Using Big Data for Decisions in Agricultural Supply Chain

by

Derik Lafayette Smith

B.S. Automotive Technology Management and Supply Chain
Brigham Young University-Idaho, 2010

and

Satya Prakash Dhavala

Master of Business Administration
Indian Institute of Science, 2005

B.Tech. Electronics and Communications
Jawaharlal Nehru Technological University, 2002

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Master of Engineering in Logistics

at the

Massachusetts Institute of Technology

June 2013

Signature of Authors..

Master of Engineering in Logistics Program, Engineering Systems Division
May 10, 2013

Certified by... ....

Dr. Bruce C. Arntzen
Executive Director, Supply Chain Master's Program,
MIT Center for Transportation and Logistics
Thesis Supervisor

Accepted by..

Prof. Yossi Sheffi
Elisha Gray II Professor of Engineering Systems, MIT
Director, MIT Center for Transportation and Logistics
Professor, Civil and Environmental Engineering, MIT

Using Big Data for Decisions in Agricultural Supply Chain

by

Derik Smith and Satya Dhavala

Submitted to the Engineering Systems Division
on May 10, 2013 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Logistics

Abstract

Agriculture is an industry where historical and current data abound. This paper investigates the numerous data sources available in the agricultural field and analyzes them for usage in supply chain improvement. We identified certain applicable data and investigated methods of using this data to make better supply chain decisions within the agricultural chemical distribution chain. We identified a specific product, AgChem, for this study. AgChem, like many agricultural chemicals, is forecasted and produced months in advance of a very short sales window. With improved demand forecasting based on abundantly-available data, Dow AgroSciences, the manufacturer of AgChem, can make better production and distribution decisions. We analyzed various data to identify factors that influence AgChem sales. Many of these factors relate to corn production since AgChem is generally used with corn crops. Using regression models, we identified leading indicators that assist to forecast future demand of the product.

We developed three regressions models to forecast demand on various horizons. The first model identified that the price of corn and price of fertilizer affect the annual, nation-wide demand for the product. The second model explains expected geographic distribution of this annual demand. It shows that the number of retailers in an area is correlated to the total annual demand in that area. The model also quantifies the relationship between the sales in the first few weeks of the season, and the total sales for the season. And the third model serves as a short-term, demand-sensing tool to predict the timing of the demand within certain geographies. We found that weather conditions and the timing of harvest affect when AgChem sales occur.

With these models, Dow AgroSciences has a better understanding of how external factors influence the sale of AgChem. With this new understanding, they can make better decisions about the distribution of the product and position inventory in a timely manner at the source of demand.

Thesis Supervisor: Dr. Bruce C. Arntzen
Title: Executive Director, Supply Chain Master's Program, Center for Transportation and Logistics

Acknowledgements

We both thank Dr. Bruce Arntzen and Josh Merrill (Dow Agro Sciences) for their valuable inputs throughout this effort. We also thank Thea Singer for assisting us in our writing efforts.

We also thank our respective families for their support, encouragement, and love during this period.

Table of Contents

List of figures

List of tables

1 Introduction

Our thesis investigates how to make better supply chain decisions using the rich data available in the agricultural industry. We begin with an analysis of various types of agricultural data and the agricultural supply chain. We then describe the role of our sponsor company in the agriculture distribution chain and present the problem they face. Subsequently, we explain the process followed to analyze the problem and identify solutions.

1.1 Agricultural Data

As the world's computing capability exponentially increases, our ability to gather, maintain, and analyze data also increases dramatically. Data is more abundant in nearly every industry and academic realm. The agricultural industry is no different. In some senses, there is more data in the agricultural industry than in most others. Agriculture is one of the world's oldest occupations and trades. For millennia, farming practices have been passed on and improved upon. For centuries, agricultural data has been collected, tracked, and analyzed.

Agricultural data ranges from crop yields in certain geographic areas to erosion, weather, and climate trends. This data is collected, analyzed, and in most cases, shared by a variety of sources: private companies, governmental agencies, and research universities. Some of this data is published in journals, available through web-based databases or sold by the data collection sources. This data varies in scope and detail; some data sources are very granular while others are more general.

In companies that produce agricultural products, employee teams of R&D engineers and researchers gather data and monitor trends to improve seeds, herbicides, pesticides, fertilizers, and agricultural equipment. The information and knowledge gained by these companies is generally patented, guarded, and used competitively within the marketplace. However, the practices of data sharing and knowledge transfer are becoming more widely accepted within some trade organizations and supplier/customer partnerships.

Additionally, government agencies and bureaus such as the United States Department of Agriculture (USDA), amass large amounts of data. The USDA consists of 17 agencies, each responsible for monitoring and assisting a different facet of the US agriculture industry. Some of these agencies include the Agricultural Research Service (ARS), the Farm Service Agency (FSA), and the National Agricultural Statistics Service (NASS). Each of these agencies gathers and tracks masses of data which relate to its specific responsibilities. Most of this data and research is published in reports and is publically available from the USDA or on its website. Universities and other research centers also add to the overwhelming amount of agricultural research and data. Land Grant universities have been called "the state-based component of the public agricultural research system" (National Research Council, 1995). These universities research a variety subjects in an effort to advance the knowledge and practices within the agricultural field.

1.1.1 Applications of this Data

Once data is collected, it is used by a variety of people and organizations to improve the effectiveness of today's agricultural industry. Individual growers use crop, weather, and soil data to make decisions about their upcoming growing season and long-term sustainability of their farms. Companies who supply seed, chemicals, and other farming necessities monitor this data to create demand forecasts and production plans. Crop insurance companies monitor factors such as climate, yield, and farm economic data to constantly fine-tune their policies.

Much of this data is widely used throughout the industry. However, there are some who feel that better operational and supply chain decisions could be made through careful compilation and analysis of particular segments of this data. Our sponsor company feels this way.

1.2 Agricultural Supply Chain

When people hear the word "agriculture," they most likely think of large fields of corn in Iowa, citrus groves in Florida, or apple orchards in Washington. Each of these scenarios is at the grower level. However, many other players are involved in the value chain of these vital products. They include producers of seeds, fertilizers, chemicals, and agricultural equipment as well as the retail organizations which market and sell these products.

Our principal interest is in the challenges faced by agricultural chemical producers. In this section, we will elaborate on those parts of the supply chain which are relevant to this sector of the agricultural industry.

1.2.1 Product Flow through the Agricultural Chemical Supply Chain

The supply chain for agricultural chemicals is large and complex. Manufacturers purchase various components from a multi-echelon supplier base. Active and non-active ingredients are combined in specific manufacturing sites. Often, one site will manufacture a chemical for distribution globally. Manufacturing sites often produce a variety of chemicals and must plan production based on forecasts and seasonality of demand. If the product is sold in bulk, it may be shipped to a manufacturer-owned terminal where it is stored until it is shipped to a distributor. If the product is to be sold in smaller quantities, it is sent to a re-packer who packages the chemical in 250 gallon totes, 2.5 gallon jugs, or other suitable retail packs for liquid applications and bags or envelopes for dry applications. After being packed, the product is sent to a distributor.

Most manufacturers do not sell directly to growers or retailers, but instead sell through major distributors. These distributors market products to retailers across the country. Retailers hold bulk or packaged product and sell it to individual growers for application on fields and crops. Figure 1 gives a representation of the product flow through this supply chain.

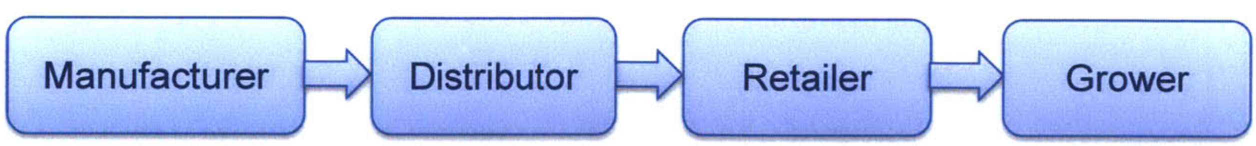

Figure 1 Agricultural chemical product flow

1.2.2 Information Flow through the Supply Chain

Traditionally, manufacturers have only received information about their sales and shipments to distributors. Distributors have been very protective of information about their sales and customers because they have feared that the manufacturers could circumvent them and sell directly to the retailers. However, in recent years, due to manufacturers' request and investment in technologies such as electronic data interchange (EDI), distributors and retailers have been sharing information with the manufacturers in an effort to improve supply availability and reduce distribution costs. This information, represented in Figure 2, includes inventory levels at distributor facilities, point-of-sale (POS) data from retailers, and grower demographics.

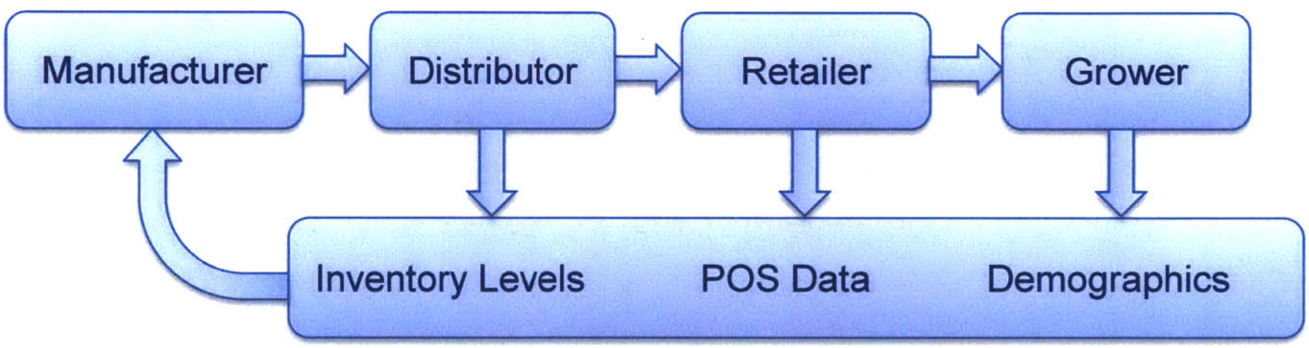

Figure 2 Agricultural chemical information flow

1.2.3 Our Sponsor Company

Dow AgroSciences (DAS) is the sponsor of our project. On their website, they explain their background in the following way:

> *We began as the plant sciences business of The Dow Chemical Company, which was founded in 1897. A joint venture with Eli Lilly and Company in 1989 created DowElanco. With a focus on sustainable agriculture, DowElanco combined the leading chemistries of The Dow Chemical Company with those of the agricultural division of Eli Lilly. In 1997, The Dow Chemical Company acquired full ownership of the business and named the business Dow AgroSciences. Today, we employ more than 7,700 people worldwide, and our 2012 global sales were $6.4 billion (U.S.)* (About Us).

DAS offers a variety of products to the agricultural industry. These products include "insecticides, herbicides, fungicides, fumigants, nitrogen stabilizers, seeds, traits, and oils" (Products).

1.3 Problem Statement

We intend to identify and analyze data which can be used by DAS to understand the changes in demand for a specific product and make better-informed decisions throughout their supply chain. For this analysis, we have selected a product called AgChem. AgChem is primarily used on corn, thus, we will focus on those factors associated with its production, such as corn prices and climate. Our intention is to identify potential influences on the demand of AgChem among the profusion of agricultural data.

1.3.1 AgChem – The Selected Product

AgChem is an agricultural chemical and is manufactured in two formulations, each under a separate name. AgChem is generally used with the production of corn, and has been produced and sold by the company for many years. This product is available in two varieties – AgChem V1 and AgChem V2, with V1 being applied predominantly during fall and V2 during spring season.

Demand for AgChem had stabilized and was believed to have matured. However, in recent years, the demand has grown significantly. Figure 3 shows the monthly sales over the past five years. As we can see there is a sudden and steep increase in sales. The company is interested in knowing what factors have influenced this trend and what to expect for future demand.

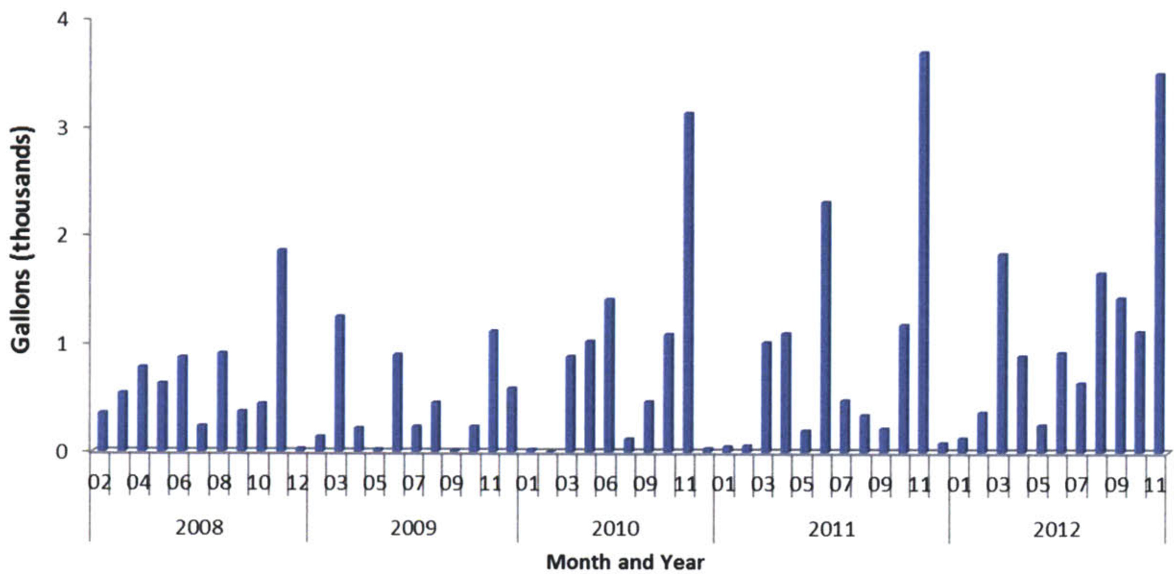

Figure 3 Increasing demand over the last 5 years

The market for agricultural chemicals and fertilizers is seasonal and has a long production and distribution lead times. Most of these products, including AgChem, are only applied during very specific periods in a crop's lifecycle. The demand for each product can vary significantly from year to year depending on the mix of crops planted, weather and climatic conditions, farm economics, as well as many other factors. Thus, companies like DAS must forecast and produce chemicals months in advance for a short period of uncertain demand.

1.4 Thesis Outline

This section presented agricultural data, analyzed the agricultural chemical supply chain, introduced our sponsor company, and stated the problem we have sought to solve. Next, in our Literature Review, we will expand on terminology pertinent to this study and cover some research that has been conducted with big data and agriculture forecasting. In Chapter 3, Methodology we describe the analytical techniques we used to evaluate factors driving the demand. Chapter 4 presents the data, our analysis, the models that were developed and the results. We conclude with Chapter 5 where we discuss the results and their significance to our sponsor company.

2 Literature Review

One of our objectives is to use the assortment of data available in the agriculture domain to develop a short-term, geographic forecast of product consumption. In order to analyze this problem better, we focus our literature review on three aspects: 1) big data and big data in agriculture, 2) forecasting in the agriculture and farming sector, and finally 3) demand sensing and short-term demand forecasting. While we were unable to find any literature that exactly addresses our thesis problem, we aim to understand the concepts surrounding big data and learn from data driven decision processes in agriculture.

2.1 What is Big Data?

While a number of definitions exist, analyst firm Gartner defines big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." Brynjolfsson, Hitt, and Kim (2011) conducted empirical research and concluded that organizations' performance is directly related to their ability to make data driven decisions. This makes it important to understand the success factors required to adopt big data techniques in organizations.

McAfee and Brynjolfsson (2012) postulate that three characteristics, volume, velocity, and variety, make the field of big data fundamentally different from analytics and hence companies require a different set of strategies to become big data enabled. In addition to the technical challenges, big data also presents a set of managerial challenges. Most important among these is the organization's newfound ability to make data-driven decisions and hence move away from intuition-based decision making. Leadership, talent management, technology, decision making, and company culture are critical factors to be successful in leveraging big data for organizational performance (McAfee, 2012).

On the agricultural side, Ashlee Vance (2012) of Business Week reports of innovative company, Climate Corp which uses localized weather data to predict crop yields and uses this data to tailor crop insurance premiums to individual farmers. Climate Corp stores and analyzes more than 60 years of weather data consisting of precipitation, temperature, and soil conditions with a granularity of 2 $^{1}/_{2}$ mile radius in order to arrive at the crop yield forecasts (Vance, 2012).

2.2 Forecasting in Agriculture: Review of Fertilizer Forecasting Studies

Since agriculture is tightly linked to the food-security of a nation, a lot of research has been conducted by governmental agencies to ensure a sustainable and productive agricultural future. In the United States, land grant universities are also a major source of research. They have been the suppliers of yearly growth forecasts such as crop yields and prices as well as short-term outlooks on crop and input prices. The majority of forecasting in agriculture uses econometric modeling. This entails using exogenous variables to forecast crop yields, fertilizer consumption, etc. We are particularly interested in research related to fertilizer consumption forecasting since it relates somewhat to the use of AgChem.

The earliest known work in fertilizer forecasting was performed by Vail (1927), and then by Mehring and Shaw (1944). They tried to investigate the relationship between nitrogen consumption per acre with crop value, soil nitrogen content, and proportion of cash generated from livestock. They all investigated the relationship between expenditure on fertilizer with lagged farm income. Mehring and Shaw concluded that farmers always tend to spend a constant proportion of their income on fertilizer. Vail did not find any significant relationship between fertilizer consumption and fertilizer price.

Zvi Griliches (1958) completed another study of fertilizer consumption with the primary objective of estimating the short-term and long-term demand elasticity. He found that technological changes that resulted in new production technologies led to a reduction in the real prices of fertilizer, thus leading to large scale adoption and increased use of fertilizer. Griliches' hypothesis was that an increase in fertilizer use was predominantly driven by a decrease in the price of fertilizer in relation to other farm inputs and the price received for output. He modeled fertilizer consumption in the current season as a function of the fertilizer price and the previous season's fertilizer consumption using the data from 1911-1956. Instead of using the weight of fertilizer consumed (i.e. total tons consumed per year) he used the USDA measure of "principal plant nutrients", which measures the total quantities of the three main components of the fertilizer – potassium, nitrogen and phosphate. The results he obtained indicated a very strong relationship between the price of fertilizer and the consumption ($R^2 \approx 0.98$) (Griliches, 1958). This study helps us understand the role price plays in factor selection of farmers.

The Food and Agriculture Organization of the United Nations (FAO), which is based in Rome, Italy, publishes a global outlook on fertilizer demand and consumption and other agricultural trends such as crop yields and grain production. Parthasarathy of FAO, in "Demand Forecasting for Fertilizer Marketing" (1994) analyzed various factors that influence fertilizer consumption along with long-term and short-term forecasting models for the same. In addition to exogenous factors such as farmers' purchasing power, which is determined by the affordability of fertilizer, and farmers' cash liquidity, availability of the fertilizer also influences the consumption. Infrastructure and better logistics management results in availability of the product when farmers need it, thus ensuring demand is converted into sales. According to Parthasarathy, in addition to land (area planted and harvested) and yield, other factors such as amount and distribution of rainfall, cropping patterns, and size of holdings also influence fertilizer consumption (Parthasarathy, 1994).

Forecasting in the context of fertilizers can be classified into three categories – assessment of the potential, forecast of the demand, and forecast of the sales. While government agencies are predominantly interested in the first two to derive their policies, companies are typically interested in the second two. Four possible methods are identified that can be employed towards the above stated objectives. They are "measurement of potential through need-oriented or agronomic method", time-series analysis, causal models, and qualitative approach. Parthasarathy recommends using time-series methods for forecasting long-term demand and qualitative approaches such as buyer surveys and sales force opinions to forecast short-term demand and sales (Parthasarathy, 1994).

In another, more recent report, the FAO forecasted global fertilizer use for 2015 and 2030. Agricultural inputs are tightly linked to crop production or yield and the growth of this crop production is related to macro factors such as an increase in population and per capita income. The positive relationship between fertilizer consumption and crop production is well established in both developing and developed countries. The base scenario for projecting fertilizer use assumed fertilizer use to be related to acres planted and yield. The model was a logistic regression model, where logarithmic transformation of original independent variables were used as inputs, since year-on-year changes in variables were used (FAO, 2000).

Finally, while Griliches's (1958) work focused only on fertilizer usage, Tenkorang (2006) attempted to forecast the long-term global fertilizer demand, similar to the work done by the FAO. In addition to forecasting fertilizer consumption, which the FAO completed, Tenkorang also attempted to estimate soil nutrient balance. Since our research will not address soil nutrient estimation, we focus on his fertilizer estimation method. He forecasted fertilizer demand across 182 countries split into nine regions, using data from 1962-2005. An interesting finding is the relationship between fertilizer usages in years following a bumper harvest. Farmers tend to have the good-year/bad-year syndrome where they feel a high yield season is usually followed by a lower yield. Hence, they possibly lack motivation to take steps to increase yield following a high harvest season. Accordingly, Tenkorang modeled fertilizer utilization in a given region and time period as a liner function of current year crop output, previous year crop output, total cultivated land, and a dummy variable to account for any structural shift in fertilizer usage. However, since both land use and crop yield are fundamentally related, his model suffers from multicollinearity. This was remedied by removing those independent variables from the equation based on the variance inflation factor (VIF). The final model, adjusted for multicollinearity, reported a high R^2, with land cultivated to be most significant variable across a majority of the regions (Tenkorang, 2006).

These studies bring out two important concepts in agricultural forecasting. First, most forecast models used are causal models that try to relate a dependent variable such as fertilizer demand to a set of exogenous variables. Second, these studies point us to some of the core factors or independent variables that we should consider in building our model; crop yield, acres planted, and acres harvested being the top three. However, we need to note that these models are yearly forecasting models, whereas our interest is also in short-term forecasting.

2.3 Demand Sensing and Short-term demand forecasting

Since our project involves developing a model to help adjust short-term forecasting in response to changing factors, we also focused on understanding the aspects of demand sensing. Demand sensing is defined by Steutermann, Scott, and Tohamy (2012) as "the translation of demand information, with minimal latency, to detect who is buying the product, what attributes are selling, and what impact demand-shaping programs are having" (Steutermann, 2012). As Robert Byrne (2012) of Terra Technology, a provider of software solutions in demand planning and management, explains, traditional demand forecasting algorithms cannot take into account the changes in consumer demand as they unfold, while demand sensing reacts to these changes as they happen. Byrne also mentions the importance of data – such as economic recessions and competitor actions – that can reflect consumer demand patterns (Byrne, 2012).

Agarwal and Holt (2005) used a short-term demand forecasting model that incorporated point of sale (POS) data in order to fine-tune the demand forecast for a consumer goods company. They proposed using the moving-average method, which is a time-series analysis technique in order to forecast the next time period sales based on the POS data of the past few time periods. While this sheds light on usage of POS data in short-term demand forecasting, we are more interested in using causal factors in addition to POS data.

2.4 Chapter Summary

This chapter explores the terminology and challenges in big data. It also reviews the previously-performed work on forecasting in agriculture with a focus on fertilizer. This helps us understand the factors that should be considered and the techniques that can be applied. In our next section, Methodology, we will explain how we approached this problem, what variables we considered, and how we analyzed the situation.

3 Methodology

To explore analytical models that allow demand sensing and forecasting for the chemical AgChem, we developed the following methodology. First, we investigated the factors that influence the sale of AgChem. Next we broke the problem of forecasting into three sub-problems. For each sub-problem, we identified the pertinent factors and developed an appropriate modeling approach.

3.1 Understanding the Factors influencing Demand

In addition to conducting a literature review, we held extensive brain-storming sessions with the supply chain and sales teams at DAS to understand the factors believed to influence the demand for AgChem. These sessions were conducted during site visits and in telephone conference calls.

AgChem is used in the production of corn. Thus, the agricultural factors we identified relate specifically to corn production. The DAS sales and supply chain teams believed that factors such as fertilizer prices, price of the product, and corn prices influence growers' desire to invest in AgChem. For example, an increase in the price of corn could encourage growers to invest more in the crop by using additional yield-boosting chemicals. Similarly, retailers' marketing efforts were believed to influence the demand. Retailers' willingness to market the product, however, is influenced by their distribution infrastructure, such as liquid bulk tanks. Other factors such as weather, including daily temperature and rainfall, were also believed to influence demand. As discussed in the Introduction, this product is available in two varieties – with fall being the primary sales season for V1 and spring for V2. Hence, the available application window also has an impact on sales. Figure 4 summarizes the identified factors in a causal loop diagram.

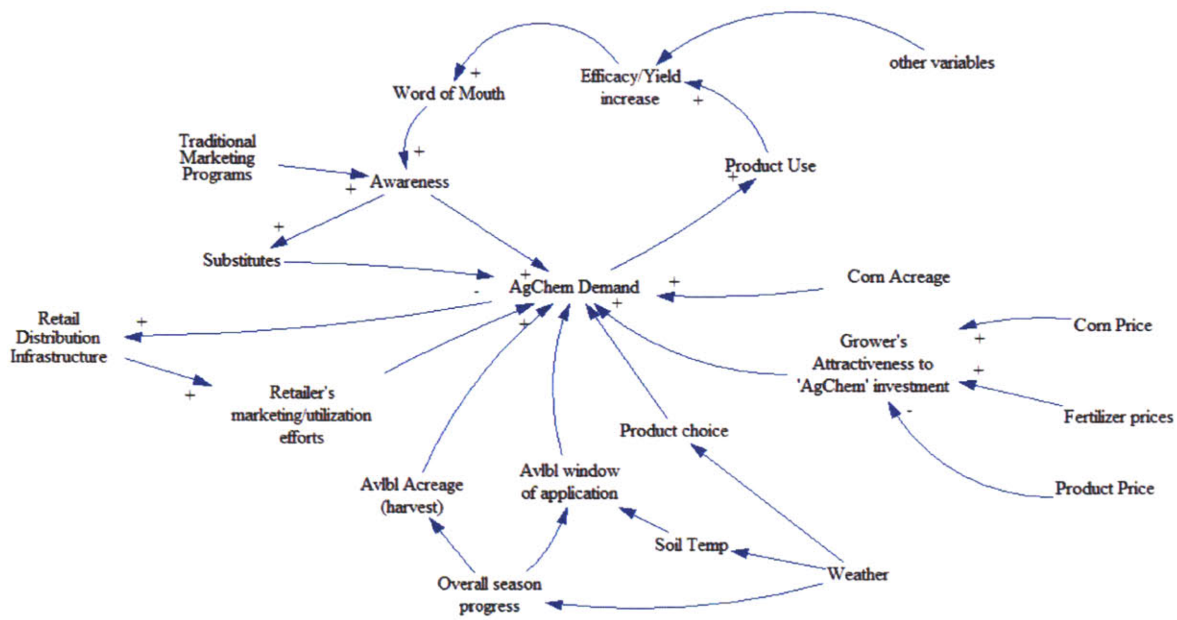

Figure 4 Factors influencing demand of the product

Some of these factors cannot be captured quantitatively. For example, while we know that growers' attraction to AgChem positively correlates with sales of the product, this attraction is dimensionless. In such cases, we have to rely on the underlying factors, such as price of fertilizer, price of the product, and price of corn, for actual analytical model development. These factors and their presumed influence on sales are shown in Table 1. Note that corn price and yield are lagging indicators. They determine the revenue that a farmer gets in the previous year and hence the money he can expend on farm inputs.

Table 1 Factors influencing the demand and their granularity

Factor	Description	Granularity[1]
Corn Price	Price of the corn. Usually measured in $ per bushel	Daily, by geography
Fertilizer Price	Price in dollars per pound of fertilizer	Annual
Product (AgChem) Price	Sales price of the product	Annual
Corn Acreage (Planted)	Number of acres planted	Annual, by county
Corn Acreage (Harvested)	Number of acres harvested	Annual, by county
Yield	Bushels per acre of corn harvested	Annual
Temperature	Degrees F	Hourly, by city
Precipitation/Rainfall	Measured in in. or mm	Hourly, by city
Marketing Budget	$ spent in marketing per year	Annual, by retailer
Distribution Infrastructure	Installed tank capacity in gallons	By retailer location
# of retailers	Number of retailers in a given county	Annual, by county

The last three of these factors relate to DAS's corporate strategy:

1. Marketing budget: As a part of a corporate effort to increase farmers' awareness of the product and the product reach, DAS collaborates with select distributors and retailers via advertisement, promotions, and discounts. This is usually measured as the amount of money spent in marketing efforts.
2. Distribution infrastructure: In order to increase the product availability in high growth geographies, DAS commissions bulk tanks at retailer and/or distributor sites. Distribution infrastructure is thus measured by the installed tank capacity.
3. Number of retailers: Since this product is generally not transported far from the retailer location, the number of retailers in a given area could lead to increased sales via increased product availability for the growers.

[1] Granularity defines the level at which data is captured by national agencies such as USDA and hence the level at which data is available for analysis.

We also tried to analyze the farmer's decision making process when deciding whether to purchase AgChem. The decision making process of a farmer can be broken into three steps:

1. Whether to buy AgChem: This is determined by factors such as his income from the previous season, current corn prices, current AgChem price, production costs in the current season, his familiarity or inclination – determined by marketing efforts and product availability.
2. When to apply: Once a decision is made to invest in AgChem, the second choice he faces is around the application window. A farmer can apply AgChem in the fall or spring. Even within a season, the farmer must consider the weather since there are specific temperature and moisture requirements for the product to be effective.
3. How much to apply: If the farmer will use AgChem and knows when, he must decide how much of his crop will receive the product.

Once we understand the various influencing factors and this decision making process, we move to the next step of dividing the problem and identifying model specifications.

3.2 Dividing the Problem

We decided to segment our problem into three sub-problems, or predictive models, to provide the greatest benefit to DAS.

1. The first model will predict total annual sales on a national basis. Macro-level, yearly variables such as average corn prices, corn acres planted, and average yield will be used to forecast the quantity for the entire season. This will help DAS plan their annual production schedules and understand future demand for the product.
2. The second model will determine the quantity of product that will be sold in a given county during the entire season. We chose the county as the geographical dimension since most of the publicly available data is measured at the county level. This will help DAS plan the distribution of AgChem commensurate with the geographical demand.
3. The third model will be a short-term demand sensing model which will determine weekly sales or product movement within a specified geography during the sales

season. Daily and weekly variables such as temperature and precipitation will be used to develop this model.

3.3 Model Development: OLS Regression

Since our objective is to investigate the relationship between a set of independent factors and the sales of AgChem, causal forecasting models were investigated. Accordingly, we chose the regression method as the modeling technique.

Simple linear regression posits a linear relationship between the independent variables and the dependent variable, whose values are to be forecasted. Equation 1 represents the simplest form where only a single independent variable is used to predict the dependent variable.

$$Y_i = \beta_0 + \beta_1 * X_i + \in \tag{1}$$

Where Y_i are the values to be forecasted and X_i, β_0, and β_1 are the coefficients that indicate how independent variables affect the dependent variable. $\in$, called error term or noise, stands for the difference between the actual value of Y_i and the predicted value. It is also called ordinary least squares (OLS) regression, since the coefficients are derived by minimizing the sum of squared errors. It is important to note that the coefficients are estimated and hence the regression equation sometimes is written as follows, where $\hat{Y}$ (read y-hat) represents the predicted value of Y and $\hat{b}_0$ and $\hat{b}_1$ are estimated values of β_0 and β_1, respectively.

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 * X_i + e \tag{2}$$

Another important concept in regression is the coefficient of determination or R^2, which measures the strength of the relationship. It represents the amount of variation in the dependent variable that is explained by independent variables. Its value can range from 0 to 1, with higher values usually representing better relationships (with the exception of non-linear relationships). This is closely linked to the concept of residual sum of squares and total variation. The derivation of R^2 is shown in Equation 3.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSR}{SST} \qquad (3)$$

SSR stands for residual sum of squares and SST stands for total sum of squares. Residual sum of squares measures the "unexplained error" whereas total sum of squares measures the total error as measured from the mean. Note that residuals are nothing but error terms indicated in Equation 2.

Other important concepts of regression are t-statistics and p-values. They measure the significance of the independent variable and usually serve as a decision criteria to include or exclude a given variable. Typically a 95% confidence interval is used, which means a p-value of less than 0.05 denotes significance of the independent variable.

4 Data and Analysis

In this chapter, we summarize the data used for the analysis. We then present the models we developed and the insights we gained from those models. The data summary is presented in two sections. The first section summarizes the data from our sponsor company and the publicly available data that was used for the analysis. The second section describes the data selection and cleaning that was applied in order to select the right data for modeling.

4.1 Data Summary

Figure 1 on page 9 of the Introduction, depicts the echelons in the supply chain that our sponsor company is part of. As explained previously, various types of sales data were obtained from the sponsor company. In addition to sales information, we also obtained the data listed below.

- Inventory data for bulk tanks: Our sponsor company invested in a bulk tank infrastructure which is monitored for inventory levels. We obtained this tank monitoring data for use in our analysis.
- Marketing spend: DAS spends marketing dollars with retailers and distributors in order to increase visibility of their product. This information is tracked by retailer and we obtained this data as well.

In addition to data provided by the sponsor company, we leveraged publicly available data predominantly from the USDA and Iowa State University. USDA sources provided us with data on annual factors such as corn acres planted, corn acres harvested, and corn yield at a county level. From Iowa State University's weather data source, we downloaded data on daily high and low temperatures, snow fall, and precipitation from September 2011 to May 2012 for various cities. We chose the season 2011-2012 since the retailer and distributor shipment data is available by county level only for this season.

4.2 Analysis and Results

The objective of our work was to use the existing public data in agriculture to identify the relationship between macro-level and micro-level factors and the sales of the product. The annual factors such as corn acreage and yield are classified as macro-level data whereas daily temperatures and precipitation are classified as micro-level data. Our expectation was that macro factors contribute to annual trends in the sales and micro factors contribute to short-term trends in the sales.

Accordingly, the following three hypotheses are tested in this project:

1. Macro factors such as corn price, fertilizer price, corn acreage, fertilizer usage, and corn yield influence the total annual sales of the product.
2. Geographical distribution of sales by county is linked to corn acres planted, corn acres harvested, average yield, sales in the first few weeks of the season, marketing dollars spent, and bulk tank capacity within that county.
3. Near-term demand (i.e. weekly or daily demand) correlates with daily weather patterns and season-to-date sales.

To test these hypotheses, we built three main regression models. The models, along with any additional sub-models, are described in detail in subsequent sections. In addition, we present the model analysis and results and discuss our findings.

4.3 Preliminary Analysis

We started with a preliminary analysis of the data to understand and find patterns if any. In addition, this analysis helped us to check the validity of our hypotheses, wherein we are proposing correlation between various factors and the sales of the product.

We began by looking at 10 years of shipment data from DAS to its distributors to get a visual understanding of the sales and the data characteristics. This is presented in Figure 5. We gained three key insights from this graph. First, sales significantly increased after 2005. While we had not yet weighed empirical evidence, we assumed this spike in sales was due to an increase in corn production, driven primarily by increased ethanol production. Second, shipments seem be fairly seasonal, with November being the peak month. This is in line with the product application characteristics. The product is applied either after harvest in the fall or before planting in spring. Third, there was a sizable dip in shipment volume during 2009, which was a drought year and during a recession. This reinforces the intuition that sales of this product could be correlated to weather and economics.

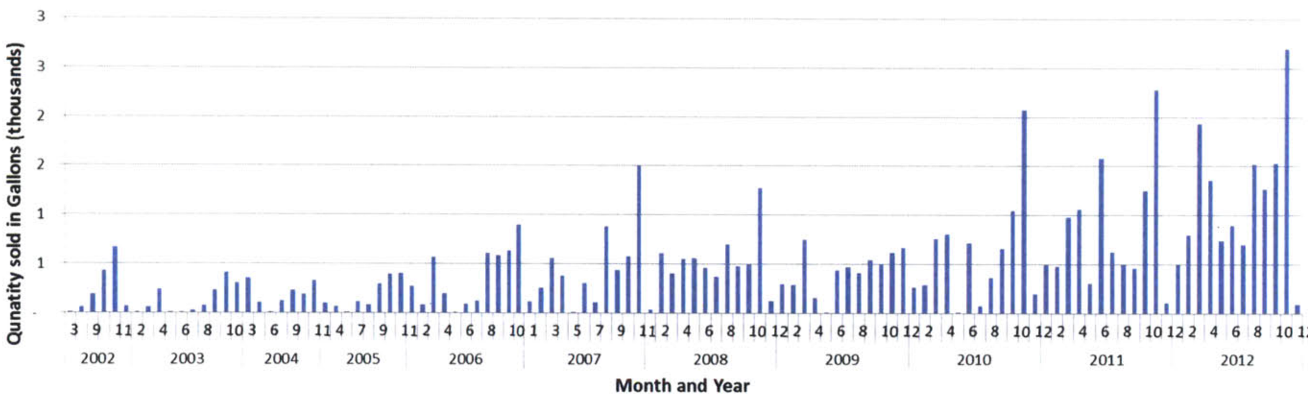

Figure 5 Increasing demand and seasonality of AgChem observed over a 10 year period

Given that we only have a few data points for each model we intend to build and that there didn't seem to be any unexplainable outliers, we chose not to ignore any data points.

Since we know that there are at least 3 echelons in the distribution chain, we started with a summary plot of shipments from DAS to distributors, distributors to retailers, and retailer to growers for the calendar year 2011. In accordance with intuition, Figure 6 shows that shipments from manufacturer to distributors happen before the other two echelons, thus re-assuring us of the data sanctity.

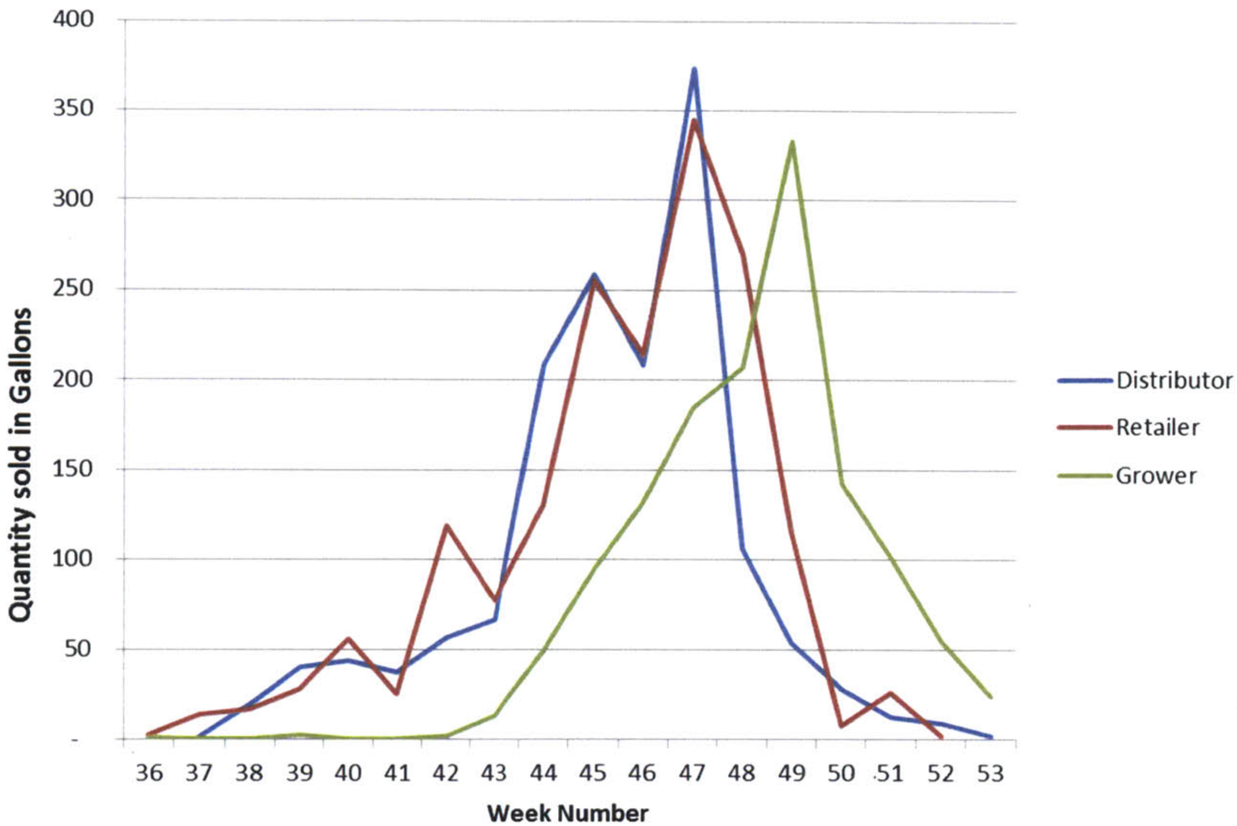

Figure 6 Product flow between echelons

4.4 Model-1: Annual Sales Model

The first model, which we call "annual model," posits a relationship between annual sales of AgChem and these independent variables: average corn price received, corn acres planted, average yield, fertilizer price, and fertilizer usage. The plot of these annual variables is shown in Figure 7. As the plot shows, corn price, fertilizer price and price of AgChem appear to be positively correlated with annual sales of AgChem.

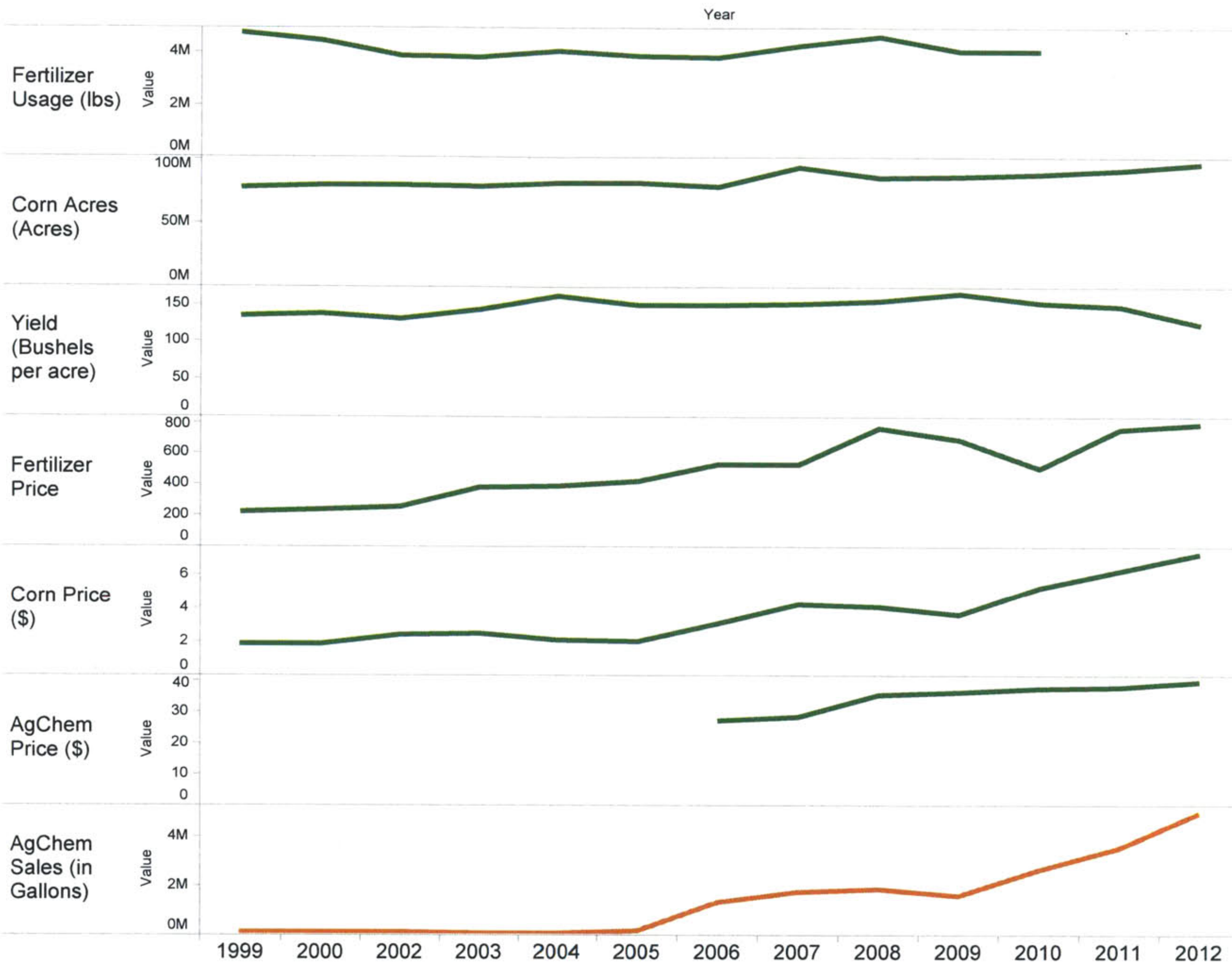

Figure 7 Annuals sales trending along with fertilizer and corn price

Next, we computed correlation coefficients for these variables using SAS' statistical software, JMP. The results are presented in Table 2 below. The correlation coefficient matrix shows that the sales are strongly correlated with corn price, fertilizer price, AgChem price, and corn acres. Out of these four independent variables, AgChem price is available from 2006 only. Hence, this variable was excluded from the regression model and only the remaining three variables were used. The results of regression developed using SAS JMP are summarized in Table 3.

Table 2 Correlation coefficient matrix for all variables

	AgChem Sales	Fertilizer Price	Corn Price	Corn Acres	Yield	Fertilizer Usage	AgChem Price
AgChem Sales	1						
Fertilizer Price	0.83	1					
Corn Price	0.99	0.82	1				
Corn Acres	0.89	0.77	0.90	1			
Yield	-0.13	0.27	-0.10	0.01	1		
Fertilizer Usage	0.06	-0.02	0.005	0.10	-0.15	1	
AgChem Price	0.72	0.67	0.73	0.52	-0.28	0.29	1

Table 3 Regression results of the annual model

	Estimate	Std. Error	t-ratio	P-value
Intercept	-874773.65	1961678	-0.45	0.6662
Fertilizer Price ($ per lb.)	443.83	653.80	0.68	0.5143
Corn Price (dollars)	872405.76	111910.7	7.8	<.0001
Corn Acres	-0.012	0.027	-0.46	0.6554
Fit Statistics				
Adjusted R^2	0.97			
Mean Squared Error	258868			

The first column lists the variables and the second column gives the value of the coefficient. We are interested in the last column, p-value, as it indicates the significance of a given variable in the regression relation. A p-value of less than 0.05 is considered to be significant. As expected, corn price turns out to be a very significant factor. But surprisingly, fertilizer price and corn acres have a p-value much greater than 0.05, indicating that they might not be significant. The variable corn acres also has a negative sign for coefficient. But given, that the p-value of this coefficient is greater than 0.05 we suspect that corn acres might be correlated to other independent variables, i.e., the model has more explanatory variables than needed.

Subsequently, we dropped the two terms that have insignificant p-values and developed the regression model with just corn price as the independent variable. The results are shown in Table 4. As can be expected, both the intercept and corn price turn out to have significant p-value and there is virtually no change in adjusted R^2 value.

Table 4 Regression results of the annual model with only significant factors

	Estimate	Std. Error	t-ratio	P-value
Intercept	-1713408.67	155083.85	-11.05	<.0001
Corn Price (dollars)	872628.97	39583.75	22.05	<.0001
Fit Statistics				
Adjusted R^2	0.97			
Mean Squared Error	241889			

This fits with our intuition and the graphical analysis we did earlier. One way to explain this relationship is to say that as corn prices go up, farmers have an incentive to produce and sell more as well as additional capital to spend on inputs. Hence, a farmer would be interested in spending more in his production and will be willing to invest in AgChem, which can boost yield.

However, it is difficult to believe that corn price is the only driver of sales and that corn acres and fertilizer price have no role to play. As we can observe in Figure 7, sales are positively correlated with the price of fertilizer. One possible explanation is that as price of fertilizer increases the grower wants to insure his expenditure and revenues and hence will buy AgChem. To evaluate whether increased corn acres planted leads to increase in sales we need to know the market penetration of the product. A higher penetration means farmers are aware of the product and hence an increase in acreage might lead to an increase in sales. While we cannot fully explain the insignificance of fertilizer price and corn acres, we need to notice that this regression was developed with just 13 data points and, therefore, might not accurately depict the relationship between dependent and independent variables. Further analysis with more data points would validate this result.

4.5 Model-2: Geographical Sales Model

Next we developed a model to explain the geographic spread of the sales. Our hypothesis is that sales in a given geographical location are correlated to acres harvested in the area, yield in the area, number of retailers in the area, and quantity sold in the first two months of the season. It is important to understand the rationale for our hypothesis.

Usually the product is applied either after harvest in the fall or before planting in the spring, for the next growing season. For example, product is applied in September 2011 – April 2012 for the corn season that starts in April 2012 and ends in November 2012. Our intuition is that if a farmer has a good yield and lots of acres planted in the current season (April 2011 – November 2011), then his income would be higher, as would his ability to spend more on farm inputs. Hence, income from this season will affect his decision to purchase the product or not. Similarly, more retailers in a given area will produce higher marketing power and accessibility of the product, thus leading to higher sales. Finally, our interviews with our sponsor team indicated that typically higher sales in the first two months indicate a higher season overall. Hence sales in the first two months is included as an independent variable. However, we understand that adding the first two months sales as explanatory variable will automatically lead to higher R^2 due to very high correlation between total season sales and the first two months' sales. Thus, we first evaluate the model without this variable and then add it. The model regresses the sales in September 2011- April 2012 as a function of these independent variables.

We considered a number of other independent variables and tested for their strength using p-values in ordinary least squares regression. We dropped the variables whose p-values were insignificant, leaving only the four variables discussed above in the model. Each one of these was tested individually in a regression model for its strength before the insignificant variables were dropped. The complete list of all variables tested and their status (included or not included) is summarized in Table 5.

Table 5 Summary of factors for geographic model

Variables	Granularity	W/ Dist to retailer data
Distributor sales by geography (i.e. to retailer)	By county, annual	Dependent (Y)
Corn acres planted (2011)	By county, annual	Not Significant - Excluded
Corn acres harvested (2011)	By county, annual	**Significant**
Yield (2011)	By county, annual	Not Significant - Excluded
Increase in yield (Y-o-Y)	By county, annual	Not Significant - Excluded
# of retail locations	By county (aggregated)	**Significant**
Bulk tanks	By county (aggregated)	Not Significant - Excluded
Annual rainfall (2011)	By county, annual (crop season)	Not Significant - Excluded
Historical sales	By county	Not Significant - Excluded
Sales in Sep-Oct 2011	By county	**Significant**
Quantity purchased by grower	By county (aggregated)	Not Significant - Excluded
Farm size (e.g. customer size)	By county (aggregated)	Not Significant - Excluded

We made two key decisions while building this model. The first one was which sales data to use. We had two sets of sales data: sales from distributor to retailer and the retailer to farmer. The second decision was the geographic level for data aggregation (e.g. city vs. county vs. state). We choose to use distributor to retailer sales since it was more complete than the retailer to farmer data and it also had the retailer ship to location. Retailer location is a fairly good approximation of actual grower application location, since growers typically don't travel very far to obtain the product. We then aggregated the data to the county level. Given that we didn't have the exact city/location of application (i.e. grower application city), we felt aggregating by county was the best choice based on our discussions with the sponsor company.

The results of the regression on the statistical analysis package SAS JMP are displayed in Table 6. In total, we have 134 counties, hence 134 data points. We can see that acres harvested and number of retailers are significant.

Table 6 Regression results of geographic model

	Estimate	Std. Error	t-ratio	P-value
Intercept	74.08	1618.46	0.05	0.96
Harvested	0.03	0.01	2.92	0.0041
Count of Retailer	311.63	48.36	6.44	<.0001
Fit Statistics				
Adjusted R^2	0.38			
Mean Squared Error	8149.8			

Next, we included the first two months' sales as an explanatory variable. The results are reported in Table 7. We can see that in addition to acres harvest in prior season and number of retailers, the sales in first two months also turns out to be significant as can be expected. While we expected certain additional factors, specifically bulk tank capacity and annual rainfall, to be significant, they are not. Even though yield and intercept term have a p-value that is higher than 0.05, they are left in the model for sake of completeness.

Table 7 Regression results of geographic model with season start sales included

	Estimate	Std. Error	t-ratio	P-value
Intercept	7187.19	3936.08	1.83	0.0701
Harvested (2011) i.e. prior season	0.026	0.010	2.41	0.0176
Yield (BU/Acre) (2011) i.e. prior season	-51.02	28.40	-1.8	0.0748
Count Of Retailer	235.56	42.18	5.58	<.0001
Quantity sold in September & October (2011)	1.72	0.23	7.34	<.0001
Fit Statistics				
Adjusted R^2	0.55			
Mean Squared Error	7116			

While the R^2 is not very high, the model makes intuitive sense. A higher harvest will leave the farmer with more money which can drive sales. Similarly, as number of retailers increases the product reach increases and we would expect sales to increase. Also note that we didn't discard any data. A careful search for outliers will help us get a better R^2. Finally since we didn't have exact location of product application, we used retailer location and then aggregated it to respective counties. This approximation might also have skewed the result we presented above. Nonetheless, the above model presents a significant finding that harvest, number of retailers, and sales in the first few weeks act as strong indicators of total season sales.

4.6 Model-3: Short-term Demand Model

Next, we developed a short-term demand model that would correlate weekly sales with weather related variables. The decision of a farmer about when to apply the product depends a lot on weather factors such as temperature and precipitation. If the weather is too cold and hence is not conducive for the farmer to go into the field, then he will not apply the product. Also, for application in the fall the temperature has to be no more than 50^0 F and trending down. In light of these details, we conducted regression analyses on weekly sales with average weekly temperature and precipitation. Since it was not possible to do this analysis for all the 1000+ cities in the data, we chose to do this only for two of the top selling cities. Before we present our detailed analyses on the weather relationships, we present a few key insights we gained while analyzing the data.

4.6.1 Timing of Product Application

As described in the Introduction, there are two variants of the products that farmers can choose to apply. The first variant AgChem V1 is applied during the fall (after harvest) or in the spring (before planting). The second variant, AgChem V2 is applied only in the spring. For this analysis we considered only the first variant, AgChem V1, since this variant accounts for more than 80% of the sales and is also the oldest.

In our first analysis we focused on identifying whether there was a pattern to when farmers chose to apply the product. We have 5 years of distributor sales data that gave us the retailer cities where the product was shipped. Using this, we plotted the quantity sold in fall and spring as a pie chart across cities over the years. The results indicate a trend that some geographies are predominantly fall application communities whereas the rest tend to be spring application communities. In Figure 8, green indicates quantity sold during fall season (September – December) and red indicates quantity sold in spring season (January – April). We see that most of Iowa and northern Illinois are predominantly green whereas southern Illinois and Indiana tend to be red indicating these are spring-application states. We plotted data for subsequent years as well (refer to appendix-1) and this pattern seems to hold true across years.

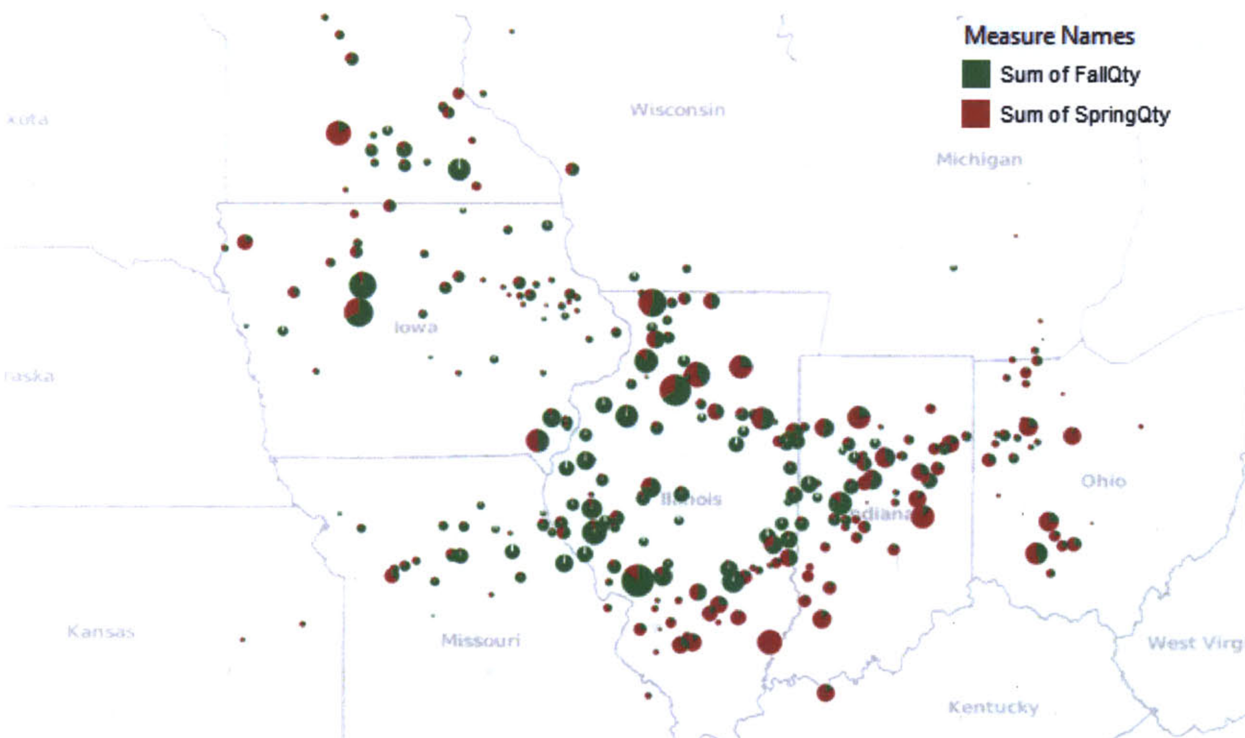

Figure 8 Geographic plot of fall and spring quantities shipped (2007)

Next, we analyzed whether we could identify any events that could potentially act as leading indicators of the sales.

4.6.2 Harvest date as a trigger point for fall application

The first insight we gained is how harvest date acts as a trigger point for application in fall. For the product to be applied, the corn crop has to be harvested. This leaves the farmer with a choice of applying it in fall or in spring. We first found the peak week of sales for each city. This will be the week where maximum sales happened. We then plotted the quantity sold in the peak week of each city from September 2011 to April 2012. This is shown in the Figure 9. Week 48 (red line) represents the week when 100% of the corn was harvested. As we can see, there is very high demand (sales) from week 49 to week 50 and significant demand around week 13. Note that week 41 through 53 represent the fall of 2011 whereas week 1 through 17 represent the spring of 2012. This means a lot of cities sell the product in the first couple of weeks following harvest. Additionally, there are a considerable number of cities where peak sales happen in the spring. Also, the total peak week sales account for 31% of the total season sales. This means that essentially $^{1}/_{3}$ of all sales happen during the peak week of each city. The manufacturer can benefit immensely by forecasting the peak week and positioning inventory to serve these spikes in demand. And harvest date could act as a trigger point for sales, essentially helping us forecasting peak weeks.

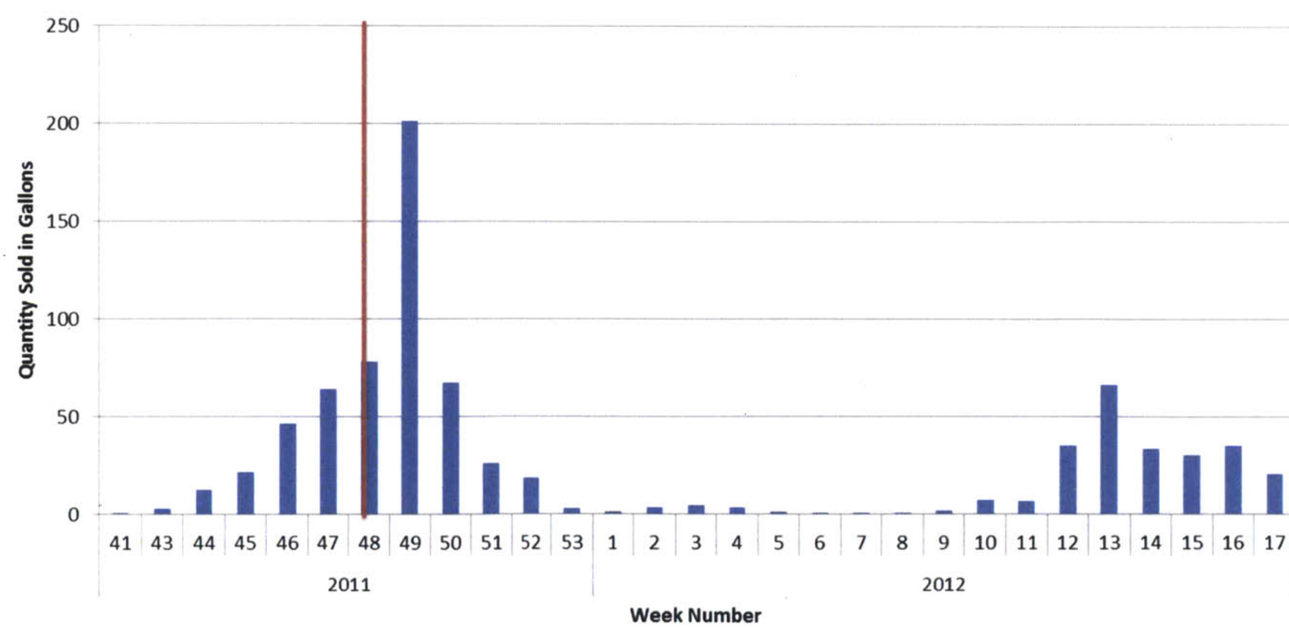

Figure 9 Peak sales occur within 1-2 weeks after harvest completion

Next, we developed a regression model to analyze how daily temperature and precipitation influence sales.

4.6.3 Temperature and precipitation as predictors of weekly sales

We have point of sale data from the retailers for the 2011-2012 season. We used this data to obtain aggregate sales by city, for each week. This data contains nearly 1000 cities. Since we could not analyze all these cities we decided to pick one city with the highest sales in fall and a second one with the highest sales in spring. Thus we end up with Ames, Iowa and Avon, Indiana.

4.6.3.1 Results for Ames, IA

Note that Ames is highest in the fall. The model uses weekly sales as the dependent variable with average weekly temperature and average weekly precipitation acting as independent variables. The regression results for Ames, Iowa are shown in Tables 8 and 9.

Table 8 Weekly quantity sold data for Ames, IA

Week Start Date	Quantity sold (quarts)	Average Temp (F)	Average Precipitation (mm)
9/25/2011	147.5	62.3	0.0
10/2/2011	75.0	65.0	0.0
10/9/2011	165.0	59.5	8.5
10/16/2011	126.9	44.0	0.0
10/23/2011	1124.7	51.7	0.0
10/30/2011	2845.4	45.4	3.6
11/6/2011	6007.0	41.7	7.5
11/13/2011	5397.6	40.4	0.0
11/20/2011	7687.4	41.8	0.0
11/27/2011	11611.7	31.6	0.0
12/4/2011	11712.7	19.4	0.1
12/11/2011	7753.3	37.0	4.7
12/18/2011	3151.0	31.3	0.1
12/25/2011	2371.6	36.6	1.3
1/1/2012	25.6	45.5	0.0
1/8/2012	204.1	31.3	0.3
1/15/2012	935.3	8.8	0.7
1/22/2012	70.9	59.5	0.0
1/29/2012	34.5	70.0	0.0
2/5/2012	337.1	58.8	0.0
2/12/2012	1290.4	59.0	1.1
2/19/2012	1882.6	47.4	2.8
2/26/2012	855.4	53.1	3.9
3/4/2012	629.9	60.1	2.8

Table 9 Regression results of weekly model using entire season data

	Estimate	Std. Error	t-ratio	P-value
Intercept	8755.97	2140.20	4.09	0.000522
Average Temp	-134.86	44.42	-3.03	0.006279
Average Precipitation	128.43	269.24	0.48	0.63828
Fit Statistics				
Adjusted R^2	0.24			
Mean Squared Error	3174.75			

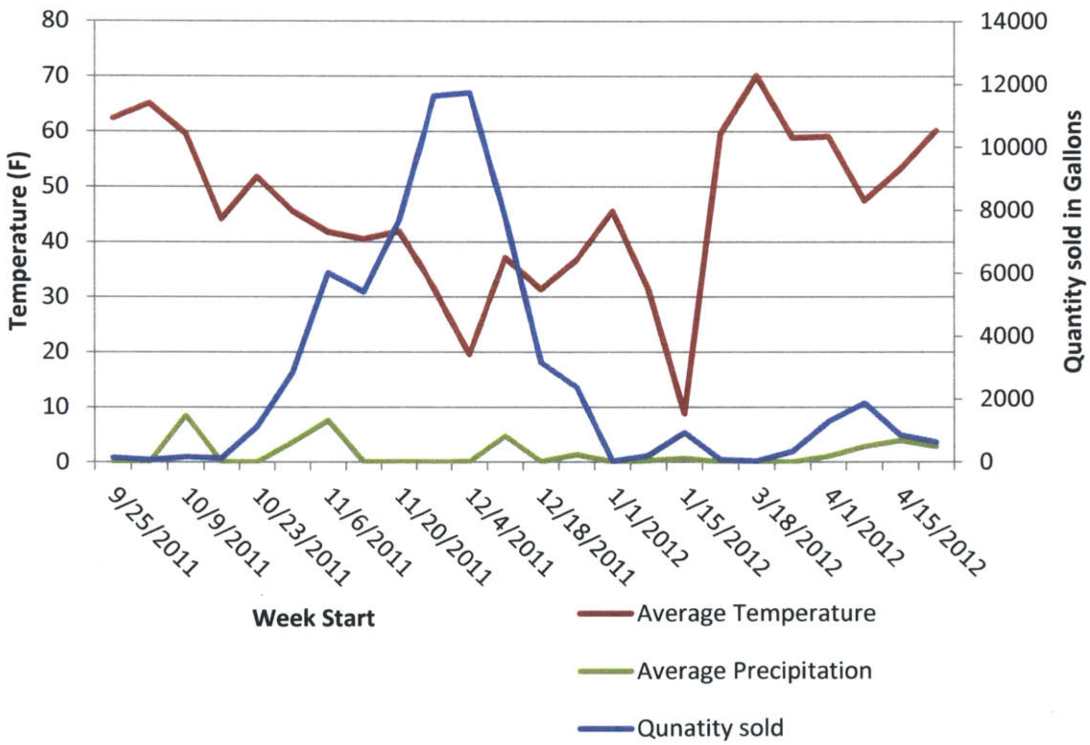

Figure 10 Weekly sales correlate with temperature for Ames, IA

Figure 10 indicates that significant reductions in air temperatures coincide with jumps in sales. Each time there is a major drop in temperature, there is a large peak in sales. We cannot, however, conclude that the colder it is the more sales will occur. If this was true, sales would be highest during the coldest part of the year. We can see in the figure that this is not the case. The peaks in sales are not proportional to the drops in temperature.

We see that the adjusted R^2 is very low and precipitation is not at all significant, as its p-value is much greater than 0.05. Since we know that Ames is predominantly a fall application city, we reran the model by including fall weeks only. The results in Table 10 indicate much better R^2 value; however precipitation still is not a significant factor.

Table 10 Regression results of weekly model using only fall data

	Estimate	Std. Error	t-ratio	P-value
Intercept	16186.97	2663.85	6.08	0.0001
Average Temp	-270.46	59.19	-4.57	0.0010
Average Precipitation	73.03	247.74	0.30	0.7741
Fit Statistics				
Adjusted R^2	0.61			
Mean Squared Error	2650.87			

4.6.3.2 Results for Avon, IN

Next we ran a similar model for Avon, IN. Figure 11 shows that Avon is a city with high spring sales. It is also one of the top-30 cities for overall sales. The data and results of the regression are shown in Tables 11 and 12.

Table 11 Weekly quantity sold data for Avon, IN

Week Start Date	Quantity sold (quarts)	Average Temp (F)	Average Precipitation (mm)
9/25/2011	2	53.5	0
10/2/2011	130	59.5	0
10/9/2011	28	62	1.4
10/16/2011	15	45	21.6
10/30/2011	20	45.5	0.5
11/6/2011	41	43.25	2.3
11/13/2011	1225	46.4	2.74
11/20/2011	680	44	0.4
11/27/2011	3561	36.4	8.88
12/4/2011	3169	33	3.04
12/11/2011	985	38	5.075
12/18/2011	541	45.25	9.4
12/25/2011	1230	39.17	0.43
1/8/2012	31	23.25	0.25
1/15/2012	123	29.5	2.05
1/22/2012	8	43.5	17.8
2/12/2012	68	29.75	0.65
2/19/2012	249	37.5	0
2/26/2012	282	41	0
3/4/2012	579	42.2	0.06
3/11/2012	2587	64.1	1.02
3/18/2012	10454	67.29	2.47
3/25/2012	16134	54.75	1.95
4/1/2012	7410	55.17	0.55
4/8/2012	11354	48.42	2.62
4/22/2012	13163	52.33	1.13

Table 12 Regression results of weekly model using entire season data

	Estimate	Std. Error	t-ratio	P-value
Intercept	-5495.09	3947.28	-1.39	0.1766
Average Temp	196.40	82.24	2.39	0.0251
Average Precipitation	-89.53	169.02	-0.52	0.6011
Adjusted R2	0.14			
Mean Squared Error	4581.12			

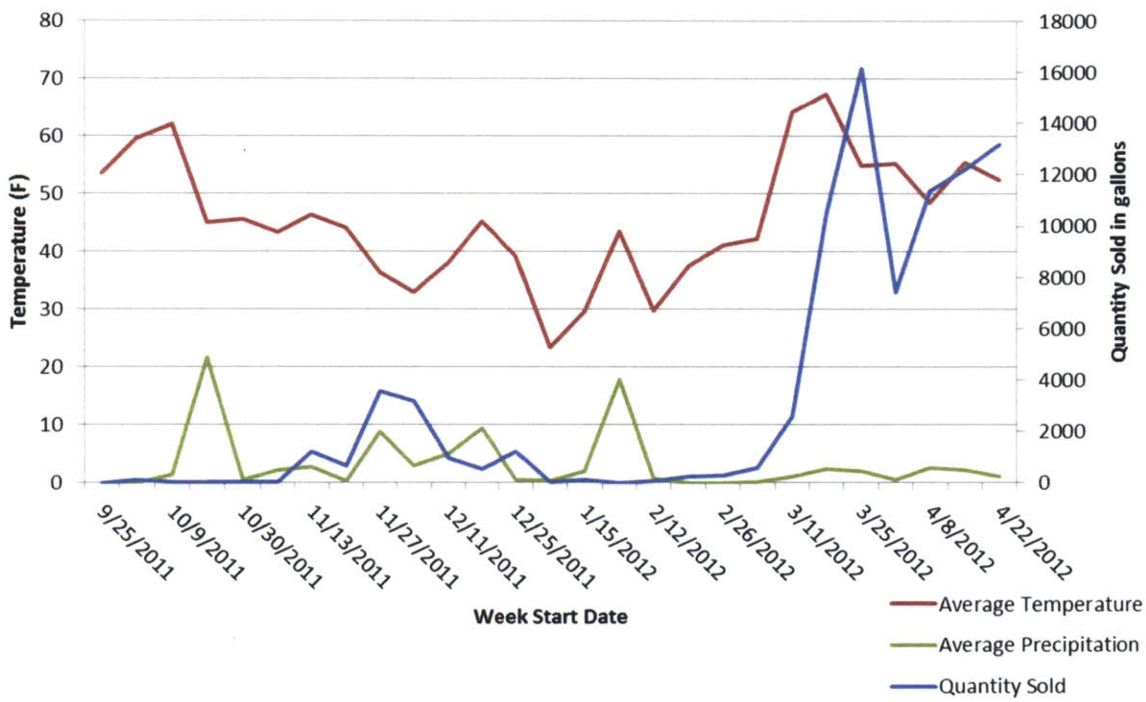

Figure 11 Peak sales occur in spring for Avon, IN

Similar to the case for Ames, IA, Figure 11 indicates that lower temperature in fall corresponds to higher sales. However, Avon is a city where peak sales happen during spring. Hence, we reran the regression model with data from only spring weeks. The results are presented in Table 13.

Table 13 Regression results of weekly model using only spring data

	Estimate	Std. Error	t-ratio	P-value
Intercept	-8560.1	4940.15	-1.73	0.1110
Average Temp	311.15	103.08	3.018	0.0116
Average Precipitation	-180.88	297.80	-0.60	0.5559
Adjusted R2	0.36			
Mean Squared Error	4881.35			

The results indicate that temperature is a significant factor but precipitation is not. While there is an improvement in the value of R^2, it is not strong enough to be conclusive in Avon, IN.

We discussed these results with our sponsor company and concluded that temperature acts as trigger point similar to harvest dates but temperature by itself not a driver for sales. For example, in fall a temperature of 50^0 F is desirable and once this threshold is reached the application can start. The application can continue until the ground freezes. However, a decreasing temperature doesn't necessarily translate to increase sales through the winter. In summary, temperature acts as an on-off switch that triggers sales rather than being a continuous variable affecting sales.

5 Conclusion

5.1 Models and Significant Variables

Our project scope was to identify how DAS can use the vast publically-available data surrounding agriculture to improve their supply chain processes and decisions. We set out to do this through the development of predictive models to forecast demand of AgChem on various time and geographical horizons. We sought to identify factors which could help us predict the total annual demand, the annual demand by geographical sub-sets (county), and a short-term demand sensing model to predict when this demand will occur. We expected that there was enough agricultural data available for us to find some related factors.

For the total annual demand, we found two factors which appear to be significantly correlated; the average price of corn received by growers the previous harvest and the price of fertilizer. We used annual AgChem sales data from DAS for the years 1999 through 2012.

For the geographical demand, we identified that the number of retailers in the area, the corn acres harvested in that area, and the first two months of AgChem sales are correlated to the amount of AgChem which is sold throughout the season. We also identified an interesting pattern about the sales during the spring versus the fall. We found that through the past five years, sales have been heavier in the spring in southern Illinois, through Indiana, and into Ohio. However, in the rest of the major corn-producing regions, AgChem is sold predominately in the fall.

For the short-term demand sensing, there are two triggers for the beginning of the sales period for fall application. The previous crop needs to be harvested and out of the field, and the temperature must be below 50^{0} F. For the year we analyzed, the temperature was not an issue as it was already cold enough when the crops were harvested. For that year, we discovered that the largest volume of fall sales at the retailer locations occurs generally within the two weeks following the corn harvest. This short window following the harvest accounted for roughly 40% of the fall sales. See Table 14 for these variables.

Table 14 Models and variables

Model	Dependent Variable	Independent Variables - Significant
Model-1: Annual Model	Annual sales nation-wide (gallons/year)	Corn price, Fertilizer price
Model-2: Geographic Model	Annual sales by county (gallons/county)	Acres harvested, Number of retailers, Sales in the first two months
Model-3: Weekly Model	Sales by week in a given city (gallons/week)	Average weekly temperature, Harvest completion date

5.2 Implied Benefits to DAS

This research and its results can benefit DAS in a variety of ways. DAS can better understand the ordering trends of each link in the distribution chain. They can quantifiably see when orders and shipments happen from them to the distributors, from the distributors to the retailers, and from the retailers to the growers. With this information, DAS can make decisions about the production and distribution of AgChem.

With a better understanding of the factors driving annual demand for the product, DAS can make production plans which balance the best utilization of their manufacturing assets and the expected profits from the sales of AgChem.

DAS can estimate where demand might be higher and lower in a given year based on the identified factors. They can also gain additional insight into how the sales period will progress within those specific geographies. Knowing this, they can make more strategic logistics decisions to lower their distribution costs and utilize their inventory more effectively.

5.3 Next Steps and Additional Research

We have identified a few additional steps which could add to the research we have done, which we did not have time to complete. We feel that these additional steps could add clarity to some of our results and provide more conclusions which will benefit our understanding of AgChem demand.

As noted above, we identified some geographic trends around whether AgChem is applied in the spring or fall. We feel that additional research in this area could uncover factors driving this trend. This research could seek to explain whether climatic factors such as moisture and temperature are affecting this trend. Other considerations might be whether the size of a retailer has any effect on the season in which they realize the most sales.

For our short-term demand sensing model, we only looked at one year of retailer sales. It would be valuable to analyze additional years as well. This might include seeing if the weather conditions ever hold back the application of AgChem even after the crops have been harvested. It would also be useful to determine whether the timing of harvest affects the volume of sales realized in the fall in comparison to those in the spring, assuming that the temperature has reached 50^{o} F.

5.4 Considerations for Future Applications

In addition to the afore-mentioned conclusions, we learned some valuable lessons which DAS should consider if they choose to use this type of analysis in the future. These considerations include identifying reliable, consistent data sources, collecting data for longer periods of time, and possibly applying the analysis to more granular geographies.

5.4.1 Data Quality

We received lots of data from DAS to use in these analyses. This data included DAS sales and shipments to distributors, distributor sales to retailers, and retailer EDI sales to growers. DAS also provided grower demographic data, retailer locations, AgChem pricing, and other applicable data. We also searched through masses of publically-available data. All of this data was valuable and useful in conducting this research. The format, granularity, and comprehensiveness of that data, however, could be improved to streamline this process in the future. We felt that excessive time was required to clean the data and that in the current formats it would be inefficient and unrealistic for DAS employees to perform the analysis.

If DAS wishes to carry this analysis forward, we recommend that they continue to work with distributors and retailers to obtain sales data which is as complete as possible. If they receive more complete data, they will be able to track exact volumes throughout the entire distribution chain. This will improve the accuracy of any analyzes they perform.

We also recommend that DAS determine the granularity at which they will analyze the data. They should consider time (daily, weekly, etc.) and geography (city, county, state). They should then identify their internal data sources and request that the data be formatted in a standard manner. While making these requests and formatting the data, DAS should consider how to incorporate external data including commodity prices and weather. Matching the data in a standard format will allow for much more efficient analysis with less risk of calculation error. For more accurate analyses, DAS could also likely find better sources for weather readings and other useful, external data. The publicly-available sources we used in our project were sufficient for our analysis. However, we had to approximate some weather data for smaller, more rural areas. If DAS would like to streamline the data gathering process while increasing accuracy, they should be able to obtain more accurate data in a more convenient format.

5.4.2 Application City not Retailer City

The most granular sales data we received was the retailers' EDI sales data to the growers. This data was helpful in establishing the last point of sale in the AgChem distribution chain. However, we had to make assumptions about the growers' locations based on the retailer location. We assumed that growers would not transport AgChem too far from the retailer. Thus, we used the retailer location as an approximation for the actual application site.
This level of granularity is not a problem if we are only trying to forecast sales through the retailer level using historic sales, commodity prices, etc. But if we are trying to use climatic factors and grower behavior to forecast specific geographic demand, we would need one of two things:

1. We would need to know how far growers were actually transporting AgChem from the retailer locations. If it is only a short distance, we can continue with our current assumption and use the retailer location for weather data etc. If we identify that growers transport the product a great distance from the retailer, the weather data for the retailer city would not be an appropriate factor for indicating short-term demand.
2. If the growers are transporting the product far enough away from the retailer that the weather patterns could be different, then we would need further information about the actual application site.

5.4.3 EDI for Multiple Years

DAS provided valuable retailer EDI sales data from September 2011 to April 2012. With this data, we were able to identify specific trends and correlations within that sales period. Since the sales of agricultural products varies from year to year due to economics, weather, etc. it would be useful to have additional years of retailer sales data. This data could be used to verify that the relationships we identified in the provided data hold true through other years.

Appendix

The following plots show the pattern described in section 4.6.1. Southern Illinois and Indiana seems to be predominantly spring application states where as Iowa and Northern Illinois are fall application states.

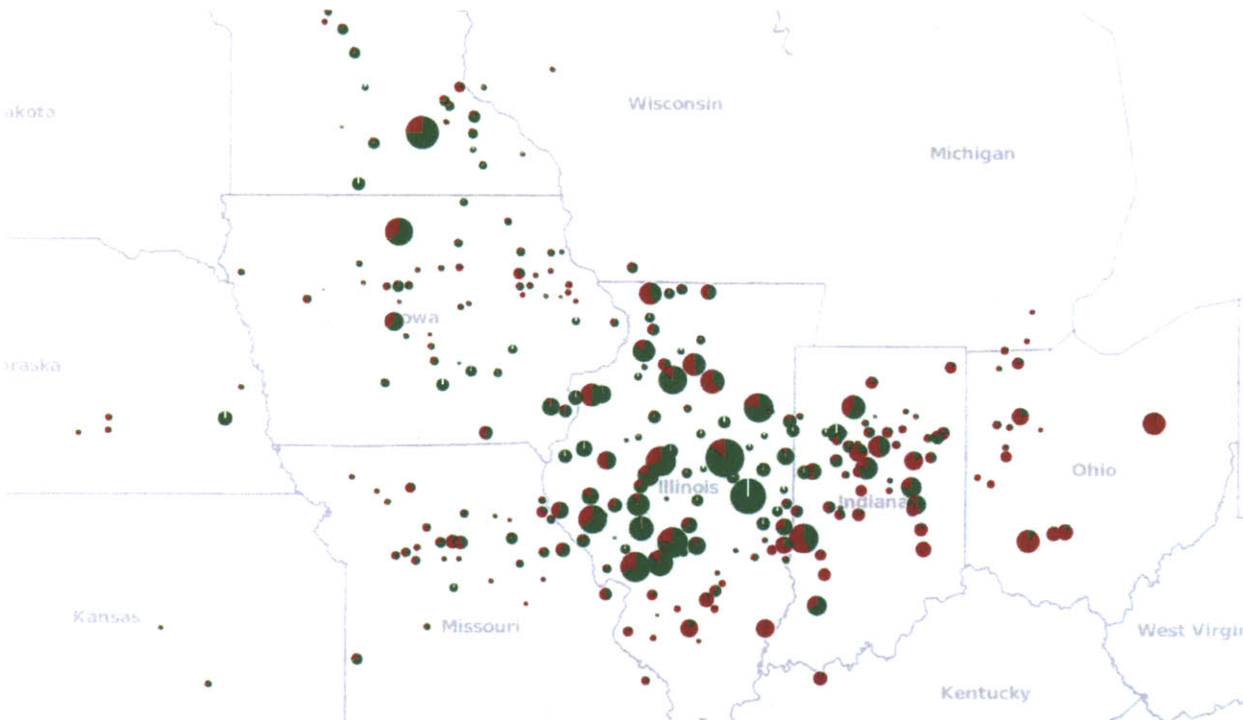

Figure 12 Geographical plot of fall and spring quantities shipped (2008-2009)

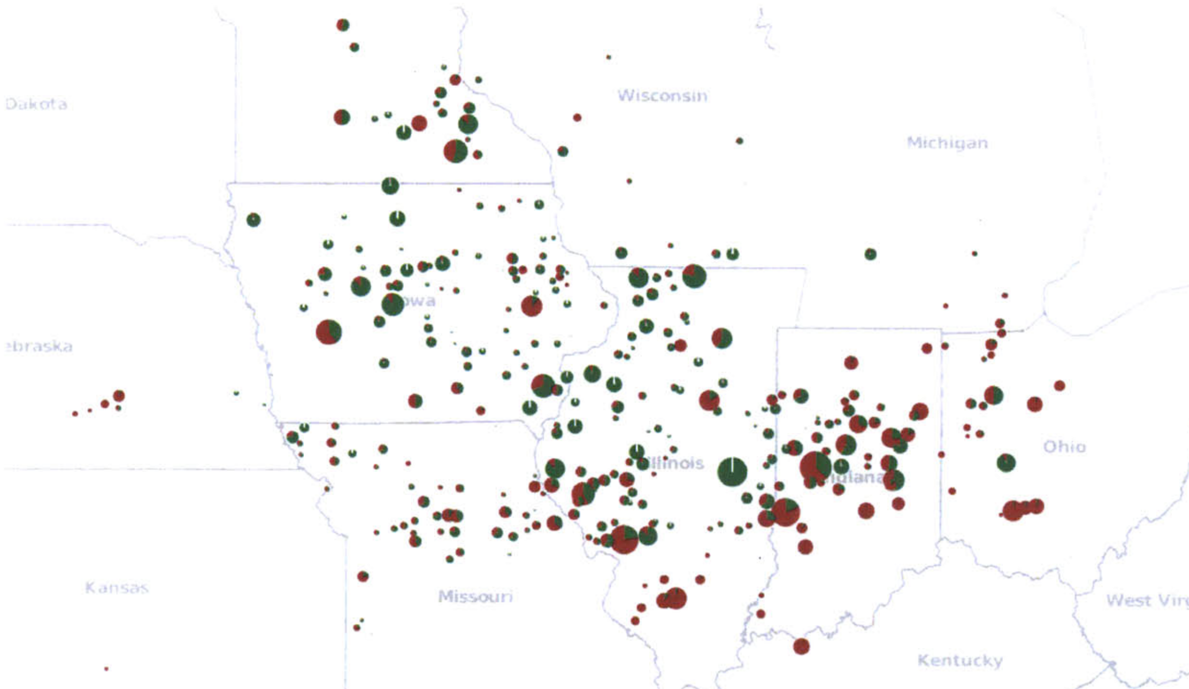

Figure 13 Geographical plot of fall and spring quantities shipped (2009-2010)

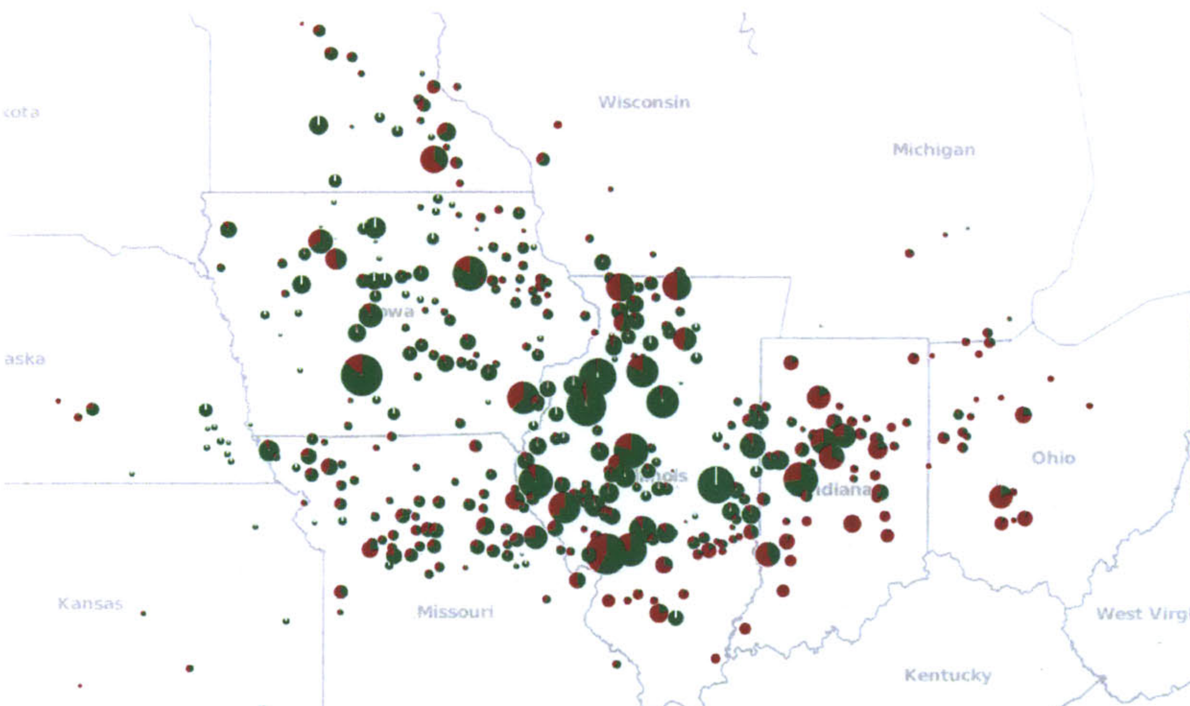

Figure 14 Geographical plot of fall and spring quantities shipped (2010-2011)

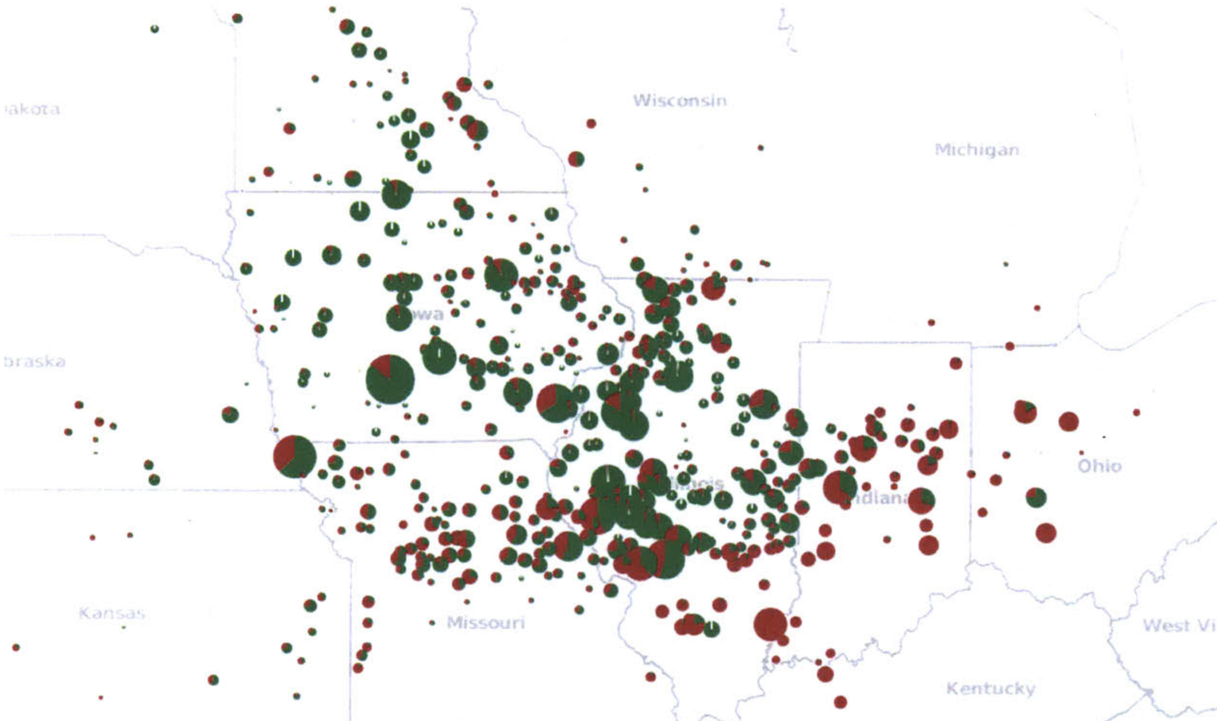

Figure 15 Geographical plot of fall and spring quantities shipped (2011-2012)

References

About Us | Dow AgroSciences. (n.d.).*Dow AgroSciences - A Dow Chemical Company*. Retrieved March 3, 2013, from http://www.dowagro.com/about/

Agarwal, A. & Holt, G. (2005). Reducing Inventory by Simplifying Forecasting and Using Point of Sale Data. MIT M. Eng. thesis

Brynjolfsson, E., Hitt, L. M. & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?. Retrieved February 10, 2013 from http://ssrn.com/abstract=1819486

Byrne, Robert. "Beyond Traditional Time-Series: Using Demand Sensing to Improve Forecasts in Volatile Times." *Journal of Business Forecasting.* Summer (2012): 13-19. Print.

FAO. (2000). Fertilizer requirements in 2015 and 2030. Food and Agriculture Organization of the United Nations. Rome, Italy

Griliches, Z. (1958). The Demand for Fertilizer: An Economic Interpretation of a Technical Change. *Journal of Farm Economics*, *40*(3), 591-606.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review, October*, 61-68.

Mehring A., & Shaw, B. (1944). Relationships Between Farm Income and Farmers' Expenditures for Fertilizer and a Forecast of the Commercial Demand for Fertilizer in 1944 and 1945, by States. American Fertilizer

National Research Council. (1995). Research at Land Grant Colleges of Agriculture. *Colleges of agriculture at the land grant universities a profile* (p. 58). Washington D.C.: National Academy Press.

Parthasarathy, N. S. (1994). Demand f*orecasting for fertilizer marketing*. Food and Agriculture Organization of the United Nations.

Products & Solutions | Dow AgroSciences. (n.d.). *Dow AgroSciences - A Dow Chemical Company.* Retrieved March 3, 2013, from http://www.dowagro.com/products/

Steutermann, S., Scott, F., & Tohamy, N. (2012). Building an Effective Demand-Planning Process. *Gartner*, *Web*. Retrieved April 2, 2013, from http://www.gartner.com/id=2085615

Tenkorang, F. (2006). *Projecting World Fertilizer Demand in 2015 and 2030.* Ann Arbor: ProQuest

Vance, A. (2012, March 22). Climate Corp. Updates Crop Insurance via High Tech. *Business Week, Web.* Retrieved February 10, 2013, from http://www.businessweek.com/articles/2012-03-22/climate-corp-dot-updates-crop-insurance-via-high-tech

Vail, E. (1927). Prices of Fertilizer Materials and Factors Affecting the Fertilizer Tonnage. Cornell University Ph.D. thesis. Ithaca