentation temperature is very important for yeast health for ethanol production. The weather has a large influence on how hot fermentation mash temperatures may get in the summer or how low in the winter. Yeast cells can start stressing if temperature gets too hot. Temperature may cause enzymes to become unstable and denature, leading to inactivity. Individual ethanol plants may handle temperature fluctuations differently, but all plants monitor temperature throughout the fermentation process to keep yeast healthy. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop DP4+ (%) – DP means degree of polymerization and describes how many glucose molecules are bonded together in a chain. Starch starts out as very long chains of glucose and gets broken down into glucose throughout the ethanol process. That being said, DP4+ is known as a void peak by lab managers in the ethanol industry because it contains any compound that doesn't stick to the HPLC column, hence the "plus" part. If there are dextrins left over in fermentation, DP4+ will be elevated at fermentation drop and identified in the HPLC fermentation drop sample when ran. This is why it is important to monitor. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop DP3 (%) – Similar to DP4+, DP3 is also a sugar. If there are dextrins leftover in fermentation, DP3 will also be elevated at fermentation drop and identified in the HPLC fermentation drop sample when conducted. This independent

variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Maltose/DP2 (%) – DP2s are known as disaccharides and common disaccharides are maltose, sucrose, and lactose. In the ethanol industry as a whole, they monitor the Maltose/DP2 HPLC peak is monitored. It's percentage decrease in fermentation and what is leftover is identified in the HPLC fermentation drop sample when ran. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Glucose (%) – Dextrose is leftover at the end of the fermentation process. The lower the percentage of dextrose left, the more that was consumed and converted to ethanol. One of the main goals is for yeast to consume the fermentable sugars to get as much carbon to turn to ethanol as possible. In the industry, it is common knowledge that one molecule of glucose once broken down becomes two molecules of ethanol and two molecules of carbon dioxide. If sugars aren't consumed in large enough quantities fermentation time has not been long enough, which lowers potential ethanol yield. In addition, extra sugar in the wetcake and DDGS distiller's products could lead to other issues, so the less sugar leftover at the end of fermentation the better. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Lactic Acid (%) – Lactic acid is an organic acid that has a three carbon chain with hydroxyl and carboxyl group and can be produced from yeast during fermentation. However, it is commonly introduced into the system by lactic acid bacteria produced as a product of carbohydrate catabolism. Lactic acid is monitored in

fermentation. If it becomes elevated, then infection will occur and stall out a fermenter early, resulting in lower ethanol yield. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Glycerol (%) – Drop glycerol is monitored because reducing the glycerol percentage at the end of fermentation as much as possible, is the desired result. Glycerol retains moisture and represents carbon that could have been converted into ethanol. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

Fermentation Drop Acetic Acid (%) – Acetic acid is also an organic acid that has a single carbon carboxyl group. Acetic acid can be found in fermentation as a product of yeast or from lactic and acetic acid bacteria. Acetic acid is monitored in fermentation because if it becomes elevated, then infection will occur and stall out a fermenter early, resulting in lower ethanol yield. Acids may also cycle up over time, so monitoring trends help prevent issues. This independent variable is located at the end of step 5 in Figure 3.1 once fermentation is complete and a sample is taken for lab analysis.

CHAPTER IV: DATA AND METHODS

An ethanol industry aimed survey was developed, conducted, and data summarized. Dependent variable and independent variables for regression analyses were analyzed to see the trends for 2010 from an Excel extract provided by Plant ABC. A linear regression model of ethanol plant data was used in this thesis to be able to examine contemporaneous dependence of fifteen different variables with the dependent variable, ethanol yield. Regression modeling was also used to determine the factors that are statistically significant in predicting ethanol yield using other types of models with alternative functional forms, including the semi-log, double-log, and quadratic. Once all the data was organized and scanned for errors and zeros, regression modeling was estimated using JMP software by SAS Institute, Inc. and GRETL software. Plant ABC fermentation data from the year 2010 was the only year used for all regressions. There were over 500 batches total in 2010, because this would have created too large of a table, the averages for each year were computed into an interpretable table for the data used in the regressions (Table 4.1). The dependent and independent variables are all represented in the table. It is important to note that this data represents the production process prior to production of corn oil and distillers grains, in order to examine the primary process for this thesis, the fermentation process.

Table 4.1: 2010 Plant ABC Regression Data

2010 Monthly Avg.	Ferm Age (Hours)	Grind Ratio Milo	Backset (%)	Slurry Solids (%)	Liq Solids (%)	Drop pH	Drop Brix (%)	Drop Temp (°F)
January	64.43	80.00	45.91	34.24	33.62	4.81	11.22	85.23
February	64.83	85.00	45.01	34.19	33.15	4.83	11.06	85.22
March	62.38	85.00	44.86	33.68	33.29	4.89	10.89	85.10
April	62.10	85.00	42.89	34.20	33.58	4.61	10.91	85.46
May	63.93	80.45	44.89	33.95	33.54	4.73	11.17	85.65
June	65.24	70.00	45.09	34.07	33.86	4.72	11.06	86.76
July	62.88	64.32	44.01	34.04	34.50	4.60	11.01	87.30
August	63.79	53.18	45.36	34.29	34.23	4.62	10.93	86.83
September	64.02	35.00	42.95	34.22	33.51	4.60	11.39	86.80
October	63.28	50.00	43.20	34.97	33.64	4.66	11.32	86.70
November	64.24	70.00	42.70	35.46	34.22	4.58	10.96	86.31
December	63.58	70.00	40.92	35.33	34.50	4.59	10.89	86.21
Minimum	62.10	35.00	40.92	33.68	33.15	4.58	10.89	85.10
Maximum	65.24	85.00	45.91	35.46	34.50	4.89	11.39	87.30
Mean	63.72	69.00	43.98	34.39	33.80	4.69	11.07	86.13
Stand. Dev.	0.94	15.99	1.46	0.56	0.46	0.11	0.17	0.77

Table 4.1: 2010 Plant ABC Regression Data Continued

DP4+ (% w/v)	DP3 (% w/v)	Maltose (% w/v)	Glucose (% w/v)	Lactic Acid (% w/v)	Glycerol (% w/v)	Acetic Acid (% w/v)	Ethanol (% w/v)
0.71	0.09	0.35	0.18	0.22	1.54	0.08	14.88
0.68	0.09	0.30	0.11	0.18	1.53	0.07	14.85
0.62	0.09	0.24	0.16	0.17	1.56	0.07	14.44
0.63	0.09	0.25	0.12	0.21	1.56	0.08	14.32
0.66	0.08	0.26	0.25	0.17	1.58	0.08	14.25
0.67	0.08	0.24	0.09	0.14	1.59	0.10	14.20
0.70	0.09	0.40	0.17	0.15	1.64	0.11	13.92
0.74	0.08	0.69	0.14	0.13	1.62	0.11	14.02
0.67	0.07	0.28	0.21	0.13	1.54	0.11	14.02
0.66	0.08	0.35	0.14	0.15	1.84	0.09	14.23
0.67	0.09	0.37	0.14	0.16	1.71	0.07	14.10
0.63	0.08	0.24	0.16	0.10	1.70	0.06	14.05
0.62	0.07	0.24	0.09	0.10	1.53	0.06	13.92
0.74	0.09	0.69	0.25	0.22	1.84	0.11	14.88
0.67	0.08	0.33	0.16	0.16	1.62	0.09	14.27
0.04	0.01	0.13	0.04	0.03	0.09	0.02	0.31

4.1 Survey

A ten question survey was developed and administered to meet thesis objectives. The survey was designed to help demonstrate the need for data analysis in ethanol production. The survey was also used to examine what kind of additional data analysis software or data analytics programs ethanol plants might be using in the industry. Lastly, the survey was designed to see how data might be used moving forward in the ethanol industry.

The survey was administered on SurveyMonkey.com in order to protect the anonymity of respondents. It was conducted over a two-week period and was advertised on LinkedIn for network ethanol industry connections as an at-will survey targeted to ethanol plant management (plant managers, CEOs, operations managers, and lab managers). The survey yielded 46 responses. It was assumed all respondents were from different ethanol plants in the United States. The responses are trackable by random number identifier as provided in SurveyMonkey.com as #1-46, but not by name or plant name. The plant that was used for regression analysis was also one response included in the survey results. (See Appendix A.2 for full survey) The survey was developed to confirm the need for data analysis and analytics at the plant level and the potential future predictive modeling. Allowing comments for each question was problematic as many respondents had additional questions or questioned the original question. The survey should have been read by experts from the industry and by coworkers prior to sending it out as there were some useful comments from colleagues made after the survey was already live and results were being tallied. Some questions yielded results in the comments that resulted in rewording some of the response options and how results were matched for the final data analysis. Allowing for

more time for data collection would allow more possible data points as well. Overall, the survey was useful, considering what it was initially designed for, but there were many lessons learned for conducting professional surveys in the future.

Overall, implications and findings from the survey were valuable. The survey reiterated the importance of analyzing data and confirmed analytics will continue to evolve in the ethanol industry. Only received partial results about what systems ethanol plants are currently using to run their plants and what software is being used in-house to analyze data due to how questions were worded at times. Certain survey data was not used due to incompleteness. An interesting finding was the amount of data analysis that was being conducted in-house versus relying on vendors, since time restrictions seem to make data analysis a low priority, historically.

4.2 Data

4.2.1 Dependent Variable

Drop ethanol yield percentage is data that is collected from an HPLC instrument and measured in weight per volume percentage in a sample collected at the end of fermentation time. Data was provided in an excel data extract and summarized in monthly averages (Table 4.1) by Plant ABC. The data was then analyzed using JMP software (Figure 4.1). Assuming that the HPLC results were dialed in and calibrated correctly, fermentation drop ethanol percentage started off higher for the first quarter of 2010 and then dropped and remained consistent the rest of the year (Figure 4.1).

Plant ABC is a batched dry mill ethanol production facility meaning even though it is a 24/7 operation, it is not a continuous type operation. Each fermentation that is completed is considered a batch. The X axis of Figure 4.1 is labeled as batches throughout the thesis and even though the time frame varies between ethanol batches, all fermenters

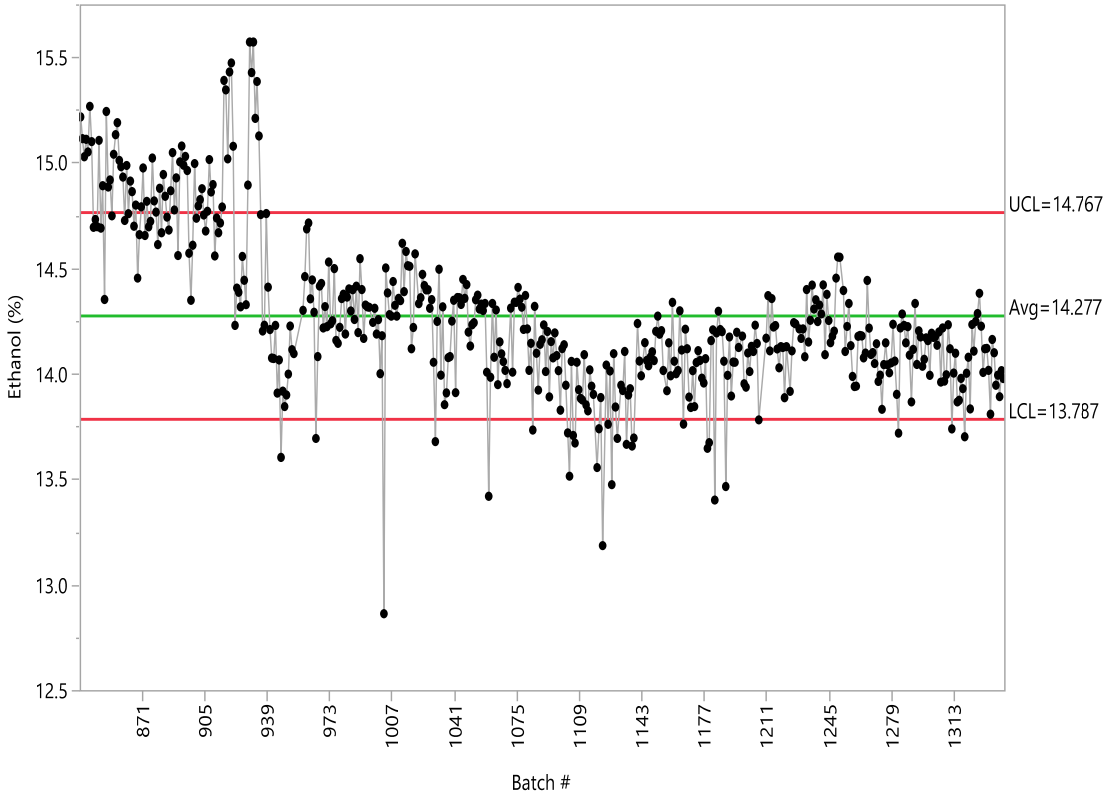
are fermenting at various time points at any given time so that batches are spaced out enough, they aren't starting and finishing at the same time. Plant ABC had four total fermenters during the time this data was documented and was completing, also known as dropping, one or two fermenters a day.

In addition to the fermenters, a lesser amount of mash is propagated in a propagation tank, where the yeast and a few other ingredients may be added depending on the recipe, that is 13,500 gallons for normally 7-9 hours and then that propagated mash is sent to a 730,000-gallon fermenter to be gradually added with the rest of mash for the rest of fermentation time. For this set of data, fermentation drop ethanol weight per volume percentage was 14.277% (Figure 4.1). This plant, in particular, did not record propagation times during this 2010 time frame and started recording them in later years, so the exact time is unknown for each batch in the data set.

Table 4.2: Data Summary, 2010

2010 Monthly Avg.	Ferm Age (Hours)	Ethanol (% w/v)
January	64.43	14.88
February	64.83	14.85
March	62.38	14.44
April	62.10	14.32
May	63.93	14.25
June	65.24	14.20
July	62.88	13.92
August	63.79	14.02
September	64.02	14.02
October	63.28	14.23
November	64.24	14.10
December	63.58	14.05
Minimum	62.10	13.92
Maximum	65.24	14.88
Mean	63.72	14.27
Stand. Dev.	0.94	0.31

Figure 4.1: Drop Ethanol (%) By Batch Number Data, 2010



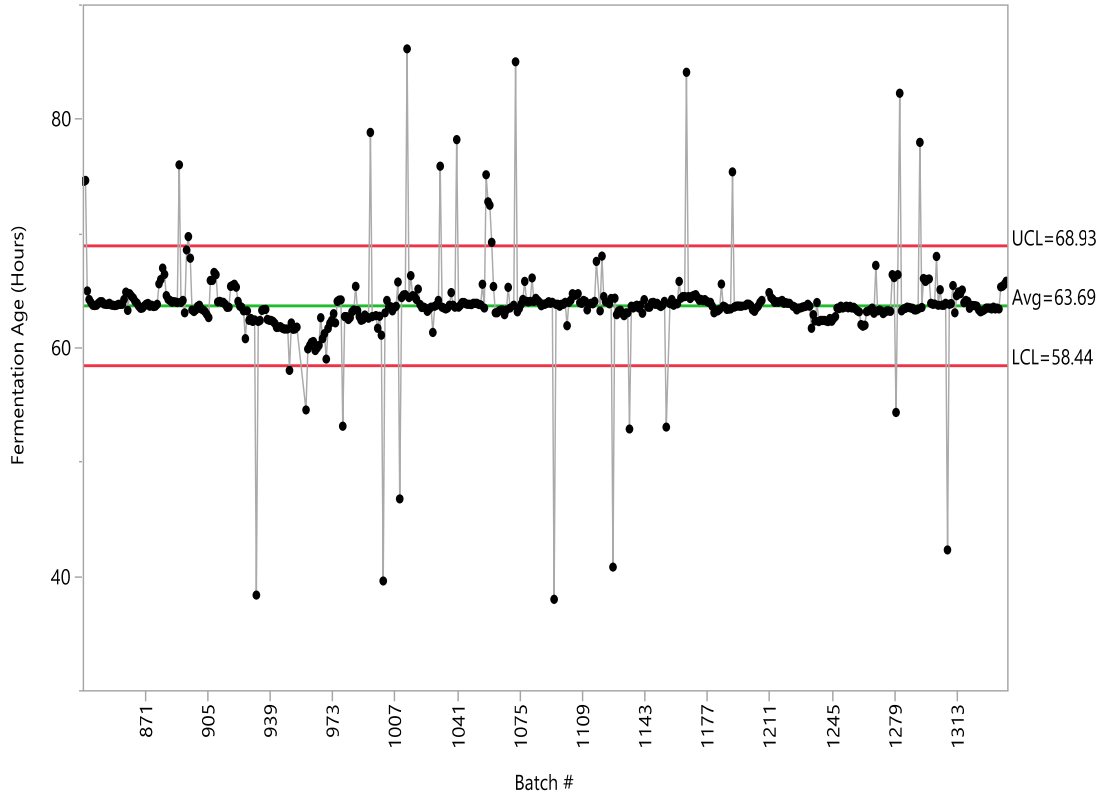
4.2.2 Independent Variables Data

All data for the independent variables was provided by Plant ABC in the form of an Excel spreadsheet. Some data was automatically inputted into their system by the HPLC and some data was hand entered data from ethanol plant employees.

4.2.2.1 Fermentation Age Data 2010

Fermentation age stayed rather consistent over 2010, averaging 63.69 hours. Any outliers were most likely due to fermentations that were longer due to plant shut downs or emergency type situations where they could not drop at normal times. This variable was an added column using the formula of fermentation drop reading date/time subtracted from the fermentation start fill time all multiplied by twenty-four hours to get the actual fermentation age time.

Figure 4.2: Fermentation Age By Batch Number Data, 2010



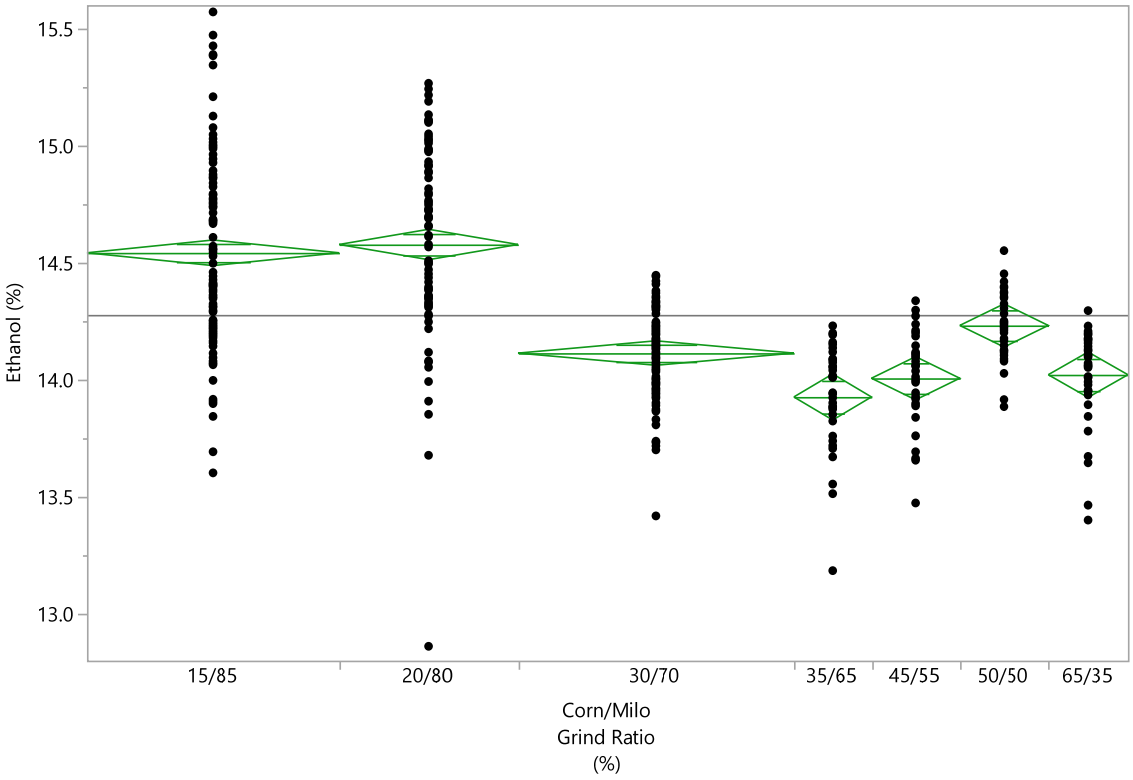
4.2.2.2 Grain Grind Ratio 2010 (Corn/Milo)

For regression purposes, only the milo ratio is used. Plant ABC accounting keeps track of grind percentages of the ground grain throughout the year with an Excel spreadsheet that is summarized in Table 4.2 for per day usage. According to Figure 4.3, higher grinds of milo at 85% and 80% resulted in the most ethanol yield for 2010.

Table 4.3: Calculated Grind Ratio Corn/Milo, 2010

	Calculated Using Gross Receipts			Calculated Using Net Receipts		
Date	**Per Day Usage**	**% Corn**	**% Milo**	**Per Day Usage**	**% Corn**	**% Milo**
01/01/10	58,744	21.34%	78.66%	57,846	21.46%	78.54%
01/04/10	59,538	19.15%	80.85%	58,938	19.21%	80.79%
02/01/10	59,002	15.49%	84.51%	58,295	15.64%	84.36%
03/01/10	59,312	14.80%	85.20%	58,701	14.86%	85.14%
04/05/10	60,739	16.04%	83.96%	60,087	16.15%	83.85%
05/03/10	59,621	22.99%	77.01%	58,911	23.12%	76.88%
06/01/10	59,062	30.60%	69.40%	58,444	30.74%	69.26%
07/02/10	59,450	34.74%	65.26%	58,944	34.93%	65.07%
07/30/10	58,458	44.07%	55.93%	58,077	44.18%	55.82%
08/30/10	59,217	65.78%	34.22%	58,805	65.85%	34.15%
10/01/10	60,409	52.39%	47.61%	60,272	52.49%	47.51%
11/01/10	61,463	28.73%	71.27%	61,268	28.82%	71.18%
11/29/10	59,405	30.05%	69.95%	59,305	30.10%	69.90%
01/03/11	59,567	28.51%	71.49%	59,520	28.51%	71.49%
Minimum	58,458	14.80%	34.22%	57,846	14.86%	34.15%
Maximum	61,463	65.78%	85.20%	61,268	65.85%	85.14%
Mean	59,571	30.33%	69.67%	59,101	30.43%	69.57%
Stand. Dev.	803	14.89%	14.89%	932	14.89%	14.89%

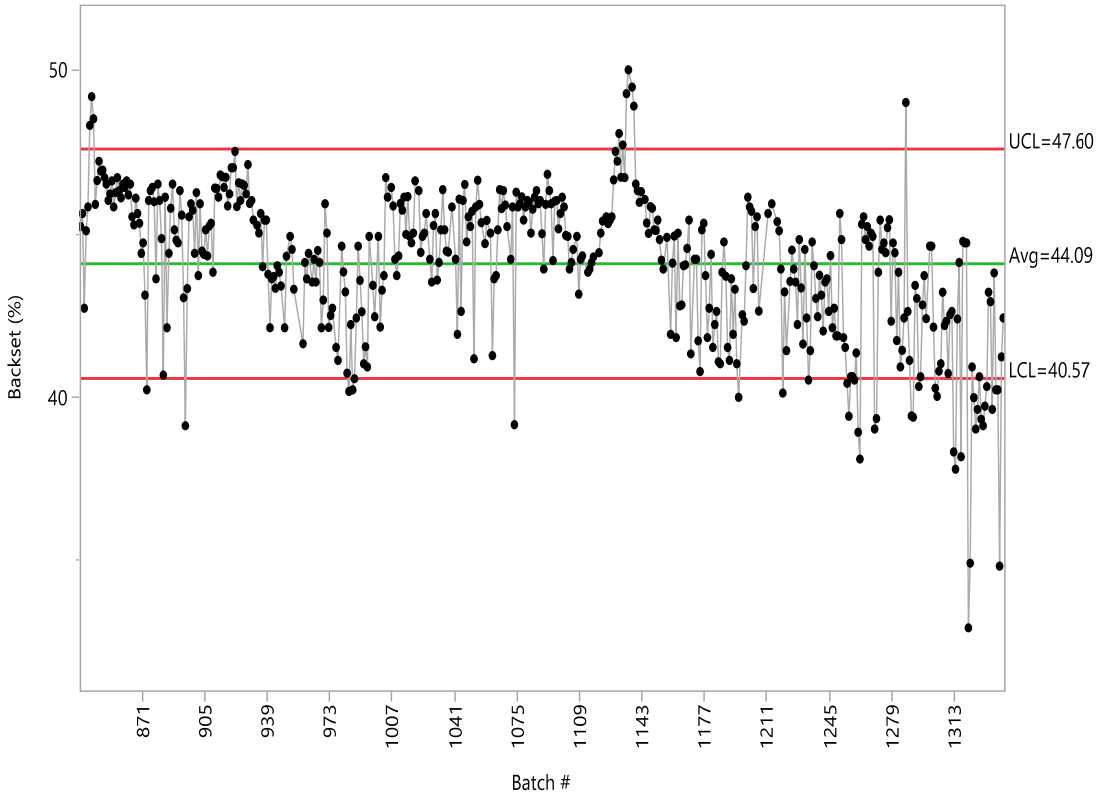
Figure 4.3: Drop Ethanol By Grind Ratio (Corn/Milo) Data, 2010



4.2.2.3 Backset Percentage Data 2010

Backset could change for many reasons. For most of 2010, fermentation batches seemed to have carried a higher than normal backset (Figure 4.4) which lowered towards the end of the year, probably an indicator of plant rates slowed down for winter months or due to lower ethanol margins. Backset percentage data came from an Excel spreadsheet sent from the plant. The data points were originally hand entered by plant operators and may carry inputting errors.

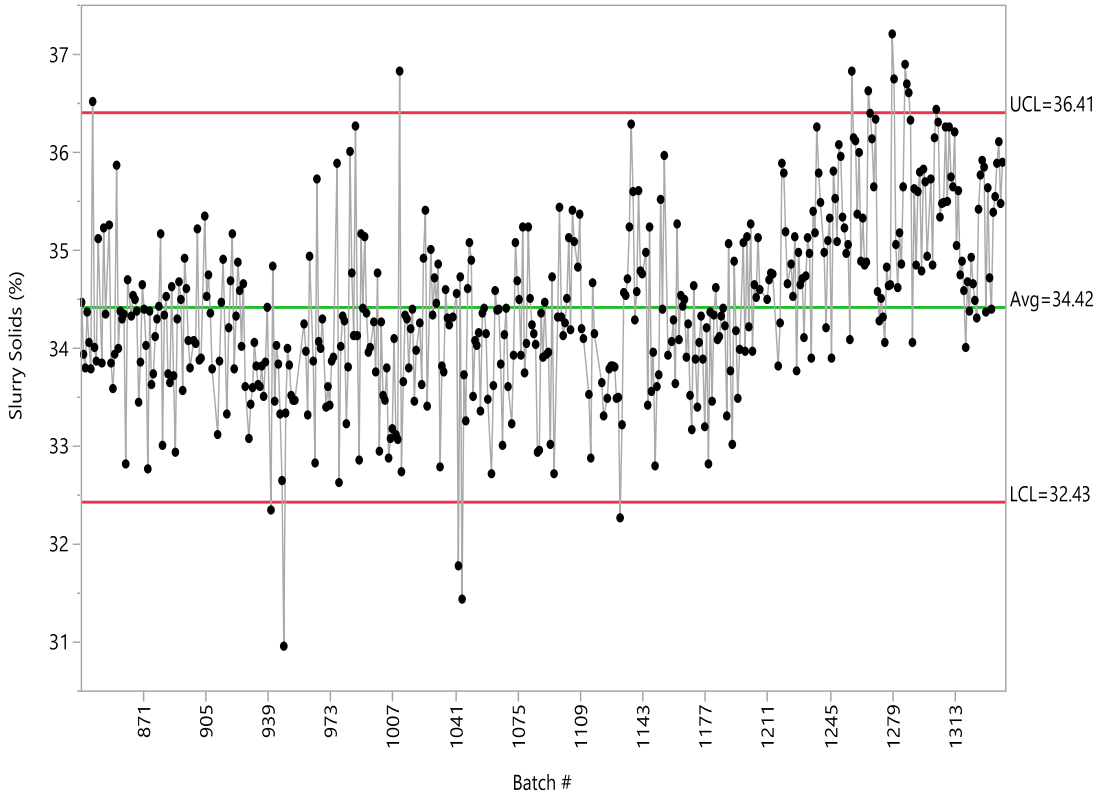
Figure 4.4: Backset % By Batch Number Data, 2010



4.2.2.4 Slurry Solids Data 2010

Overall for the year, slurry solids maintained a 34.42% average. Based on previous lab experience, the slurry solids were measured by plant operators in the ethanol plant lab by NIR (near-infrared), or a microwave type instrument. Slurry solids data came from an Excel spreadsheet sent from the plant. The data points were originally hand entered by plant operators and may carry inputting errors. According to Figure 4.5, the slurry solids had an increasing trend in the last quarter of 2010.

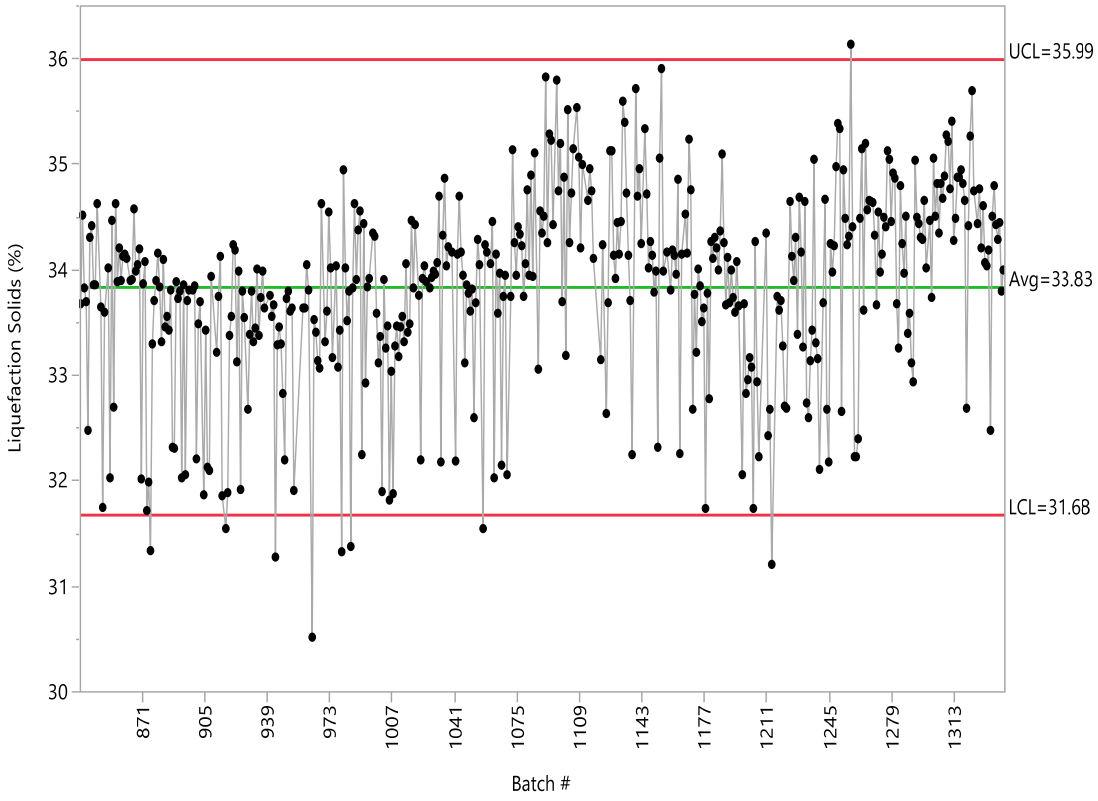
Figure 4.5: Slurry Solids By Batch Number Data, 2010



4.2.2.5 Liquefaction Solids Data 2010

Overall for the year, liquefaction solids maintained a 33.83% average. These solids will normally be slightly lower than the slurry solids. Based on previous lab experience, the liquefaction solids were also measured by plant operators in the ethanol plant lab by NIR (near-infrared), or a microwave type instrument. Liquefaction solids data came from an Excel spreadsheet sent from the plant. The data points were originally hand entered by plant operators and may carry inputting errors. According to Figure 4.6, the liquefaction solids increased in the middle of the year from about the end of June to the beginning of September and then also in the last quarter of 2010. This could have been due to grind ratio changes for fall harvest new crop grain coming in.

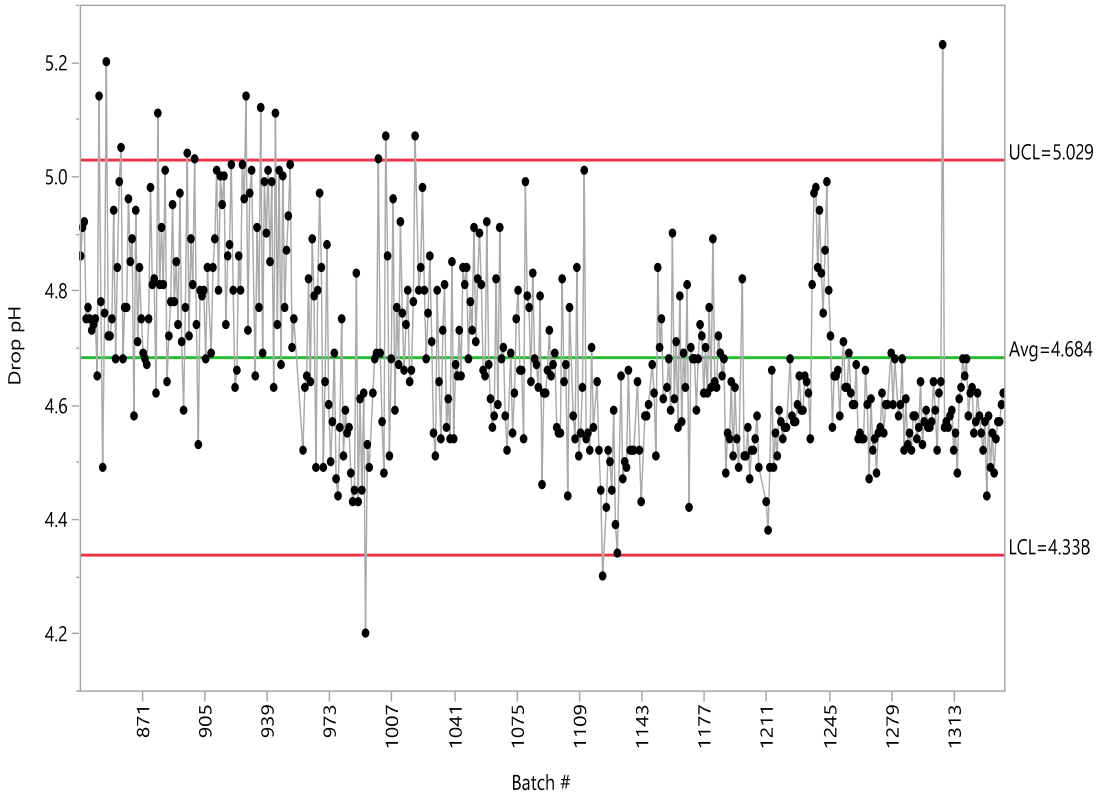
Figure 4.6: Liquefaction Solids By Batch Number Data, 2010



4.2.2.6 Fermentation Drop pH Data 2010

In the lab, pH is measured using a countertop pH probe setup. Plant operators run the fermentation drop sample for pH results. If the probe is not calibrated correctly or it is going out, then the results will be skewed. Drop pH data came from an Excel spreadsheet sent from the plant. The data points were originally hand entered by plant operators and may carry inputting errors. There seemed to have been very variable results the first half of the year, which then became more consistent and controlled later in the year.

Figure 4.7: Drop pH By Batch Number Data, 2010

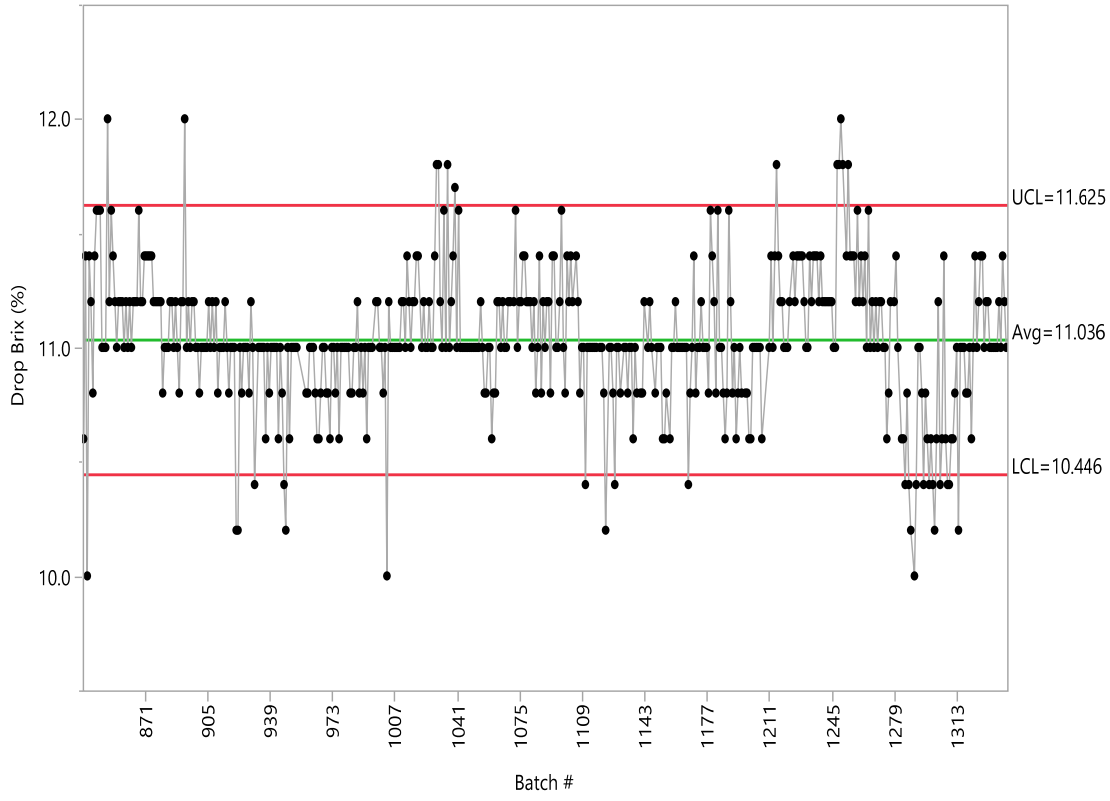


4.2.2.7 Fermentation Drop Brix Data 2010

Today, they are newer and more consistent ways of measuring brix with a digital refractometer, but in this years' worth of data, a handheld version that is held up to the light

was used to interpret the results, which may vary from plant operator to another shift plant operator. Brix percentage data came from an Excel spreadsheet sent from the plant. The data points were originally hand entered by plant operators and may carry inputting errors. Drop brix results did seem to fluctuate some during the year, but had a normal average of 11.036% according to Figure 4.8.

Figure 4.8: Drop Brix (%) By Batch Number Data, 2010

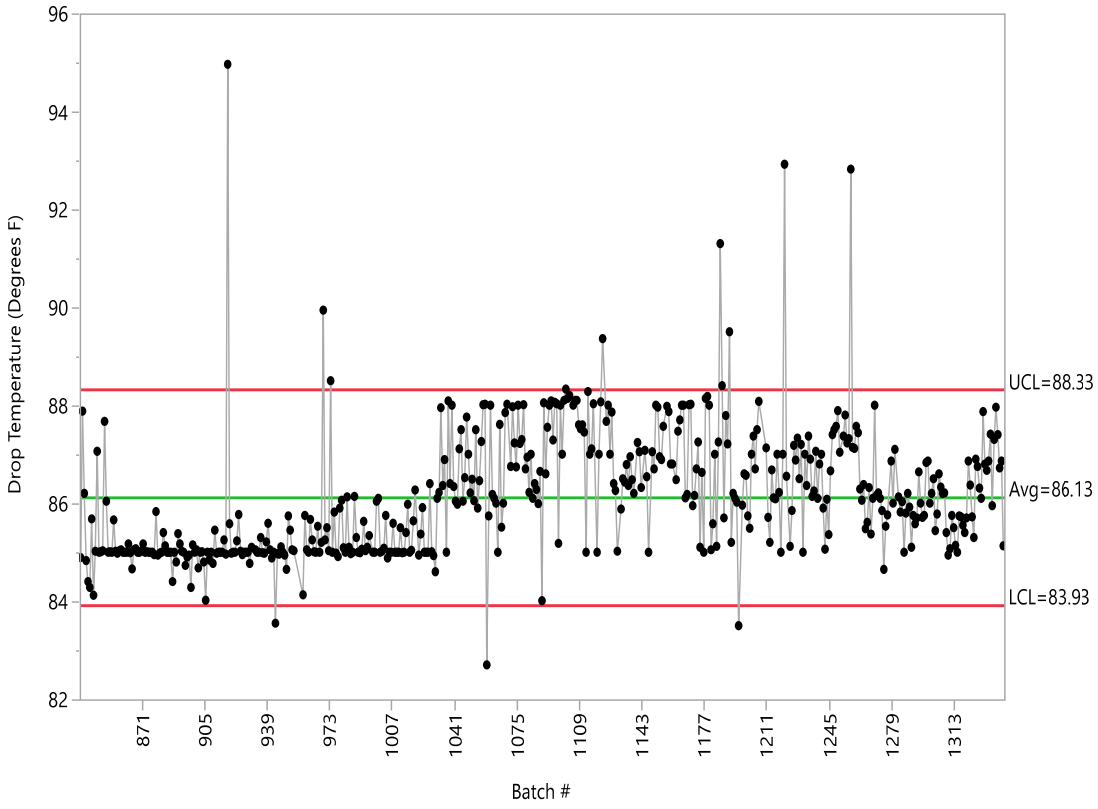


4.2.2.8 Fermentation Drop Temperature Data 2010

Figure 4.9, shows that drop temperature was very consistent the first part of 2010. Around summer time, temperatures seemed to increase and follow a more variable trend. Temperature is read off of temperature probes that are installed in fermentation tanks, so similar to lab probes, they are only as good as their calibration and maintenance. The

operators then record the probes results. Temperature data came from an Excel spreadsheet sent from the plant. The data points were originally hand entered by plant operators and may carry inputting errors.

Figure 4.9: Drop Temperature By Batch Number Data, 2010

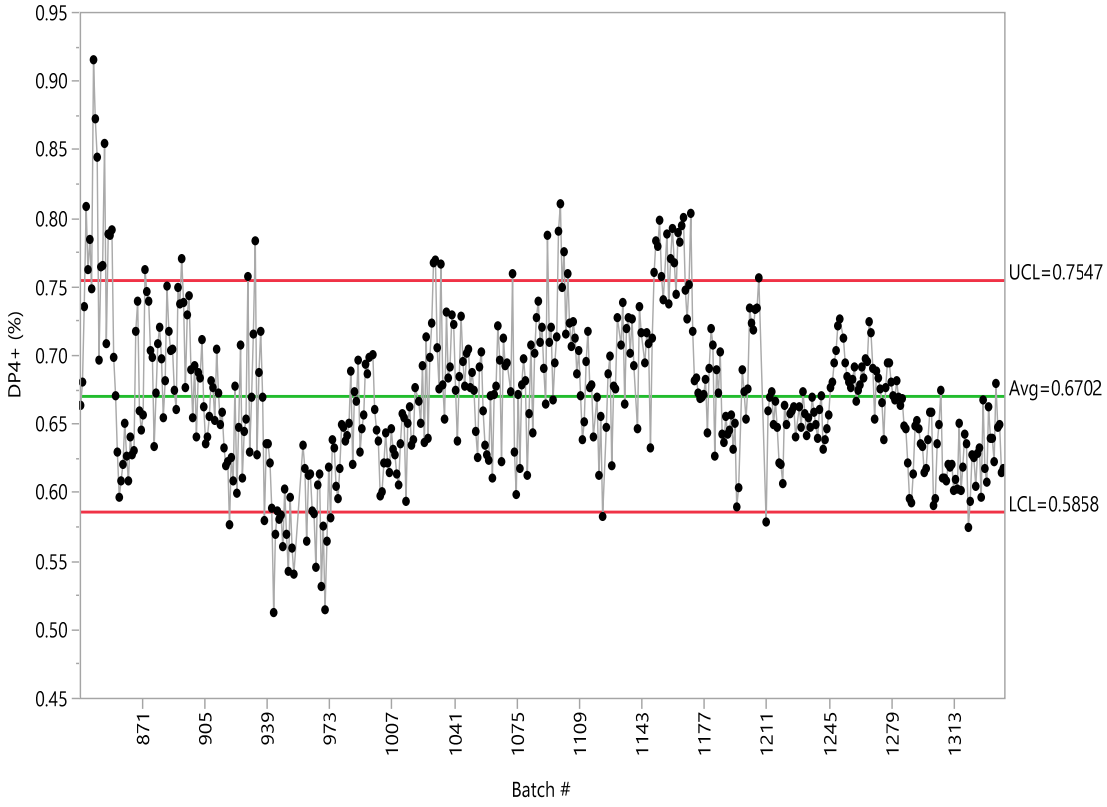


4.2.2.9 Fermentation Drop DP4+ Data 2010

DP4+ percentage is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab filtration procedure and running the sample on the HPLC. Figure 4.10 shows that the HPLC results fluctuated for DP4+ throughout the year, but just like the probes, an HPLC is only as good as its calibration and standards used to calibrate it with. The HPLC does

automatically send the results to the system, from which an Excel spreadsheet was extracted.

Figure 4.10: Drop DP4+ By Batch Number Data, 2010

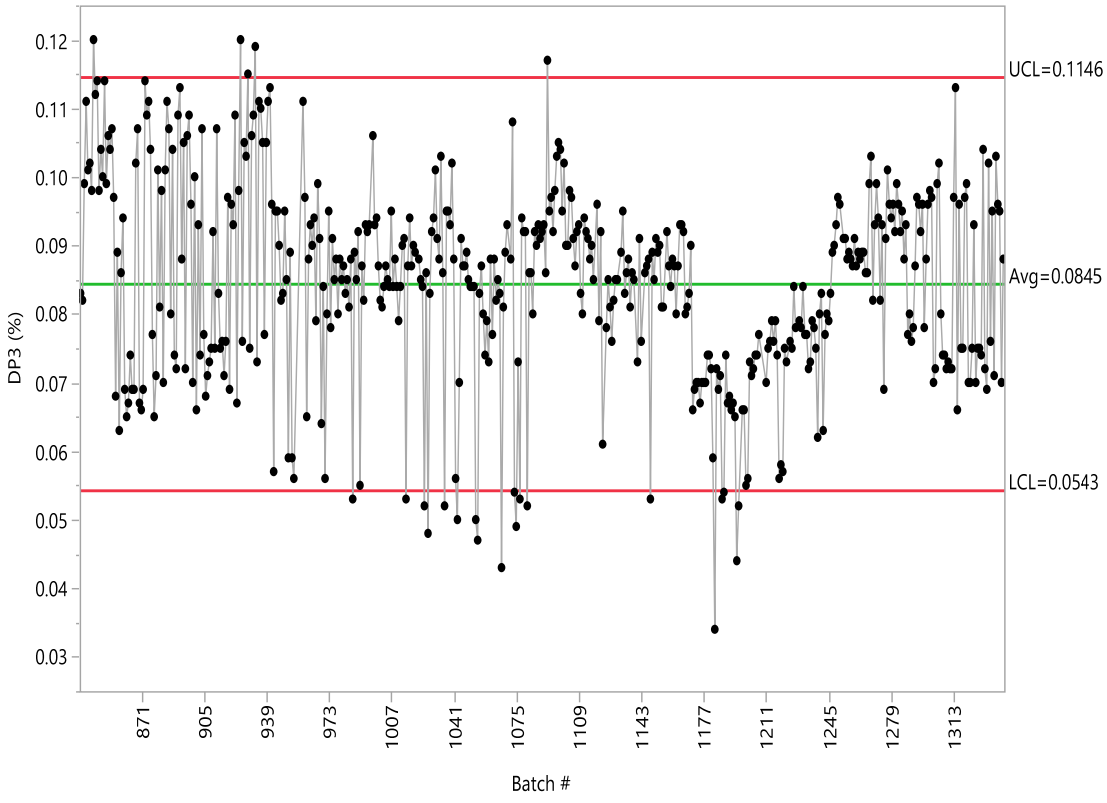


4.2.2.10 Fermentation DP3 Data 2010

DP3 percentage is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab filtration procedure and running the sample on the HPLC. Figure 4.11 shows that the HPLC results fluctuated for DP3 throughout the year, but just like the probes, an HPLC is only as good as its calibration and standards used to calibrate it with. Due to the drastic up and down variable pattern, it could be that a calibration issue is present or that the DP3 component wasn't as calibrated as it should have been on the HPLC, meaning the DP3

sugar peak may not be registering correctly and creating an incorrect or slightly skewed result. It is not known the frequency of calibrations or calibration standards that were used at this time, but it was assumed it was being done correctly to ethanol industry standards. Daily calibration check standards should be ran on the HPLC to determine it is still calibrated. If it is not, then a full calibration should be ran before running anymore fermentation samples. DP3 may also change if the plant is trialing, switching enzymes or any other ingredients, which was information that was not able to be obtained for thesis purposes.

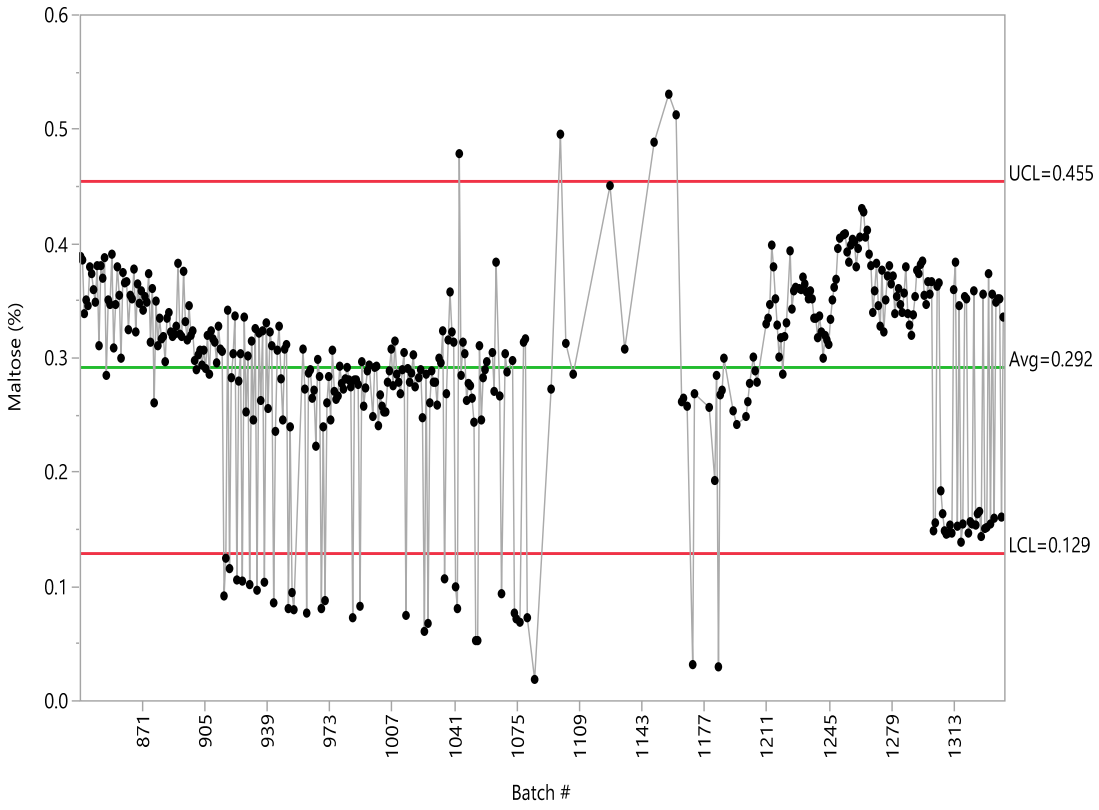
Figure 4.11: Drop DP3 By Batch Number Data, 2010



4.2.2.11 Fermentation DP2 (Maltose) Data 2010

DP2 is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab filtration procedure and running the sample on the HPLC. Figure 4.12 shows that the HPLC results fluctuated for DP4+ throughout the year, but just like the probes, an HPLC is only as good as its calibration and standards used to calibrate it with. It is concluded that the middle section of around batch number 1109 to batch number 1150 was all zero readings meaning the HPLC was not reading these results correctly during these times and may have needed calibration or the detection was not working properly on the instrument itself.

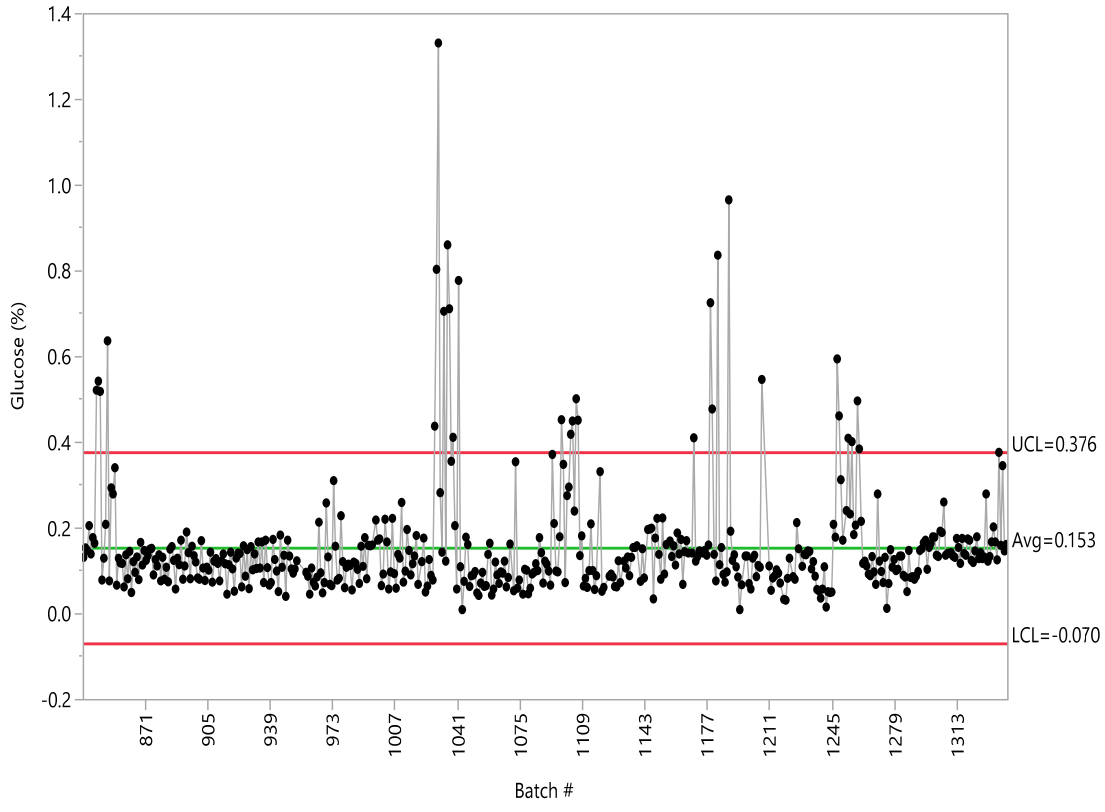
Figure 4.12: Drop DP2/Maltose By Batch Number Data, 2010



4.2.2.12 Fermentation Drop Glucose Data 2010

Glucose is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab filtration procedure and running the sample on the HPLC. Overall, Figure 4.13 shows a rather consistent year for drop glucose. There are random higher spikes, indicating there may have been a lot of different reasons those fermentations did not convert glucose completely, such as infections or temperatures getting too high. The average for 2010 was 0.153% glucose at the end of fermentation.

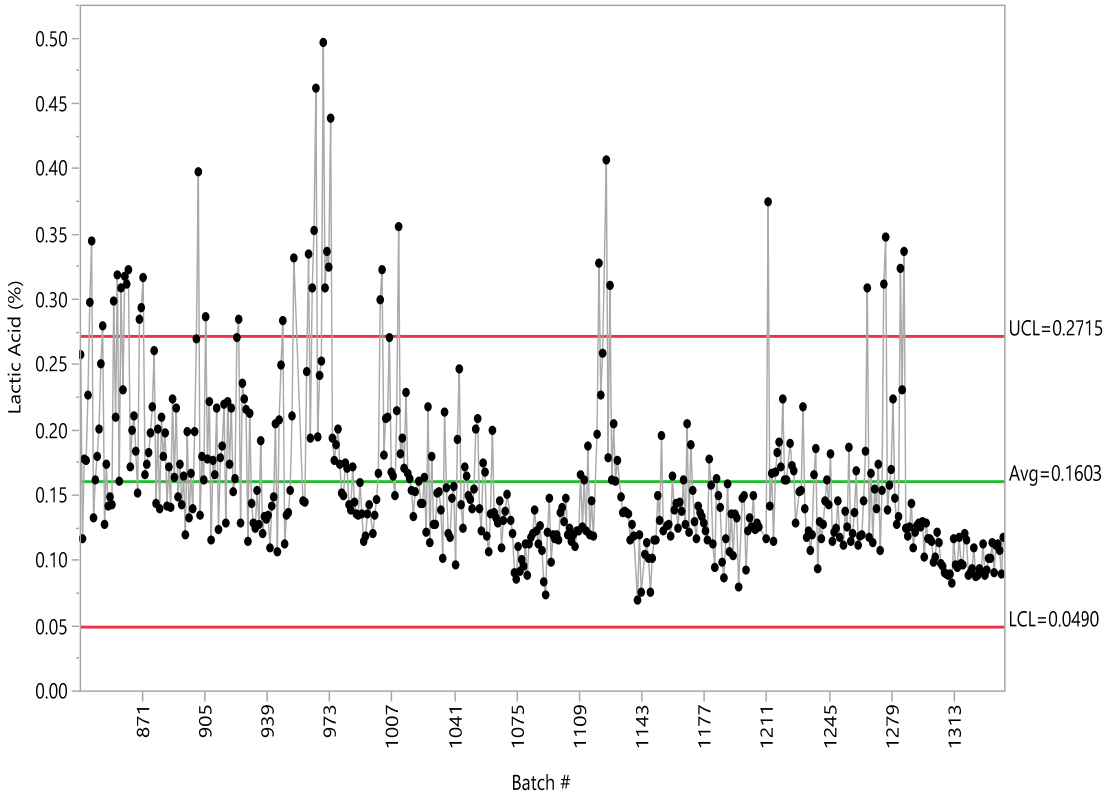
Figure 4.13: Drop Glucose By Batch Number Data, 2010



4.2.2.13 Fermentation Drop Lactic Acid Data 2010

Lactic acid is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab filtration procedure and running the sample on the HPLC. Figure 4.14 shows that lactic acid overall decreased and became more consistent over the year. A decrease in lactic acid is good news, as an increase may show an infection or some sort of negative impact on fermentation.

Figure 4.14: Drop Lactic Acid By Batch Number Data, 2010

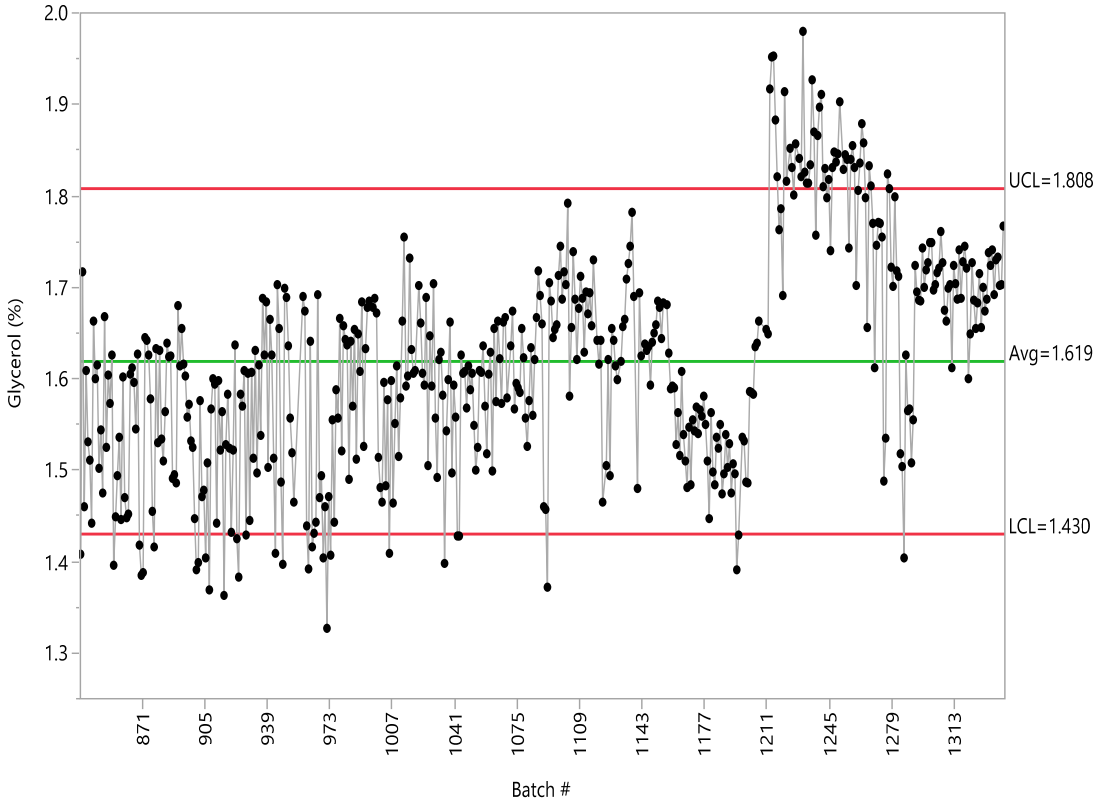


4.2.2.14 Fermentation Drop Glycerol Data 2010

Glycerol is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab

filtration procedure and running the sample on the HPLC. Glycerol was rather steady in 2010, as shown by Figure 4.15. There was some sort of excursion in the latter part of the year, which could have been caused by many different things such as the fall crop coming in from fall harvest or HPLC calibration issues that weren't addressed right away. The plant did seem to be able to get it under control, but it did not go back down to the baseline level during the year.

Figure 4.15: Drop Glycerol By Batch Number Data, 2010

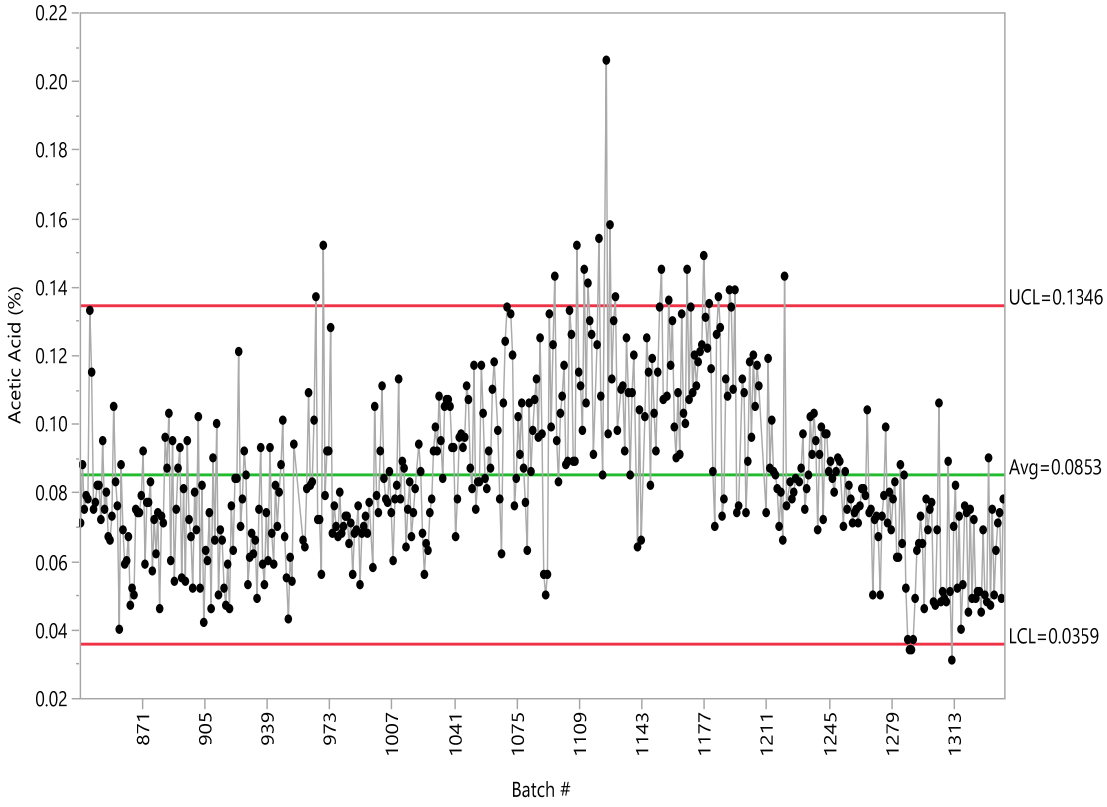


4.2.2.15 Fermentation Drop Acetic Acid Data 2010

Acetic acid is measured from the fermentation drop mash sample collected from an ethanol plant employee. Said employee extracts the liquid from the mash using a lab filtration procedure and running the sample on the HPLC. Like lactic, acetic acid has a

negative effect on ethanol production and should be monitored extensively. Figure 4.16 indicates that there seemed to be an acetic issue in the middle of the year, which may have been caused by warmer weather or HPLC issues. It is more difficult to keep both acetic and lactic acids in calibration because of how low their concentrations are. Both acids have gotten better detection wise in today's industry compared to what they might have been in 2010, nine years ago.

Figure 4.16: Drop Acetic Acid By Batch Number Data, 2010



4.3 Linear Regression Modeling

Regression analysis provides additional understanding for ethanol industry vendors and service suppliers to continue to improve yields and efficiency, as well as, be competitive with their services. Regression modeling allows one to capture the contemporaneous relationships between variables in the ethanol production process. Production type industries are able to generate a significant amount of data points. There are several software programs available for regression modeling. For the multiple linear regression analysis conducted here, JMP was used to estimate the model.

The multiple linear regression estimated was as follows:

$$Ethanol\ Yield\ (\%) = \beta 0 + \beta 1_{AGE} + \beta 2_{MILO} + \beta 3_{BACKSET} + \beta 4_{SLURRY} + \beta 5_{LIQ} + \beta 6_{pH} + \beta 7_{BRIX} + \beta 8_{TEMP} + \beta 9_{DP4+} + \beta 10_{DP3} + \beta 11_{MALTOSE} + \beta 12_{GLUCOSE} + \beta 13_{LACTIC} + \beta 14_{GLYCEROL} + \beta 15_{ACETIC} + u$$

where u is a normally distributed, mean zero IID error term.

4.4 Prediction Analysis

Regression modeling helps to be able create prediction models to be able to predict yields before they take place. To be able to take all the historical data points and be able to conduct accurate prediction analysis is not only useful, but can change the way a production plant, such as an ethanol plant, runs. There are various functional forms that can be used for prediction models.

For prediction, the regression model data was randomly sorted in Excel using the RAND function. The data was then sorted from smallest to largest and was randomly split into two datasets, representing 80% and 20% of the data. The 80% of the dataset was used to estimate the regressions, while the remaining 20% was used to test the out-of-sample

predictive power of the alternative models. All four predictive models were estimated using GRETL software.

Functional forms used for estimation were:

Linear Regression: $Y = b_0 + b_1*X_1 + b_2*X_2 + u$

Semi–Log Model: $\ln Y = b_0 + b_1*X_1 + b_2*X_2 + u$

Double–Log: $\ln Y = b_0 + b_1*\ln(X_1) + b_2*\ln(X_2) + u$

Quadratic: $Y = b_0 + b_1*X_1 + b_2*X_2 + b_{11}*X_1^2 + b_{12}*X_1*X_2 + b_{21}*X_2^2$

Once, all regression models were estimated, the remaining 20% of data was used as a cross validation to test actual ethanol yield versus predicted ethanol yield. The actual values were subtracted from the predicted values to calculate the difference. The differences were then squared and summed. Using the number of observations and the difference sums, I was able to estimate the out-of-sample predictive ability for each model using the root mean square error (RMSE). RMSE shows how far off the prediction was from the actual and the model with the lowest RMSE is the best predictive model. RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{TS}\sum_{i=1}^{T}(Y_{True} - Y_{Predicted})^2}$$

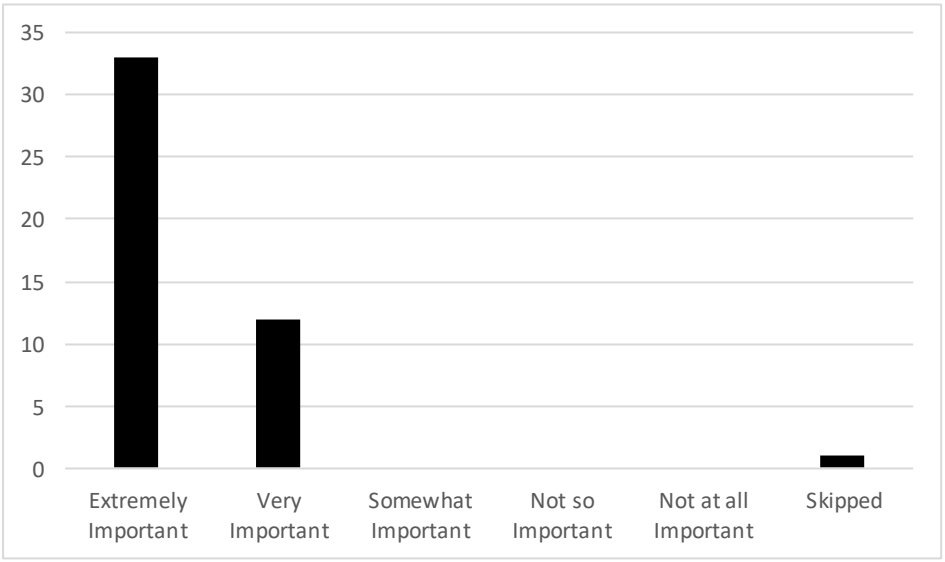
CHAPTER V: RESULTS

Data analysis has been offered from various vendors as a customer service bid or as a product in itself in the industry for many years from ethanol vendors such as Novozymes, Phibro, and Lallemand. Plants look at various data every day, but seemed to rarely be able to fully analyze it due to time and work constraints. Personal experience has showed that more and more plants are taking the time to do this internally to be able to make immediate or large decisions and the survey confirmed this as well. Linear regression and predictive modeling was examined to determine the relationship the different independent variables from the production process might have on fermentation ethanol yield.

5.1 Survey Results

The first question that was asked on the survey was “How would you rate the importance of the plant data analysis in today’s biofuels industry?”. The main intent of the question was to solidify the hypothesis that data analysis is indeed important and something to be looking at on a daily basis.

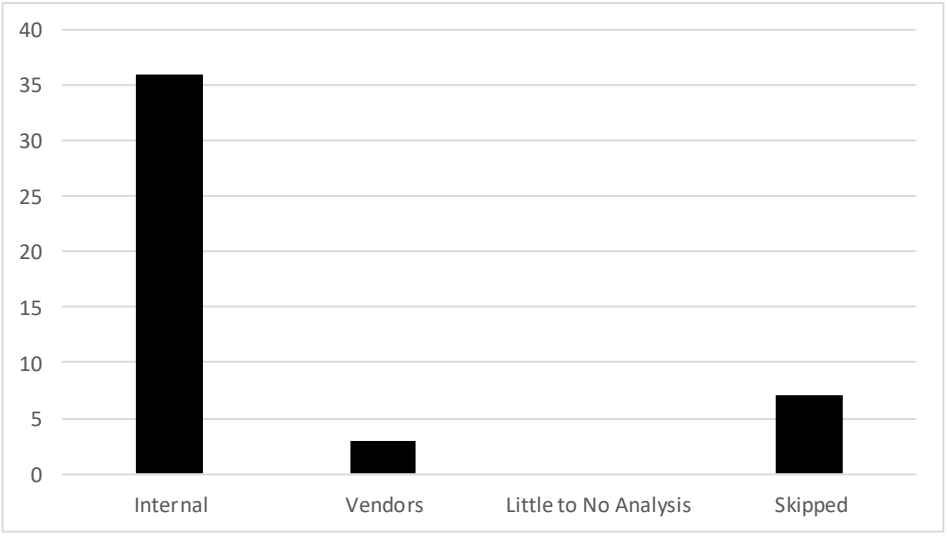
Figure 5.1: Question 1 Responses: "How would you rate the importance of the plant data analysis in today's biofuels industry?"



As quantified in Figure 5.1, besides the one skipped result, 100% of the respondents feel that data is either extremely or very important in the ethanol industry.

This leads to the second question, "Does your plant have an internal employee(s) analyzing plant data, rely on vendors, or use little to no analysis?".

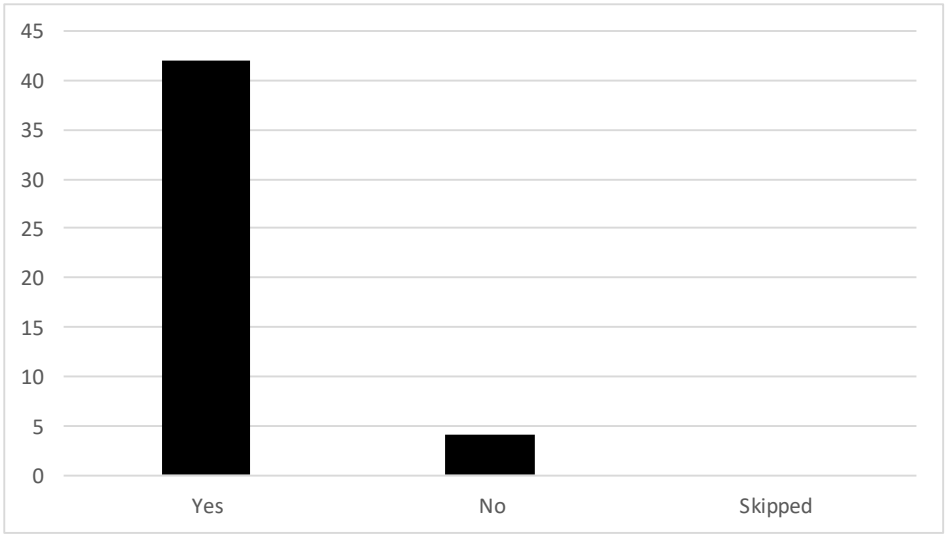
Figure 5.2: Question 2 Responses: "Does your plant have an internal employee(s) analyzing plant data, rely on vendors, or use little to no analysis?"



Making plant changes is not only a large decision, it may also be a costly one. This is a main reason, in Figure 5.2, that plants have resulted in using more internal data analysis to confirm, deny, or back up decision-making. This is a possible explanation, but may not be definitive.

The third question on the survey was, “Have you made large impacting plant decisions based on results from data analysis? If Yes, please provide examples in comments if willing to share.”.

Figure 5.3: Question 3 Responses: “Have you made large impacting plant decisions based on results from data analysis?”



Based off the Figure 5.3, some of the comments that stood out with the plants who did make decisions based on data include: “~$1,000,000 in annualized enzyme savings by just analyzing data!”; “We use data to drive almost every decision that is made within our organization. For example: When trialing new products, you compare production data (baseline vs trial) to measure trial success. If the production data suggest better performance, then we have to look at the financial data and see how it impacts our

conversion cost."; and "We do all new product trials in house and don't use vendor data. Every decision we make, large and small, is dependent on the data analysis.".

Question 4, "Does your plant use Excel or another program for data analysis? (For example, JMP)" was written to understand what software plants are using for analysis tools. A popular software among vendors is the SAS program, JMP. Excel was used heavily prior, but there are more predictive modeling options when using JMP.

Figure 5.4: Question 4 Responses: "Does your plant use Excel or another program for data analysis? (For example, JMP)"

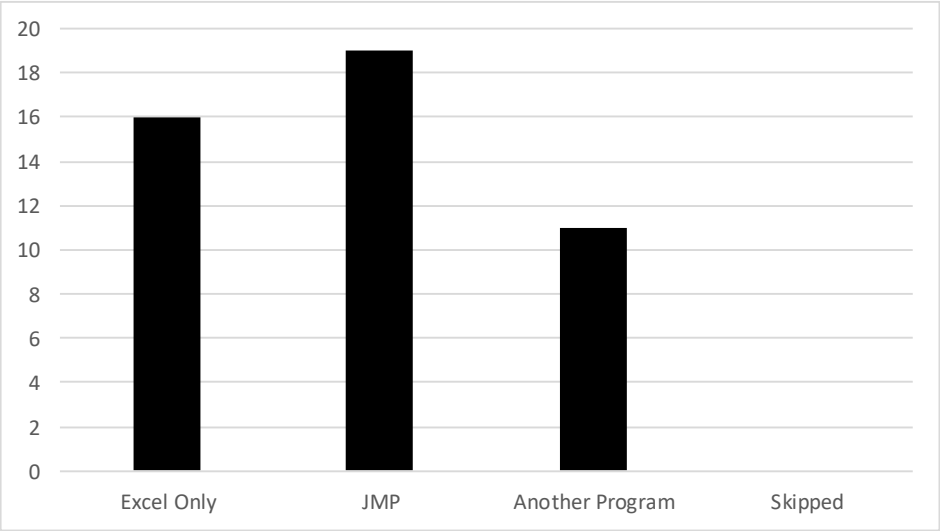
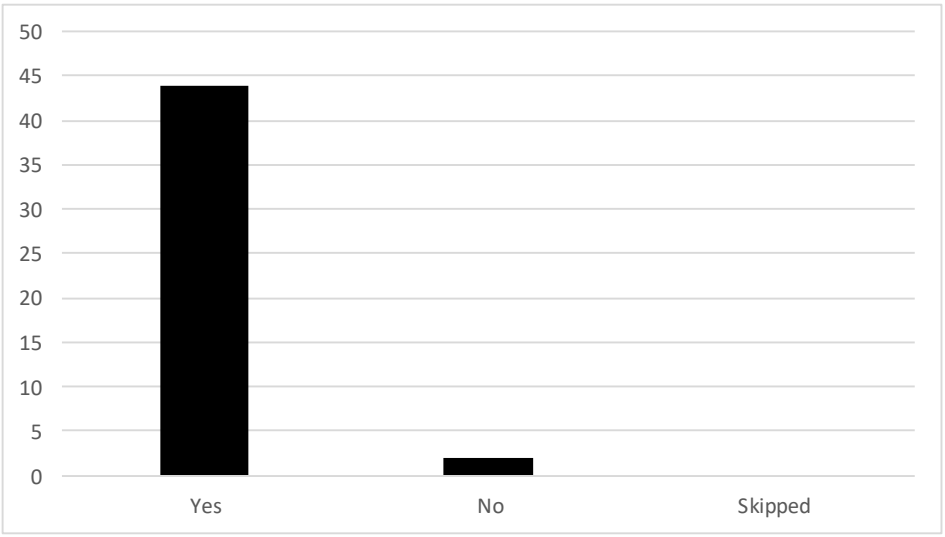


Figure 5.4, is surprising to see how many plants don't invest in newer, advanced software. Analysis can be a lot faster in software programs other than Excel. Excel can be time consuming sometimes. Based off personal experience, plant personnel have time constraints.

Question 5 was, "Do you see the industry as a whole continuing to use data analysis in the future?". There was 100% response to yes on this question, so no graph was depicted.

Question 6 was, “Would you find predictive modeling for ethanol yield useful as long as the modeling is relatively accurate?”. This question was used to examine if predictive modeling research and analysis for the industry would be valuable for the future of the ethanol industry.

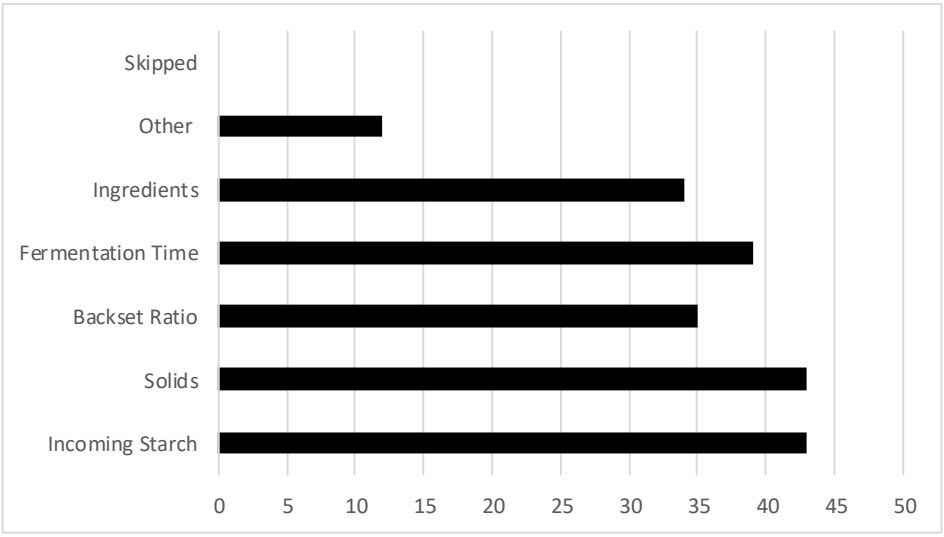
Figure 5.5: Question 6 Responses: “Would you find predictive modeling for ethanol yield useful as long as the modeling is relatively accurate?”



In Figure 5.5, it is encouraging to see that the industry overall believes that if accuracy is able to be upheld, that all the work going into predictive analytics currently could be an innovation for the industry. All comments that were associated with that question all recognize this as a benefit if the data is accurate.

Question 7 was, “What inputs do you consider important while looking at the output of ethanol yield?”.

Figure 5.6: Question 7 Responses: "What inputs do you consider important while looking at the output of ethanol yield?"



Given the high responses for some of these options in Figure 5.6, including them as variables was important in regression analysis.

Question 8 was, "Does your plant have additional control/instrumentation systems beyond the DCS? (For example, Trident, Direct Automation, Pavilion, DataParc, etc.)". DCS is the data control system used in the control room of an ethanol plant. The results from this question are questionable due to lack of information and knowledge when the survey was developed, so no analysis was conducted here.

Question 9 was, "Do you plan to purchase new systems in the near future?", was developed to confirm if plants would be willing to purchase new technology or software if market conditions improved.

Figure 5.7: Question 9 Responses: "Do you plan to purchase new systems in the near future?"

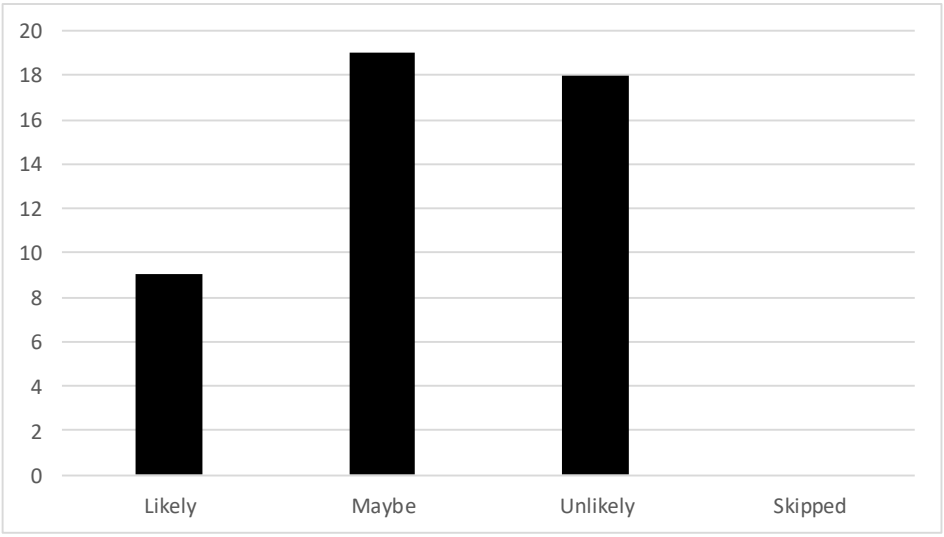


Figure 5.7, depicts that the industry will likely or maybe purchase newer systems or technology in the future, assuming the high number of unlikely responses is related to the poor ethanol markets the industry has endured over the past 2 years. It would be interesting to poll this question again under different ethanol market conditions.

Question 10 was, "Overall, how proactive is your plant when it comes to new technology that is available for the industry? (It can be big bolt on technologies, DCS and instrumentation technologies, Maintenance technologies, Safety technologies, etc.)", examines if a plant is willing to make a plant change that may change the process significantly.

Figure 5.8: Question 10 Responses: "Overall, how proactive is your plant when it comes to new technology that is available for the industry?"

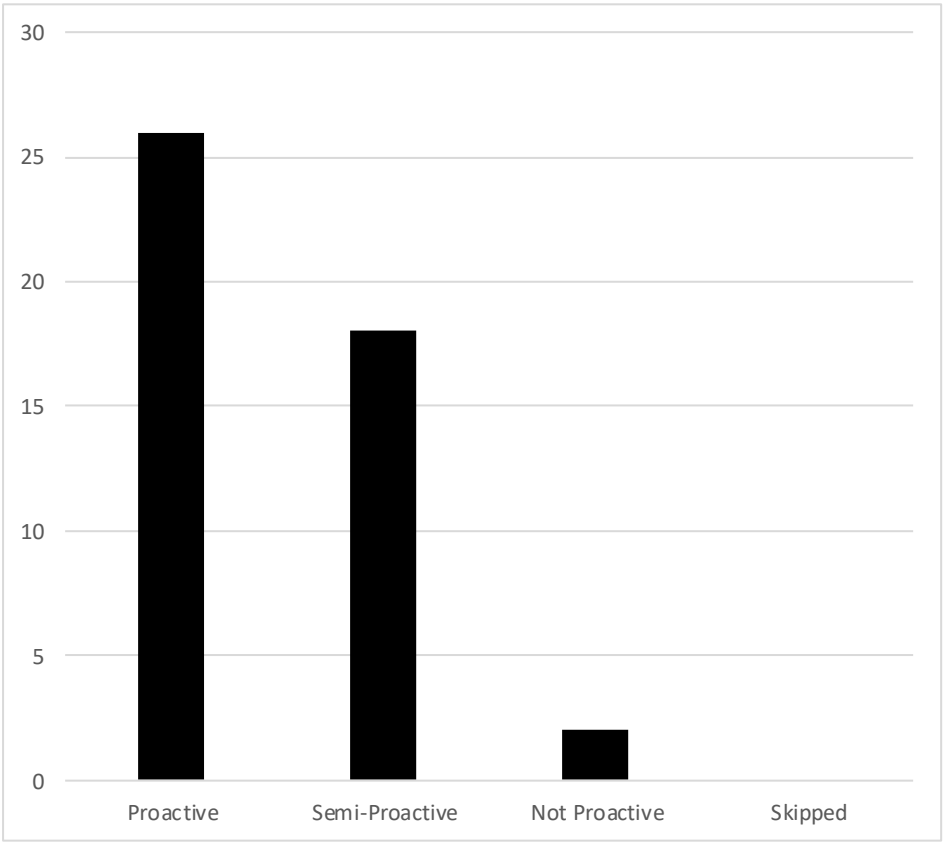


Figure 5.8 confirms, that overall, the ethanol industry is willing to make changes for innovation and to improve efficiency.

5.2 Ethanol Yield Linear Regression

The linear regression model was developed using the plant data from Plant ABC that was summarized in the monthly averages in Table 4.1. Results from the ethanol yield linear regression model are provided in Table 5.1. The independent variables of Ratio Milo, Drop pH, Drop DP4+, Drop Glucose, Drop Lactic Acid, and Drop Acetic Acid have *p*-values < 0.01 (i.e. they are statistically significant at a 1% level of significance) (Figure 5.1), meaning there is a very strong correlation with the dependent variable ethanol yield. The independent variables of Fermentation Age, Drop Brix, and Drop Glycerol are

statistically significant at a 5.0% level of significance (Figure 5.1), meaning there is correlation with the dependent variable ethanol yield. Lastly, the independent variables of Backset, Slurry Solids, Liquefaction Solids, Drop Temperature, DP3, and Maltose have *p*-values >0.10 (Figure 5.1), meaning there is a little to no correlation with the dependent variable ethanol yield (i.e. they are not statistically different from zero). In the last column of Figure 5.1 is the standardized beta, which represents each independent variables relative importance. "The larger the absolute value of the standardized beta coefficient, the more important the variable" (Klimberg and McCullough 2016, p.95). The *p*-value for the *F* test is <0.0001, indicating that one or more of the fifteen independent variables is significantly correlated with ethanol yield.

"Each independent variable regression coefficient represents an estimate of the change in the dependent variable to a unit increase in that independent variable while all the other independent variables are held constant" (Klimberg and McCullough 2016, p.94). Based on this the interpretation of the slope coefficients on the independent variables is as follows:

- Fermentation Age – Ethanol yield may increase by 0.0036424 and may be expected to result from a unit increase in fermentation drop age, holding all other variables constant.
- Ratio Milo – Ethanol yield may increase by 0.0057071 and may be expected to result from a unit increase in ratio milo, holding all other variables constant.
- Backset – Ethanol yield may increase by 0.0031789 and may be expected to result from a unit increase in backset percentage, holding all other variables constant.

- Slurry Solids – Ethanol yield may increase by 0.0139931 and may be expected to result from a unit increase in slurry solids percentage, holding all other variables constant.
- Liquefaction Solids – Ethanol yield may decrease by 0.013676 and may be expected to result from a unit increase in liquefaction solids percentage, holding all other variables constant.
- Drop pH – Ethanol yield may increase by 0.6272752, which may be expected to result from a unit increase in drop pH, holding all other variables constant.
- Drop Brix – Ethanol yield may increase by 0.1244393 and may be expected to result from a unit increase in drop brix percentage, holding all other variables constant.
- Drop Temperature – Ethanol yield may decrease by 0.010327 and may be expected to result from a unit increase in fermentation drop temperature in degrees Fahrenheit, holding all other variables constant.
- Drop DP4+ – Ethanol yield may increase by 3.3567161 and may be expected to result from a unit increase in drop DP4+ percentage, holding all other variables constant.
- Drop DP3 – Ethanol yield may decrease by 2.430599 and may be expected to result from a unit increase in drop DP3 percentage, holding all other variables constant.
- Drop Maltose – Ethanol yield may decrease by 0.001593 and may be expected to result from a unit increase in drop maltose percentage, holding all other variables constant.

- Drop Glucose – Ethanol yield may decrease by 0.898206 and may be expected to result from a unit increase in drop glucose percentage, holding all other variables constant.
- Drop Lactic Acid – Ethanol yield may increase by 1.325188 and may be expected to result from a unit increase in drop lactic acid percentage, holding all other variables constant.
- Drop Glycerol – Ethanol yield may decrease by 0.391335 and may be expected to result from a unit increase in drop glycerol percentage, holding all other variables constant.
- Acetic Acid – Ethanol yield may decrease by 4.255121 and may be expected to result from a unit increase in drop acetic acid percentage, holding all other variables constant.

Referring back to the theoretical ethanol yield model in Table 3.1, six out of the fifteen variables did not meet the expected signs of their relationship with ethanol yield. Backset resulted in a positive coefficient and did not have a negative impact on ethanol during the time period examined. Backset percentage can be very variable depending on the process at a given time. Liquefaction solids resulted in a negative coefficient. With the large quantities of milo being used during most of 2010, it is possible that the liquefaction solids were not dialed in to reach the target percentage, having a negative impact on ethanol yield. Drop pH tends to be more controlled during the production process and it was no surprise that it had a positive relationship with ethanol yield. Drop Brix can fluctuate for unknown reasons and it was unexpected to see an overall positive relationship for an entire

year's worth of data. Drop DP4+ had large percentage variation throughout the year. Experience suggests it can have a negative impact on ethanol yield, and in this case it did not. DP4+ had a significant number of peaks registering on the HPLC. It is hard to know why the peaks occurred without further analysis. In this case, it is likely these peaks resulted in the positive relationship found. Lactic acid is an organic acid and in too high of concentrations it is is very inhibitory to higher levels of ethanol yield. For most of 2010, lactic acid levels varied up and down and were never consistent until the very end of the year. While it was expected there would be a negative relationship between lactic acid and ethanol yield, a positive relationship resulted and further analysis would be needed to determine why.

Table 5.1: Ethanol Yield Linear Regression, 2010

Term	Coefficient	Std Error	*t* -Ratio	*p* -Values	Std Beta
const	8.9731477	1.47289	6.09	<.0001*	0
Ferm Age	0.0036424	0.001619	2.25	0.0251*	0.096278
Ratio Milo	0.0057071	0.001711	3.34	0.0009*	0.208092
Backset	0.0031789	0.007565	0.42	0.6746	0.019644
Slurry Solids	0.0139931	0.017708	0.79	0.43	0.034314
Liq Solids	-0.013676	0.016348	-0.84	0.4034	-0.03282
Drop pH	0.6272752	0.106316	5.9	<.0001*	0.265147
Drop Brix	0.1244393	0.051985	2.39	0.0172*	0.114666
Drop Temp	-0.010327	0.013191	-0.78	0.4343	-0.0347
DP4+	3.3567161	0.403871	8.31	<.0001*	0.496934
DP3	-2.430599	1.630186	-1.49	0.1369	-0.09752
Maltose	-0.001593	0.224392	-0.01	0.9943	-0.00038
Glucose	-0.898206	0.116992	-7.68	<.0001*	-0.35419
Lactic Acid	1.325188	0.334903	3.96	<.0001*	0.229471
Glycerol	-0.391335	0.173746	-2.25	0.0250*	-0.12858
Acetic Acid	-4.255121	0.993123	-4.28	<.0001*	-0.22972
Number of Observations = 504					
$R^2 = 0.558196$					

5.3 Regression Predictive Models

The following section examines each of the functional forms for the predictive analysis described in section 4.4.

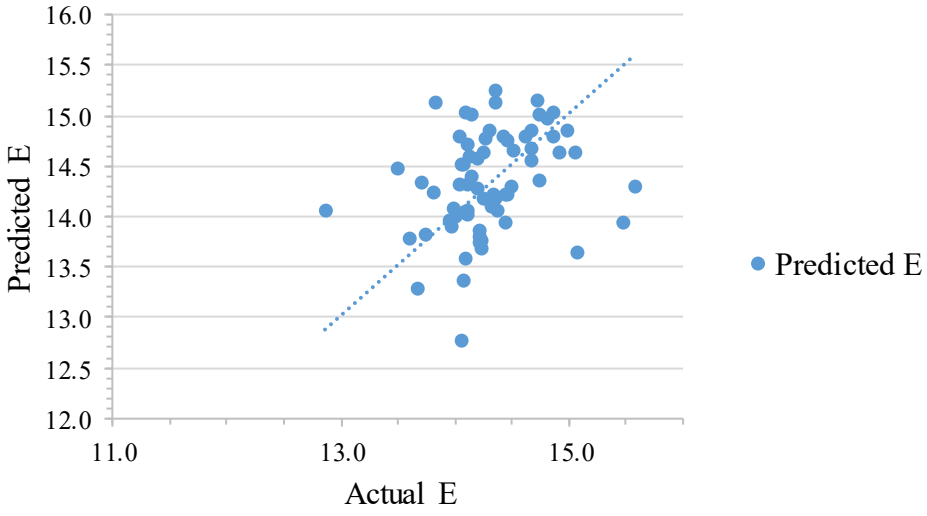
5.3.1 Linear Predictive Model

Using 80% of the randomly selected data points, table 5.2 presents the estimation results for the linear predictive model estimated with 402 observations. Fermentation Age, Ratio Milo, Drop DP4+, Drop DP3, Drop Glucose, Drop Lactic Acid, Drop Glycerol, and Drop Acetic Acid were all significant at a 1% level of significance. Figure 5.11 shows the comparison of the out-of-sample predictive power of the model using remaining 20% of the data.

Table 5.2: Linear Predictive Model, 2010

Term	Coefficient	Std. Error	*t* -Ratio	*p* -Value	
const	−0.822384	0.399317	−2.059	0.0401	**
FermAge	0.0209484	0.00540963	3.872	0.0001	***
RatioMilo	0.0182547	0.00317316	5.753	<0.0001	***
Backset	−0.00142220	0.00332005	−0.4284	0.6686	
SlurrySolids	0.00204775	0.00555199	0.3688	0.7125	
LiqSolids	−0.00100358	0.00499596	−0.2009	0.8409	
DroppH	0.273997	0.210508	1.302	0.1938	
DropBrix	0.0126099	0.0337277	0.3739	0.7087	
DropTemp	0.00736396	0.00855568	0.8607	0.3899	
DP4	9.6939	0.766617	12.65	<0.0001	***
DP3	−16.8521	3.03437	−5.554	<0.0001	***
Maltose	−0.111828	0.0754503	−1.482	0.1391	
Glucose	−1.15923	0.317918	−3.646	0.0003	***
LacticAcid	5.66694	0.500802	11.32	<0.0001	***
Glycerol	3.16713	0.335725	9.434	<0.0001	***
AceticAcid	−5.01729	1.54385	−3.250	0.0013	***
Number of Observations = 402					
$R^2 = 0.933461$					

Figure 5.9: Linear Model Actual Ethanol vs. Predictive Ethanol



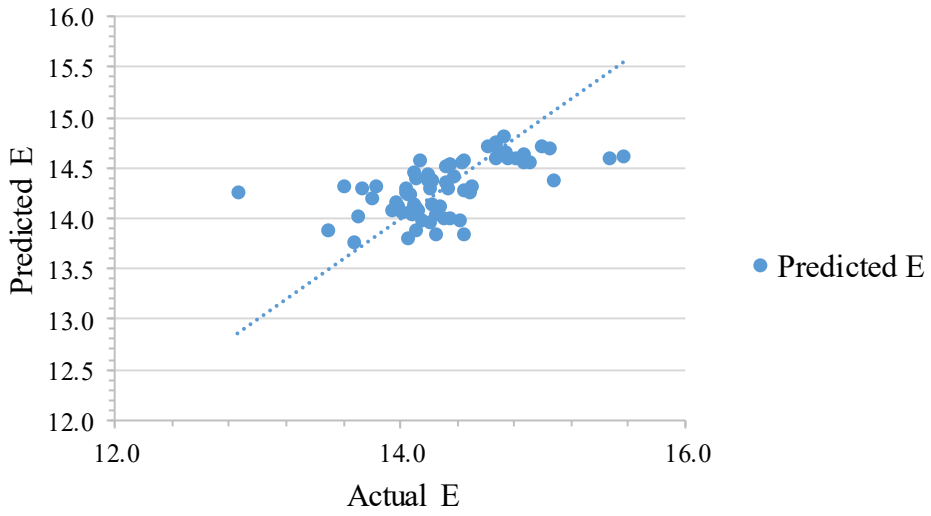
5.3.2 Semi-Log Predictive Model

Using 80% of the randomly selected data points, figure 5.12 presents the results for the Semi-Log Predictive model estimated with 402 observations, with 9 observations dropped due to missing data. Fermentation Age, Ratio Milo, Drop DP4+, Drop Glucose, Drop Lactic Acid, Drop Glycerol, and Drop Acetic Acid were all statistically significant at the 1% level of significance. Figure 5.13 shows the comparison of the out-of-sample predictive power of the model using remaining 20% of the data.

Table 5.3: Semi-Log Predictive Model, 2010

Term	Coefficient	Std. Error	*t* -Ratio	*p* -Value	
const	2.52473	0.0290176	87.01	<0.0001	***
FermAge	0.000339409	0.00010839	3.131	0.0019	***
RatioMilo	0.000550546	9.97E-05	5.523	<0.0001	***
Backset	0.000251327	0.00014087	1.784	0.0752	*
SlurrySolids	4.25E-05	0.00021574	0.197	0.844	
LiqSolids	−0.000213640	0.000192	−1.113	0.2665	
DroppH	0.00559119	0.0055867	1.001	0.3176	
DropBrix	0.00113525	0.00085812	1.323	0.1867	
DropTemp	−0.000215193	0.00029582	−0.7274	0.4674	
DP4	0.201372	0.0209059	9.632	<0.0001	***
DP3	−0.131304	0.0999411	−1.314	0.1897	
Maltose	0.00728803	0.00833574	0.8743	0.3825	
Glucose	−0.0482434	0.00643224	−7.500	<0.0001	***
LacticAcid	0.0910548	0.0213413	4.267	<0.0001	***
Glycerol	−0.0350937	0.0111576	−3.145	0.0018	***
AceticAcid	−0.321101	0.058819	−5.459	<0.0001	***
Number of Observations = 393					
R^2 = 0.507417					

Figure 5.10: Semi-Log Model Actual Ethanol vs. Predictive Ethanol



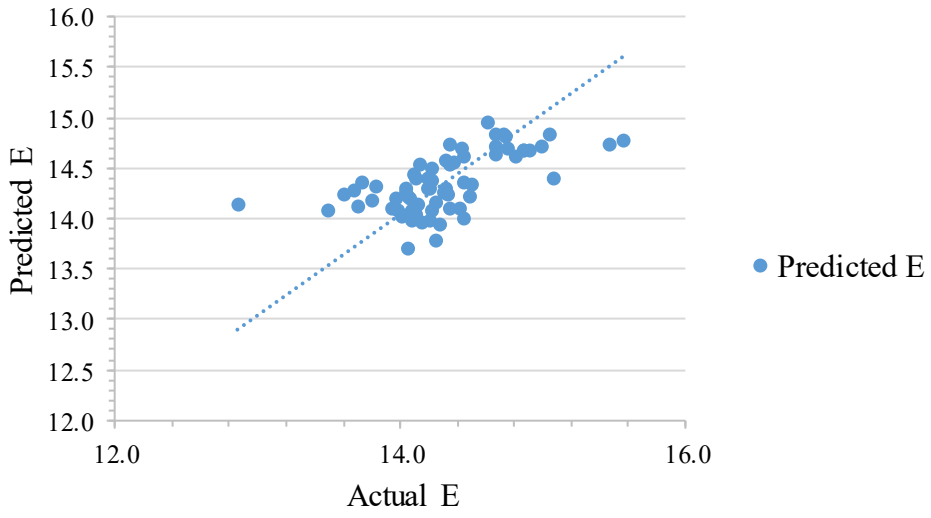
5.3.3 Double-Log Predictive Model

Using 80% of the randomly selected data points, figure 5.14 presents the estimation results for the Double-Log Predictive Model estimated with 402 observations, with 123 dropped because of missing or incomplete data. lnRatio Milo, lnDrop pH, lnDrop DP4+, lnDrop Glucose, lnDrop Lactic Acid, and lnDrop Acetic Acid were statistically significant at the 1% level of significance. Figure 5.15 shows the comparison of the out-of-sample predictive power of the model using remaining 20% of the data.

Table 5.4: Double-Log Predictive Model, 2010

Term	Coefficient	Std. Error	*t* -Ratio	*p* -Value	
const	2.44192	0.542758	4.499	<0.0001	***
lnFermAge	0.0216526	0.0105764	2.047	0.0416	**
lnRatioMilo	0.0217615	0.00802787	2.711	0.0072	***
lnBackset	0.0213831	0.0232825	0.9184	0.3592	
lnSlurrySolids	−0.000661499	0.041147	−0.01608	0.9872	
lnLiqSolids	−0.0105790	0.043615	−0.2426	0.8085	
lnDroppH	0.172249	0.0376984	4.569	<0.0001	***
lnDropBrix	0.0117447	0.0541672	0.2168	0.8285	
lnDropTemp	−0.0623964	0.113006	−0.5521	0.5813	
lnDP4	0.147254	0.0192087	7.666	<0.0001	***
lnDP3	−0.00215558	0.0114894	−0.1876	0.8513	
lnMaltose	−0.00284923	0.00340201	−0.8375	0.4031	
lnGlucose	−0.00621492	0.00231395	−2.686	0.0077	***
lnLacticAcid	0.0216141	0.00452191	4.78	<0.0001	***
lnGlycerol	−0.0358737	0.0196552	−1.825	0.0691	*
lnAceticAcid	−0.0300725	0.00739641	−4.066	<0.0001	***
Number of Observations = 279					
R^2 = 0.543118					

Figure 5.11: Double-Log Model Actual Ethanol vs. Predictive Ethanol



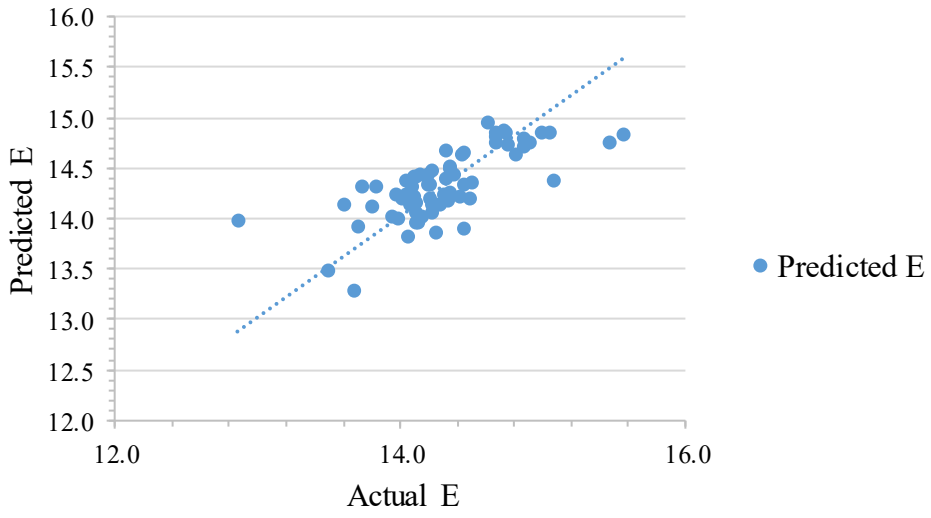
5.3.4 Quadratic Predictive Model

Using 80% of the randomly selected data points, figure 5.16 presents the estimation results for the Quadratic Predictive Model estimated with 402 observations with 123 dropped because of missing or incomplete data. Ratio Milo, Drop Glucose, and lnRatio Milo were statistically significant at the 1% level of significance.

Table 5.5: Quadratic Predictive Model, 2010

Term	Coefficient	Std. Error	*t* -Ratio	*p* -Value	
const	199.795	231.515	0.863	0.389	
FermAge	0.00075846	0.00482004	0.1574	0.8751	
RatioMilo	0.037947	0.011712	3.24	0.0014	***
Backset	0.317482	0.180515	1.759	0.0799	*
SlurrySolids	−0.581137	0.614463	−0.9458	0.3452	
LiqSolids	0.180471	0.831943	0.2169	0.8284	
DroppH	0.0444397	3.57892	0.01242	0.9901	
DropBrix	−2.09141	1.18236	−1.769	0.0782	*
DropTemp	0.731954	0.687431	1.065	0.288	
DP4	−1.81935	3.62486	−0.5019	0.6162	
DP3	2.71295	7.76467	0.3494	0.7271	
Maltose	0.817799	0.631178	1.296	0.1963	
Glucose	−1.26777	0.221536	−5.723	<0.0001	***
LacticAcid	1.31966	1.28177	1.03	0.3042	
Glycerol	5.38901	2.99728	1.798	0.0734	*
AceticAcid	0.512198	4.01816	0.1275	0.8987	
lnFermAge	0.337417	0.417471	0.8082	0.4197	
lnRatioMilo	−1.89455	0.663304	−2.856	0.0047	***
lnBackset	−13.3722	7.67973	−1.741	0.0829	*
lnSlurrySolids	20.2121	21.2401	0.9516	0.3422	
lnLiqSolids	−6.00053	27.7897	−0.2159	0.8292	
lnDroppH	2.41254	16.8002	0.1436	0.8859	
lnDropBrix	23.8824	13.36	1.788	0.0751	*
lnDropTemp	−64.4154	60.7989	−1.059	0.2904	
lnDP4	3.49335	2.45587	1.422	0.1562	
lnDP3	−0.398331	0.605192	−0.6582	0.511	
lnMaltose	−0.166934	0.0969395	−1.722	0.0863	*
lnGlucose	0.122662	0.0497742	2.464	0.0144	**
lnLacticAcid	0.0143599	0.238962	0.06009	0.9521	
lnGlycerol	−9.05259	4.93283	−1.835	0.0677	*
lnAceticAcid	−0.441504	0.347222	−1.272	0.2047	
Number of Observations = 279					
$R^2 = 0.638176$					

Figure 5.12: Quadratic Model Actual Ethanol vs. Predictive Ethanol



5.3.5 Predictions Summary

The root mean square error (RMSE) was estimated for each model for out-of-sample prediction comparisons. Results are presented in Table 5.6. The quadratic predictive model (Table 5.5) had the lowest RMSE indicating the best predictive model. The quadratic RMSE was 0.3002. The double-log and semi-log predictive models were not that far behind in out-of-sample predictive performance.

Table 5.6: RMSE Results Prediction Models, 2010

Model	RMSE
Linear	0.560056717
Semi-Log	0.350227401
Double-Log	0.339503963
Quadratic	0.300233359

CHAPTER VI: CONCLUSIONS AND FUTURE RESEARCH

6.1 Conclusions

The purpose of this thesis is to examine the importance and impact of data analytics in ethanol production and the value predictive modeling of ethanol yield can have for an ethanol plant. The objectives were to survey the industry, identify if plant changes were being made on the plan level based on data analytics, formulate simple regressions, and develop predictive regressions. Estimating multiple linear regression can help identify variables that have strong correlations for ethanol yield. Regression modeling may lead to prediction models that can help identify ways to fine-tune processes throughout the ethanol plant, helping to improve plant efficiency. The independent variables of ratio milo, Drop pH, Drop DP4+, Drop Glucose, Drop Lactic Acid, and Drop Acetic Acid were all statistically significant, pointing to processes that may help improve ethanol production and confirming that in 2010 there was a very strong correlation between these indicators and ethanol yield. Out of the four predictive regression models that were estimated, the quadratic model was the most accurate and best fit model for Plant ABC 2010 data set for prediction analysis.

Many of the data points that are being collected on a daily basis at the ethanol plant level can be analyzed beyond control chart trends to investigate current plant conditions for ways to optimize and better improve ethanol plant efficiencies. I believe there is some value in exploring milo dosage and its effect on ethanol yield as well as fermentation age. Sugars and the organic acids concentrations in fermentation will always be a work in progress to continually make them lower at the end of the fermentation process. Regression and predictive modeling exemplify advancing steps during the ethanol production process

that could have economic value for plant managers. These models could provide valuable diagnostic information in addition to their expertise in managing their plants. These type of analyses can provide plant managers with weekly reports of plant optimal efficiency. Overall, it may be difficult when dealing with so many variables at once, but statistical and analytical tools can help to provide opportunities to examine many facets of the ethanol production process and bring value-added modeling to ethanol plant customers.

6.2 Future Research

Being a production process, there are more areas of an ethanol plant to utilize regression analysis and other data analytic tools. Co-products could be another area to focus analysis and regression on. It might be worth analyzing Plant ABC's newer data sets to include the latest technologies and how those might compare to a traditional set of data that was used. Plant ABC's data representing year 2010 was from before their installment of corn oil extraction. If an updated model was estimated with newer plant data there could be differences in efficiency, depending where in the production process these technical advances have occurred. There is potential to look at this type of ethanol plant modeling in a sequential, step-by-step process as the production process advances. Future research can be used to conduct analysis for other ethanol plants, as well. There are opportunities to add in additional alternative functional form models such as the translog predictive model for predictive assessments. Another prospect, could be to focus solely on the fermentation step variables and possibly even do this for a full round of each fermenter or to help reduce bias from a specific fermenter. For example, most plants have around four fermenters and one of them may be performing at a lower level due to infection, a pipe leak, or some other

reason, resulting in the poorer performance. Data analytics may provide a way to identify this.

WORKS CITED

2010. "Confidential Data from Plant ABC." *Ethanol Plant Data.*

Contributors, Wikipedia. 2019. *High-performance liquid chromatography.* March 1. Accessed March 13, 2019. https://en.wikipedia.org/w/index.php?title=High-performance_liquid_chromatography&oldid=885686915.

Coward-Kelly, Guillermo. 2011. "Ethanol Producer Magazine." *ethanolproducer.com.* July 21. Accessed March 2, 2019. http://ethanolproducer.com/articles/7990/helping-ethanol-producers-operate-in-the-undefinedsweet-spotundefined.

Ge, Zhiqiang. 2017. *Review on data-driven modeling and monitoring for plant-wide industrial processes.*

Jessen, Holly. 2014. "Ethanol Producer Magazine ." *ethanolproducer.com.* December 15. Accessed January 14, 2019. http://ethanolproducer.com/articles/11708/putting-data-to-work.

Klimberg, Ron, and B.D. McCullough. 2016. *Fundamentals of Predictive Analytics with JMP, Second Edition.* Cary, NC: SAS Institute Inc.

n.d. *Renewable Fuels Association.* Accessed April 4, 2019. https://ethanolrfa.org/how-ethanol-is-made/.

Sadati, Najibesadat, Ratna Babu Chinnam, and Milad Zafar Nezhad. 2017. *Observational data-driven modeling and optimization of manufacturing processes.*

Urbanchuk, John M. 2017. "CONTRIBUTION OF THE ETHANOL INDUSTRY TO THE ECONOMY OF THE UNITED STATES IN 2017." Prepared for the Renewable Fuels Association, Doylestown, 15. Accessed February 2019. https://ethanolrfa.org/wp-content/uploads/2018/02/RFA-2017-Ethanol-Economic-Impact-01_28_17_Final.pdf.

APPENDIX A

A.1 Survey Description

Once again, the ten question survey was conducted as a proof of concept to confirm the significance that data and data analysis still played in plants. It is more important than ever and is being observed by both internal employees and sourced out to vendors. The survey was also conducted to see what all was software and technology that might being utilized. The survey was anonymous, meaning all responses may be traced by computer IP address, but do not know the name, position, or plant the information came from. Survey was conducted through Survey Monkey and analyzed using Excel and it had 46 responses total.

A.2 Survey Questions (As Conducted)

1. How would you rate the importance of the plant data analysis in today's biofuels industry?

- ☐ Extremely important
- ☐ Very important
- ☐ Somewhat important
- ☐ Not so important
- ☐ Not at all important

2. Does your plant have an internal employee(s) analyzing plant data, rely on vendors, or use little to no analysis?

- ☐ Internal
- ☐ Vendors
- ☐ Little to No Analysis

Other (please specify)

3. Have you made large impacting plant decisions based on results from data analysis? If Yes, please provide examples in comments if willing to share.

- ☐ Yes
- ☐ No

4. Does your plant use Excel or another program for data analysis? (For example, JMP)

- [ ] Excel Only
- [ ] JMP
- [ ] Another Program

5. Do you see the industry as a whole continuing to use data analysis in the future?

- [ ] Yes
- [ ] No

6. Would you find predictive modeling for ethanol yield useful as long as the modeling is relatively accurate?

- [ ] Yes
- [ ] No

7. What inputs do you consider important while looking at the output of ethanol yield?

- [ ] Incoming Starch
- [ ] Solids
- [ ] Backset Ratio
- [ ] Fermentation Time
- [ ] Ingredients
- [ ] Other (please specify)

8. Does your plant have additional control/instrumentation systems beyond the DCS? (For example, Trident, Direct Automation, Pavilion, DataParc, etc.)

- [ ] Yes, Please Comment Below Which Ones
- [ ] No

9. Do you plan to purchase new systems in the near future?

- [ ] Likely
- [ ] Maybe, if market conditions get better
- [ ] Unlikely

10. Overall, how proactive is your plant when it comes to new technology that is available for the industry? (It can be big bolt on technologies, DCS and instrumentation technologies, Maintenance technologies, Safety technologies, etc.)

- ☐ Proactive
- ☐ Semi-Proactive
- ☐ Not Proactive