Applications of Machine Learning to Agricultural Land Values: Prediction and Causal Inference

by

Emrah Er

B.A., Anadolu University, 2007

M.A., Ankara University, 2010

M.A., North Carolina State University, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agricultural Economics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Abstract

This dissertation focuses on the prediction of agricultural land values and the effects of water rights on land values using machine learning algorithms and hedonic pricing methods. I predict agricultural land values with different machine learning algorithms, including ridge regression, least absolute shrinkage and selection operator, random forests, and extreme gradient boosting methods. To analyze the causal effects of water right seniority on agricultural land values, I use the double-selection LASSO technique.

The second chapter presents the data used in the dissertation. A unique set of parcel sales from Property Valuation Division of Kansas constitute the backbone of the data used in the estimation. Along with parcel sales data, I collected detailed basis, water, tax, soil, weather, and urban influence data. This chapter provides detailed explanation of various data sources and variable construction processes.

The third chapter presents different machine learning models for irrigated agricultural land price predictions in Kansas. Researchers, and policymakers use different models and data sets for price prediction. Recently developed machine learning methods have the power to improve the predictive ability of the models estimated. In this chapter I estimate several machine learning models for predicting the agricultural land values in Kansas. Results indicate that the predictive power of the machine learning methods are stronger compared to standard econometric methods. Median absolute error in extreme gradient boosting estimation is 0.1312 whereas it is 0.6528 in simple OLS model.

The fourth chapter examines whether water right seniority is capitalized into irrigated agricultural land values in Kansas. Using a unique data set of irrigated agricultural land sales, I analyze the causal effect of water right seniority on agricultural land values. A possible concern during the estimation of hedonic models is the omitted variable bias so we use double-selection LASSO regression and its variable selection properties to overcome the omitted variable bias. I also estimate generalized additive models to analyze the nonlinearities that may exist. Results show that water rights have a positive impact on irrigated land prices in Kansas. An additional year of water right seniority causes irrigated land value to increase nearly $17 per acre. Further analysis also suggest a nonlinear relationship between seniority and agricultural land prices.

Applications of Machine Learning to Agricultural Land Values: Prediction and Causal Inference

by

Emrah Er

B.A., Anadolu University, 2007

M.A., Ankara University, 2010

M.A., North Carolina State University, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agricultural Economics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Nathan P. Hendricks

Copyright

Abstract

This dissertation focuses on the prediction of agricultural land values and the effects of water rights on land values using machine learning algorithms and hedonic pricing methods. I predict agricultural land values with different machine learning algorithms, including ridge regression, least absolute shrinkage and selection operator, random forests, and extreme gradient boosting methods. To analyze the causal effects of water right seniority on agricultural land values, I use the double-selection LASSO technique.

The second chapter presents the data used in the dissertation. A unique set of parcel sales from Property Valuation Division of Kansas constitute the backbone of the data used in the estimation. Along with parcel sales data, I collected detailed basis, water, tax, soil, weather, and urban influence data. This chapter provides detailed explanation of various data sources and variable construction processes.

The third chapter presents different machine learning models for irrigated agricultural land price predictions in Kansas. Researchers, and policymakers use different models and data sets for price prediction. Recently developed machine learning methods have the power to improve the predictive ability of the models estimated. In this chapter I estimate several machine learning models for predicting the agricultural land values in Kansas. Results indicate that the predictive power of the machine learning methods are stronger compared to standard econometric methods. Median absolute error in extreme gradient boosting estimation is 0.1312 whereas it is 0.6528 in simple OLS model.

The fourth chapter examines whether water right seniority is capitalized into irrigated agricultural land values in Kansas. Using a unique data set of irrigated agricultural land sales, I analyze the causal effect of water right seniority on agricultural land values. A possible concern during the estimation of hedonic models is the omitted variable bias so we use double-selection LASSO regression and its variable selection properties to overcome the omitted variable bias. I also estimate generalized additive models to analyze the nonlinearities that may exist. Results show that water rights have a positive impact on irrigated land prices in Kansas. An additional year of water right seniority causes irrigated land value to increase nearly $17 per acre. Further analysis also suggest a nonlinear relationship between seniority and agricultural land prices.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my deep and sincere gratitude to my major professor, mentor, and committee chair, Dr. Nathan P. Hendricks. I am grateful for his kindness, patience, support and encouragement during my hard times. Numerous Skype sessions with him gave shape to this dissertation. I am deeply indebted for his guidance and availability. I also would like to thank to all members of the Hendricks family for their warm hospitality and friendship. Thank you Lindsay for your delicious meals, thank you Charli, Piper, and Ben for letting me to read you stories and playing games with me.

I would like to thank my other committee members, Dr. Jason Bergtold, Dr. Marcellus Caldas, and Dr. Mykel Taylor for their help and advice on my dissertation and Dr. Huston Gibson for agreeing to be part of the committee in the last minute. I also thank Dr. Peri da Silva for accepting being a proxy for Dr. Marcellus Caldas during my defense.

During data collection process numerous people helped me. I wish to acknowledge the help provided by Dr. Mykel Taylor and Dr. Leah Tsoodle. Thank you for providing me the Property Valuation Division data set and answering my questions about the data. I also would like to thank Dr. Rich Llewelyn for his assistance in the basis data collection.

I would like to extend thanks to my professors, friends and colleagues. I would like to thank Dr. Allen M. Featherstone, Dr. John Crespi, and all other members of the faculty for their kind help and support at various times of my study. I greatly appreciate the encouragement of Dr. Barry Goodwin, Dr. Hasan Şahin, and Dr. İrfan Civcir to pursue my Ph.D. degree. I am also thankful to Dr. Kemal Akoğlu, Dr. Tülay Ayyıldız Akoğlu, Dr. Krishna Pokharel, and all other fellow PhD students in the department for being great friends during my time in the US.

I thank the Turkish Fulbright Commission and Ankara University for financial support that allowed me to start my studies in the US and pursue it.

Last but not the least, I would like to thank the most important people in my life. I am grateful to my beloved Mother for her unconditional love and support through all my education life. I am thankful and indebted to my wonderful and lovely wife, Ilgım Seval Kurt Er, for her patience and love.

Dedication

To my beloved Mother and lovely Wife

Chapter 1

Introduction

Recent advancements in computer technology have allowed researchers and policymakers to analyze various big data sets. Even though conventional statistical and econometric techniques usually work well, researchers have developed new econometric methodologies to better handle this kind of data. Machine learning methods such as regression trees, ridge regression, and gradient boosting have become popular in the economics and agricultural economics literature in recent years.

Machine learning models are not designed to determine causal impacts, but to make predictions. Kleinberg et al. (2015) argue that many policy problems generally do not require causality and machine learning methods can easily be applied. Besides, recent methodological developments incorporate causality into machine learning models and allow researchers to make causal inferences (Athey and Imbens 2017).

One potential application of machine learning methods in agricultural economics is the prediction of agricultural land values. Accurate prediction of agricultural land prices is important to future land owners, farmers, and other agricultural land market participants. Better prediction models may help market participants and increase the market efficiency.

Recent commerical applications of machine learning algorithms aim to help consumers and farmers to make purchasing decisions. One reputed commercial application of machine learning algorithms in real estate markets is done by Zillow (2018)[1]. Using more than 100 million observations of houses, Zillow estimates a metric called as Zestimate, showing the valuation of a house. Similar to Zillow, AcreValue (2018)[2] of Granular Inc. and Accuacre (2018)[3] of Peak Soil Indexes are commercial applications of machine learning methods to predict agricultural land values.

As stated in Nickerson et al. (2012), farm real estate accounts for 84% of U.S. farm assets in 2009. Because of this, profitability in agricultural production is mostly capitalized in land values and therefore changes in land values will have a big impact on the financial well-being of agricultural producers. Land values are significant both to farmers and landowners. Therefore, understanding the determinants of land values and predicting them is important (Nickerson and Zhang 2014).

In this dissertation, we use recently developed machine learning algorithms to predict agricultural land values in Kansas and compare our results with standard econometric techniques. We also use double-selection least absolute shrinkage and selection operator (LASSO) to determine whether water right seniority is capitalized into land values.

This dissertation is laid out as follows. The second chapter presents the data used in the dissertation. A unique set of parcel sales from Property Valuation Division of Kansas constitutes the backbone of the data used in the estimation. Along with parcel sales data, we collected detailed basis, water, tax, soil, weather, and urban influence data. This chapter provides detailed explanations of various data sources and variable construction processes.

The third chapter presents multiple machine learning models for irrigated agricultural land price predictions in Kansas. Researchers, and policymakers use different models and data sets for price prediction. Recently developed machine learning methods have the power to improve the predictive ability of land valuation models. In this chapter we estimate several machine learning models for predicting irrigated agricultural land values in Kansas. Our results indicate that the predictive power of the machine learning methods are stronger compared to standard econometric methods. We find that median absolute error in extreme

[1] https://www.zillow.com/

[2] https://www.acrevalue.com/

[3] http://accuacre.com

gradient boosting estimation is 0.1312 whereas it is 0.6528 in simple OLS model.

The fourth chapter examines whether water right seniority is capitalized into irrigated agricultural land values in Kansas. Using a unique data set of irrigated agricultural land sales, we analyze the causal effect of water right seniority on agricultural land values. A possible concern during the estimation of hedonic models is omitted variable bias, so we use double-selection LASSO regression and its variable selection properties to overcome omitted variable bias. In order to estimate a casual relationship, in double-selection LASSO, we estimate two different LASSO regressions, a regression of land prices on all controls, and a regression of priority date on all controls, and one OLS regression with all the selected controls from LASSO regressions. LASSO regressions allow us to keep controls that have moderate sized effects on both land prices and priority date which might be dropped from the regression if we used simple LASSO.

We also estimate generalized additive models to analyze the nonlinearities that may exist. Our results show that water rights have a positive impact on irrigated land prices in Kansas. An additional year of water right seniority causes irrigated land values to increase nearly $17 per acre. Further analysis also suggest a nonlinear relationship between seniority and agricultural land prices.

Chapter 2

Data Description

In this chapter, we provide a detailed explanation about the construction of data used in the analysis. The data set is constructed by using multiple data sources and various data sets. The following subsections will provide information about how each data set is obtained, processed, merged, and used to construct variables.

2.1 Property Valuation Division Data

The Property Valuation Division (PVD) of the Kansas Department of Revenue (KDOR) is the source of land sales data. The data set covers sales between 1985 and 2015 and contains information on county code, parcel identification number, property class, parcel type, sales price, sales validity codes, sale date, agricultural use type, soil type, acres per soil type, irrigation, well depth, improvement values on the parcel and location. The raw data set includes 1,913,704 observations for 211,660 unique parcels. Due to missing information and duplicate entries, a cleaning of the data is required. An example of parcels data obtained from PVD is shown in Figure 2.1. Yellow dots on the figure denote the center point of the parcels. First we drop the parcels without any parcel identification number. Parcel identification number is required to find a parcel's location, so missing parcel identification numbers make finding data for these parcels difficult. The validity code in the data provides

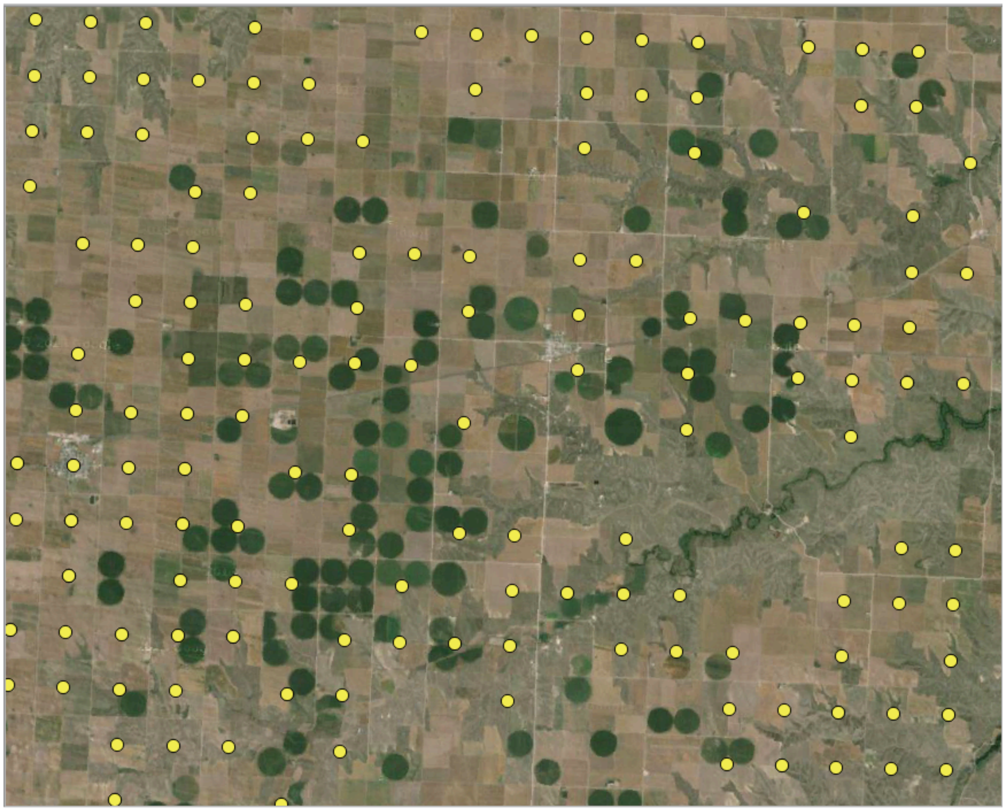

Figure 2.1: Parcels Data

information about the type of the sale. We opt to keep the arms-length sales, which are the transactions that occur based on self interest without any outside pressure, like government. Therefore, we exclude sales that are coded as not-open sales, forced sales, etc. PVD considers all parcel sales as valid unless there is sufficient information to show otherwise. Some sales in the data included multi-parcels sales, which are also considered as valid sales. If a parcel had significant changes after the sale, these parcels are coded as 3 in the data and also considered valid arms-length sales. We also include Partial Interest, Other, Immediate Family, and Absolute Auction sales into arms-length sales. Summary statististics in Table 2.1 show that Multiparcel sales and Valid sales constitute nearly 80% of the total sales. Following the literature, we dropped parcels that are smaller than 35 acres in total. Since it is not very likely to do farming on these small parcels, removing these data is reasonable. Some parcels had abnormal resales. We dropped parcels that are sold within the same month. Since our interest is agricultural land values, parcels that have only buildings are also dropped.

Table 2.1: Summary Statistics for Selected Variables in PVD Data

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Valid Sale	0.3166	0.4652	0.0000	1.00
Multiparcel Sale	0.4985	0.5000	0.0000	1.00
Change After Sale	0.1393	0.3462	0.0000	1.00
Partial Interest Sale	0.0085	0.0919	0.0000	1.00
Other Sale	0.0060	0.0770	0.0000	1.00
Immediate Family Sale	0.0259	0.1589	0.0000	1.00
Absolute Auction Sale	0.0052	0.0718	0.0000	1.00
Percent of Dryland	0.4807	0.4080	0.0000	1.00
Percent of Irrigated Land	0.0586	0.2030	0.0000	1.00
Percent of Nativegrass	0.3981	0.4004	0.0000	1.00
Percent of Tamegrass	0.0626	0.1884	0.0000	1.00
Price per Acre	1786.2468	2586.6789	33.5093	21551.72

Probably due to problems in record keeping, there were some duplicate entries in the data. We deleted these duplicate observations. Some of the numeric variables, such as total dryland acres, were recorded as rounded. Because of this rounding problem, the total acres values did not add up. We used acres per soil type to calculate the areas for different land types (dryland, irrigated land, etc.). We used these new values to calculate total acres of the parcels. The total acres variable shows the agricultural land acres of the parcel and does not include the home acres. For every parcel, we calculated irrigated land, dryland, native grassland, and tame grassland acres and their percentages along with their soil properties. Using QGIS, we created buffers around each parcel to calculate the average nearby sale prices for different buffer sizes and time periods. We also calculated average land values within counties for various time periods to use as predictors in our estimations.

After cleaning the data, we calculated price per acre of the land. "[I]n the United States, land is not usually traded separately from the structures placed upon it, so the observed prices reflect the values of both the land and its structural improvements. This causes no problems at the theoretical level, but it does require that the hedonic price equation adequately control for structural characteristics" (Freeman III, Herriges, and Kling 2014, 317). To overcome this problem, previous researchers used different methods. For example,

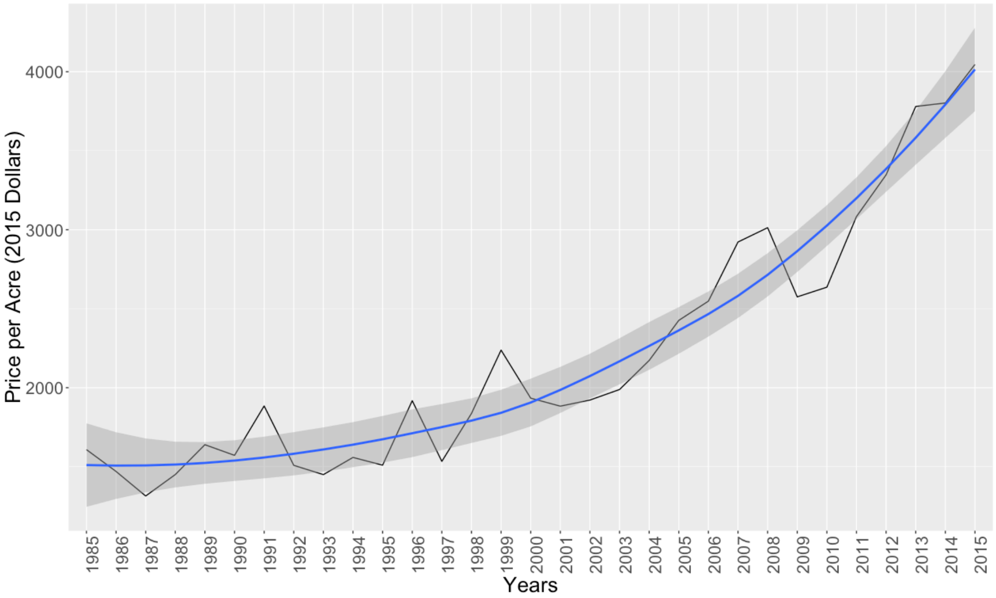

Figure 2.2: Average Land Prices in Kansas

Palmquist and Danielson (1989) used dummy variables representing structures on the land. Guiling, Brorsen, and Doye (2009) and Zhang and Nickerson (2015) subtracted the value of improvements from the sales price. Following a similar approach to Guiling, Brorsen, and Doye (2009) and Zhang and Nickerson (2015), we subtract total improvements from the sale price. Due to errors in data reporting, we ended up with some parcels having an extremely high price per acre and some having negative price per acre. At this point we chose to drop the upper and lower 1% of the observations. The final PVD data has 841,105 observations and 93,500 unique parcels.

Figure 2.2 shows the price per acre in 2015 dollars which were calculated using PVD data. As seen from the figure, there is an increasing trend. The average price per acre in 1985 was around $1,600, but it increased to $4,000 in 2015. Figure 2.3 shows the trend for irrigated agricultural lands only. The average price per acre for irrigated land in 1985 was around $1,550, but it increased to $6,700 in 2015.

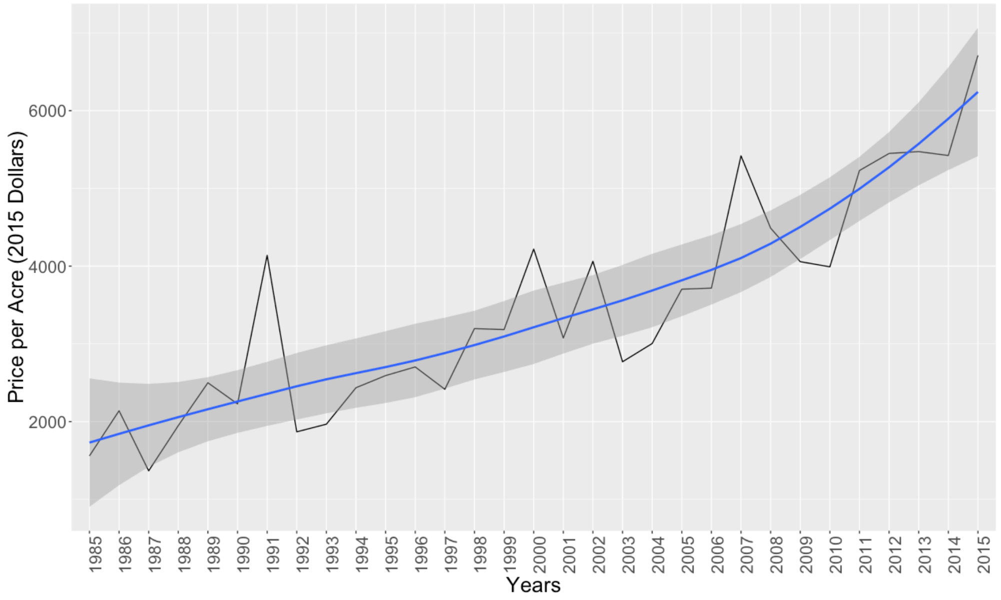

Figure 2.3: Average Irrigated Land Prices in Kansas

2.2 Basis Data

Basis data is obtained from AgManager.info. Data is collected from different grain elevators around Kansas and neighboring states. Basis data includes the basis for corn and wheat. We use this data to create rasters using kriging methods.[1]

Figure 2.4 shows average corn basis kriging results for 2015. Basis data for each parcel is extracted from this raster. We use a 3 year moving averages of the basis data. Since basis data do not go back to 1985, we use 3 year moving average of 1999, 2000, and 2001 for all years prior to 2001. Basis is defined as the difference between the cash price and future price of a commodity. Basis, an indicator for the current local demand of the commodity, is usually negative for agricultural commodities. In southwestern Kansas, large numbers of cattle feeding operations create a huge demand for corn. Therefore, we see a better basis in this region (Figure 2.4). Table 2.2 shows the average corn and wheat basis for 1985 and 2015. We see that both corn and wheat basis increased between 1985 and 2015. McNew and Griffith (2005) showed that new ethanol plants increased the corn basis prices in the area, which in turn can translate into higher land values. So we expect to see higher land prices

[1] We used `autoKrige` function in `automap` (Hiemstra 2013) package in R (R Core Team 2018).

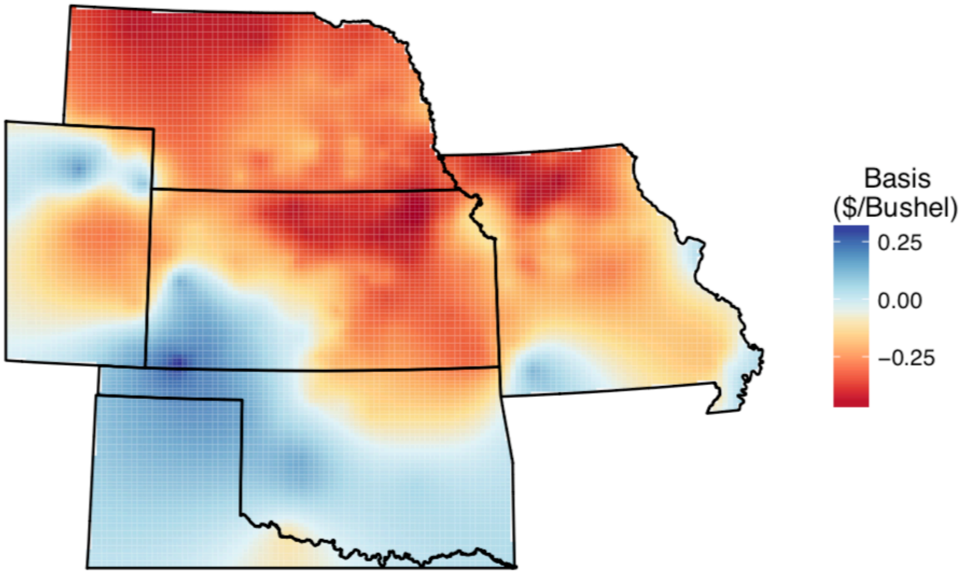

Figure 2.4: Corn Basis in 2015

Table 2.2: Summary Statistics for Basis in 1985 and 2015

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Corn Basis in 1985	-0.3203	0.0584	-0.4317	-0.1871
Wheat Basis in 1985	-0.5274	0.0254	-0.5925	-0.4211
Corn Basis in 2015	0.0250	0.1424	-0.1909	0.4920
Wheat Basis in 2015	-0.3513	0.0600	-0.5361	-0.1938

in areas with higher basis values.

2.3 High Plains Aquifer Data

High Plains Aquifer Section-Level Data is obtained from Brownie Wilson via personal communication. The data set contains two important section-level variables: predevelopment depth to water and predevelopment saturated thickness. Since data is at the section-level these variables are merged using section levels of each parcel.

Saturated thickness is the vertical thickness of the aquifer which is full of water and it is an approximation for the amount of water available. Depth to water, on the other hand, is defined as the depth of the water table below the earth's surface.

While setting or managing water use policies and regulations, predevelopment saturated thickness is commonly used. This "value is the estimated saturated thickness before the withdrawal of significant amounts of groundwater, and is taken as the starting point against which the amount and rate of any depletion is measured" (Schloss and Buddemeier 2000). The averaged 2013 - 2015 saturated thickness of the High Plains aquifer in Kansas ranges from nearly zero to over 400 feet. Estimated decreases in saturated thickness for 1997 - 1999 varies between 0 to more than 60 percent (Schloss and Buddemeier 2000). Table 2.3 shows the selected summary statistics for the High Plains Aquifer data. As seen from the table, predevelepment saturated thickness varies between 6.3 and 602.2 feets for the sample we have. Similarly, predevelopment depth to water varies between 0 and 261.4 feets. We use predevelopment values rather than the current values to avoid any endogeneity problems.

Other two closely related variables from this data set used are hydraulic conductivity and specific yield. Hydraulic conductivity shows how easily water can move through pores or fractures in the soil. Specific yield shows the volume of water that can be released from a unit volume of saturated ground. In other words, it shows how much water is available for use (Heath 1983).

Depletion of the high plains aquifer causes irrigated lands to return to dryland production. As stated in Torell, Libbin, and Miller (1990) this causes land values to decline. Depth to water is directly related to the cost of pumping. If depth to water is high then we expect pumping costs to increase, which will lower the profits from farm. This in turn will cause land prices to fall. Since saturated thickness shows the abundance of water in the land we expect to see high land prices where saturated thickness is also high. Hydraulic conductivity and specific yield both show how quickly the water replenishes. Therefore we expect to see higher land values in regions with high hydraulic conductivity and specific yield.

Table 2.3: Summary Statistics for High Plains Aquifer Data

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Predevelopment Saturated Thickness	161.8682	116.9625	6.3301	602.1499
Predevelopment Depth to Water	88.6600	57.6912	0.0000	261.3501
Hydraulic Conductivity	73.8449	28.0427	10.0000	196.0000
Specific Yield	16.2744	3.6892	5.0000	25.0000

2.4 Macroeconomic Data

We use Consumer Price Index data from U.S. Bureau of Labor Statistics to deflate prices. We also use population data from U.S. Census Bureau to calculate population densities and population growth for counties. Population densities and population growth are associated with urban influence. We expect to see high land prices in regions with high population density and high population growth rates. Table 2.4 shows selected population data for 1985 and 2015. As seen from the table, mean population density in Kansas increased, whereas the population growth has decreased from 1985 to 2015.

Table 2.4: Summary Statistics for Population Data

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Population Density - 1985	23.8774	46.5368	2.1518	384.4255
Population Density - 2015	26.2034	61.4275	1.6615	512.8500
Population Growth - 1985	-0.0117	0.0152	-0.0695	0.0373
Population Growth - 2015	-0.0073	0.0129	-0.0568	0.0223

2.5 Mill Levies Data

Mill Levies data are also from PVD of KDOR. This data set includes county level tax levies for the years considered in the analysis. Since levies data only goes back to 1987, 1987 values are used for 1986 and 1985. Mill levy is a tax rate applied to assessed value of a property and used by local governments to cover annual expenses. One mill is $1 per $1,000 of assessed valuation so for example, if assessed property value is %11.5 of appraised value, for a property appraised at $100,000, the assessed value will be $11,500. If the mill rate is 10 mills, then the total property tax amount will be $10,000/1,000 X 10 = $100.

The property tax is one of the fiscal instruments that governments can use to control land use patterns. Usually we expect to see a reduction in population density in urban areas when there is an increase in property taxes. So we can say that mill levies are an indicator for urban influence on land prices. We expect to see higher prices of land where tax rates are higher. On the other hand, high mill levies and land prices might be negatively correlated. Higher tax rates may lead land prices to decrease since the buyers of the lands will pay the tax.

Table 2.5 shows the average mill levies for Kansas for 1985 and 2015. The mean mill levies increased from 116.6 to 150.6 from 1985 to 2015.

Table 2.5: Summary Statistics for Mill Levies Data

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Mill Levies - 1985	116.5618	24.9535	39.120	176.38
Mill Levies - 2015	150.5502	22.6226	89.416	211.64

2.6 Google Maps Data

Distance and commute time data are calculated using the Google's "Google Maps Distance Matrix API". First, using the geolocation of the parcels in the PVD data and populated areas data from Census, we calculated distances from each parcel to different populated areas. Figure 2.5 shows an example map of the distances calculated for closest city with a population greater than 10,000. Using the Google Maps Distance Matrix API (Google 2018), driving distances and commute times to these populated areas are downloaded. Figure 2.6 shows the driving time to the closest city with a population greater than 10,000. Driving distances to closest population areas shows the urban pressure on land. We expect to see higher land prices for parcels that are easier to access and close to urban areas. Table 2.6 shows summary statistics for Google Maps data. All distances are in kilometers and all

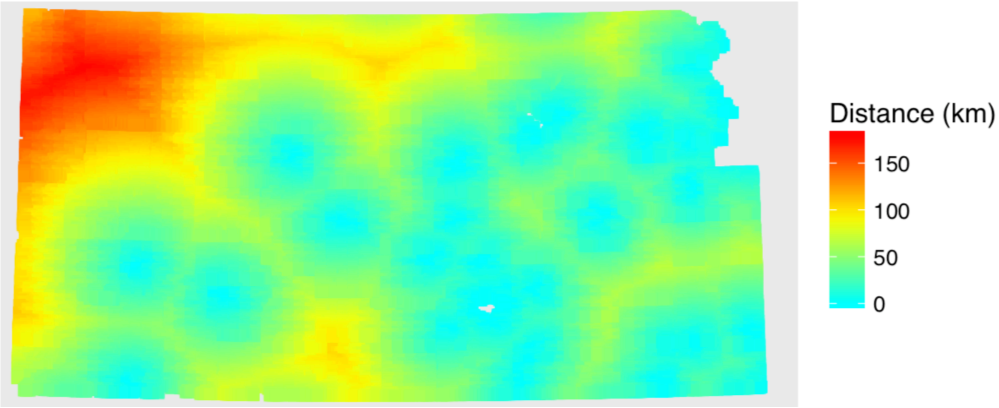

Figure 2.5: Distance to Closest City with Population greater than 10,000

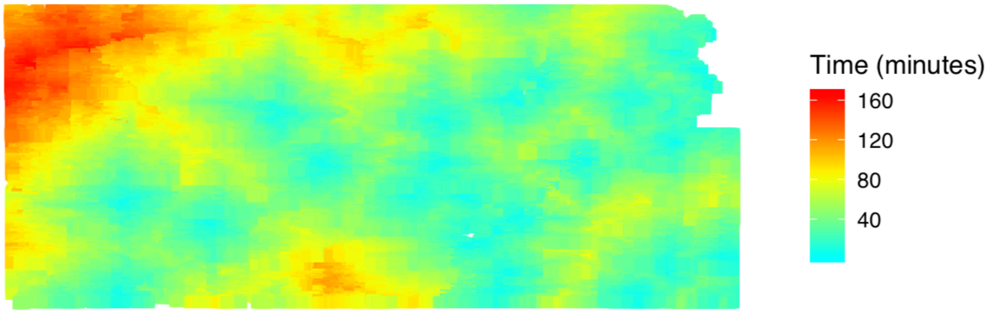

Figure 2.6: Driving Time to Closest City with Population greater than 10,000

times are in hours. As seen from the summary statistics, mean distance from parcels to a city with population greater or equal to 10,000 is around 71 kilometers and time is around 1 hour. Average distance to a city with population greater or equal to 1 million is around 778 kilometers and time is around 7 and a half hours.

2.7 Soil Data

The source for soil data is the Soil Survey Geographic Database (SSURGO) (Soil Survey Staff 2016) and it includes estimated and measured data on physical and chemical soil properties, and soil interpretations. SSURGO database provides data for map units (**mapunit**), soil components (**component**), and soil horizons (**chorizon**). Figure 2.7 shows an example of the map unit and Figure 2.8 shows the relationship between **mapunit**, **component**,

Table 2.6: Summary Statistics for Google Maps Data

Variable	Mean	Std.Dev.	Min.	Max.
Distance to 10K	71.5520	46.6635	0.5330	268.7100
Distance to 20K	81.5055	45.6279	1.5540	283.8270
Distance to 40K	144.9264	93.7076	1.6700	408.3590
Distance to 50K	172.8108	109.1627	4.7580	478.2660
Distance to 100K	189.6162	105.9231	7.1600	478.2660
Distance to 200K	208.5001	105.4531	12.1310	484.7440
Distance to 500K	421.5448	88.4851	181.2560	676.6920
Distance to 1M	777.9321	142.6432	512.2430	1210.2120
Time to 10K	0.8518	0.4721	0.0147	2.7800
Time to 20K	0.9557	0.4548	0.0550	3.0439
Time to 40K	1.5870	0.9322	0.0606	4.0733
Time to 50K	1.8289	1.0322	0.1169	4.3750
Time to 100K	1.9670	0.9945	0.1633	4.3750
Time to 200K	2.1432	0.9991	0.2281	4.9789
Time to 500K	4.0418	0.8020	1.6967	6.0650
Time to 1M	7.4050	1.1907	4.7664	11.0547

and **chorizon**. As seen from Figure 2.7, each map unit polygon is comprised of different soil components and each soil component is associated with multiple horizons. Map unit polygons are connected to a record in the map unit table via a key called **mukey**. Each map unit is linked to multiple records in the component table and referenced with the key **cokey**. Each component, on the other hand, is linked to multiple records in the horizon table and referenced with the key **chkey** (UC Davis California Soil Resource Lab 2018; Esri 2018).

We merge PVD and SSURGO data sets and calculate soil variables for all parcels. For a given parcel, we calculate average soil chracteristics for each agricultural land type. For some string variables in the soil data, we create dummy variables. For example, `hydgrp` shows groups of soils having similar runoff potential under similar storm and cover conditions and it is a string variable with 7 different categories. We convert this data into 7 different dummy variables. `weg`, which shows susceptibility to soil blowing, is also a string variable in the SSURGO data set. We aggregate these kind of variables by the dominant characteristic for the **mapunit**. After calculating soil properties per agricultural land type, we need to aggregate the data since we want to get rid of missing observations in soil characteristics per

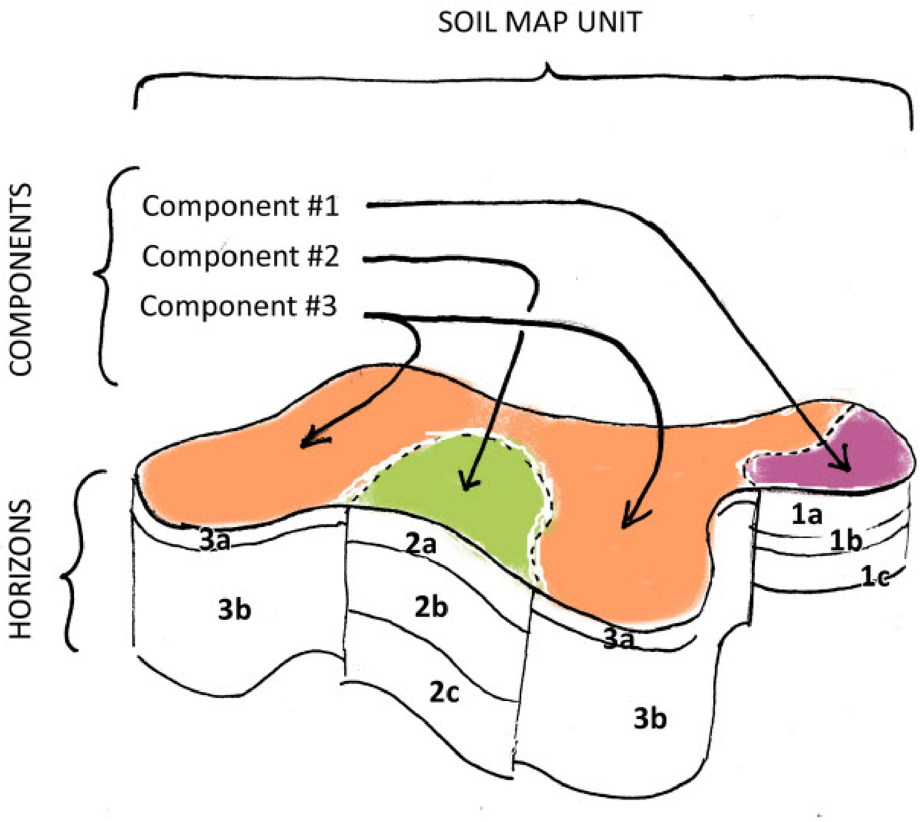

Figure 2.7: Soil Map Unit (Esri, 2018)

agricultural land type.

The total number of soil variables in our data set is 985. Soil data include many different variables such as elevation, soil organic carbon, bulk density, clay percentage, silt percentage, slope, etc. Table 2.7 shows the summary statistics for selected soil variables. We expect soil variables to have different effects on land price. For example, soil organic carbon, which is measured in gram C per square meter, is very important for plant growth since it affects available nutrients in the soil; therefore our expectation for soil organic carbon is to increase land price. On the other hand, for some soil variables we expect negative impacts. For example, bulk density affects the movement of air and water in the soil and a larger value indicates poorer quality soils so we expect high bulk density to decrease the price of land (Hendricks 2018). Slope, which is defined as the difference in elevation between two points and expressed as a percentage of the distance between those points, of the parcels varies

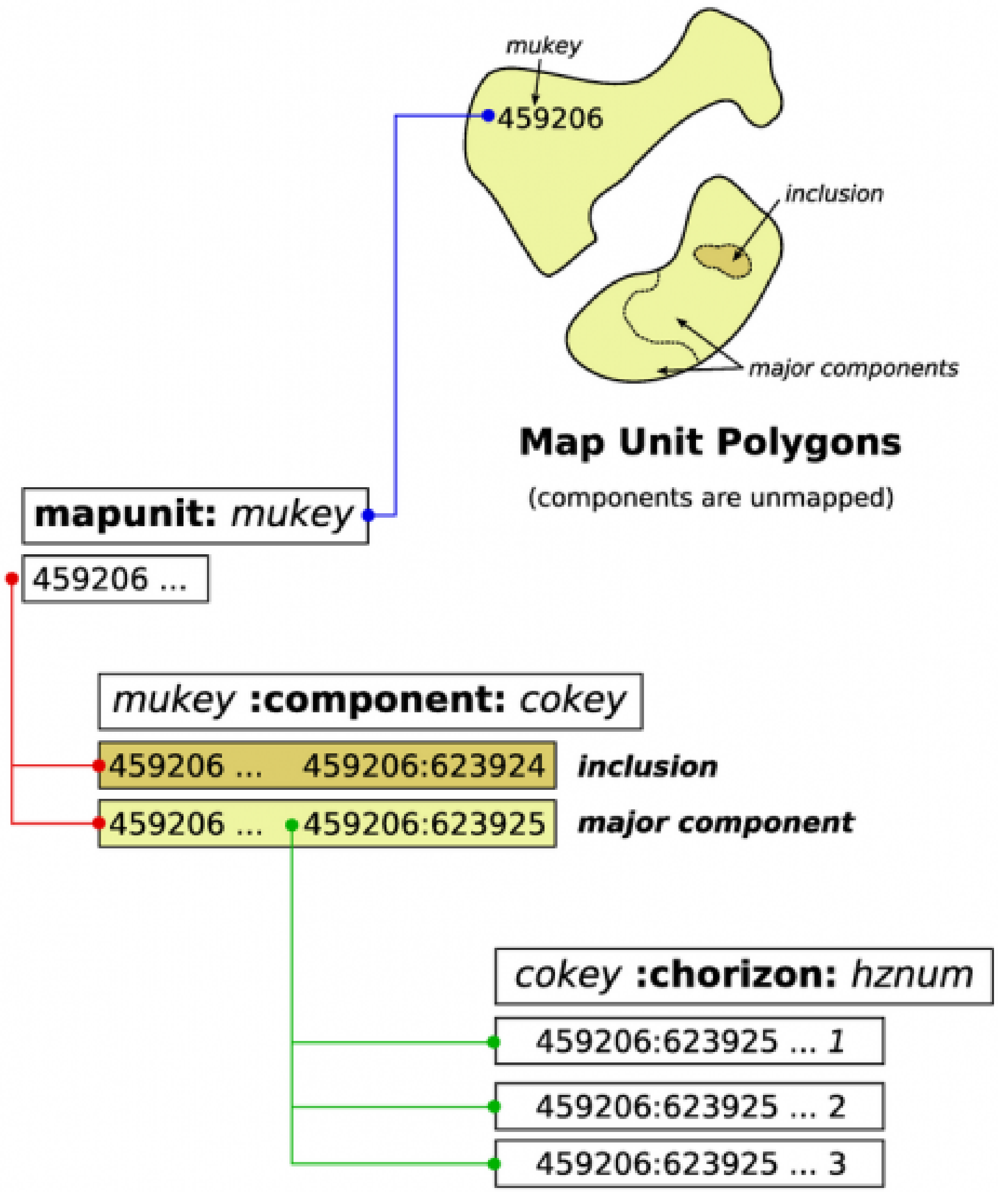

Figure 2.8: SSURGO Table Diagram (UC Davis California Soil Resource Lab 2018)

Table 2.7: Summary Statistics for Soil Data

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Elevation	564.0949	258.2822	160.0000	1631.000
Soil Organic Carbon	866.0577	241.2276	90.0305	1988.646
Bulk Density	1.4000	0.0821	1.1807	1.758
Slope	4.0770	2.8472	0.0000	33.000

between 0 and 33. Even though it is not an indicator of soil quality, it has some impacts on crop productivity and on land prices. We expect high slopes to decrease land prices. In estimations we use the log of slope since the distribution is highly skewed (Hendricks 2018).

2.8 Water Rights Data

Water rights data are obtained from Water Rights Information System (WRIS). “A notable complexity with water rights is how they can overlap each other. For example, a single water right may have multiple uses of water and multiple points of water diversion. Likewise, a single point of diversion may be associated with multiple water rights” (Wilson et al. 2005, 9). Since water rights data is very complex, it will be better to provide a detailed explanation of the construction of this data set.

Water rights data that is associated with parcels is constructed from 2 different databases; Place of Use (POU) and Water Information Management and Analysis System (WIMAS). From these databases, we extract 4 different data sets; Place of Use, WIMAS Acres Irrigated, Points of Diversion, and Group Summary.

Water rights data is depicted in Figure 2.9. Blue dots in this figure are the points of diversion and blue squares are the place of use areas. First we extract data from Points of Diversion which combines several relational tables from Water Rights Information System (WRIS) dealing with water rights, uses made of water, points of diversion, authorized rates and quantities. In this data set, we only keep irrigated and active wells and then we create dummy variables for source, right type, and priority. Then we merge the data with the WIMAS SIT table which is a collection of WRIS tables listing the place(s) of use for all water rights and the 40-acre tract designations associated with irrigation-based water rights. By merging these two data sets we found the water rights on specific Place(s) of Use. Finally we merge group summary data and Place(s) of Use polygons data and then aggregate all data to township-section-range (TRS) level using weighted means.

After merging all water rights data, we create buffers around the parcels. We use total agricultural land area of each parcel to calculate the radius of the buffer specific to that

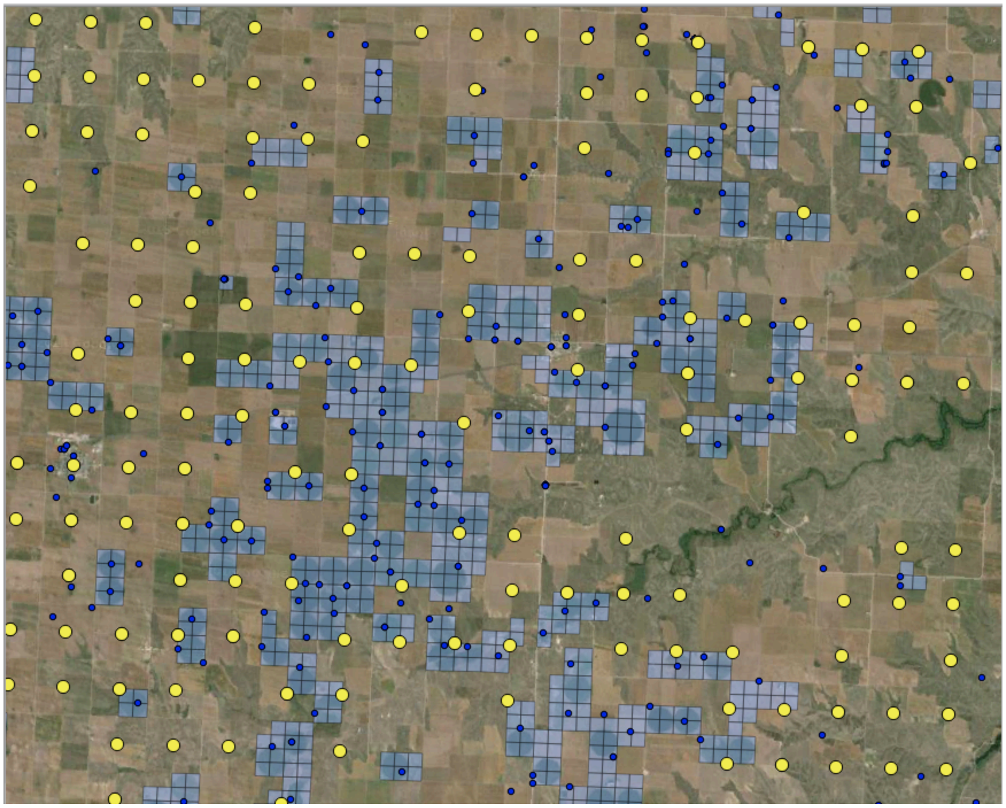

Figure 2.9: Water Rights Data

parcel. Using these buffers we can get intersections between parcels and water rights data. Figure 2.10 shows the intersections (red areas) of parcel buffers and water rights data. Then we calculate the areas for these intersecting polygons and aggregate water rights to parcel level. We use weigted mean (area of intersections) to calculate these values.

Table 2.8 shows the summary statistics for the selected water rights variables. In 1992, enactment of an order created two groups (senior and junior) of water rights, with 1 October 1965 being the dividing priority date (Peck 2002). Priority years are the number of years after the enactment of Kansas Water Appropriation Act. 88% of the parcels in our sample use groundwater whereas 12% use surface water for irrigation. As seen from the table, 17% of the parcels have senior water rights. Average priority years for the sample is around 32 years and authorized irrigation varies between 0 inch to 1875 inches.

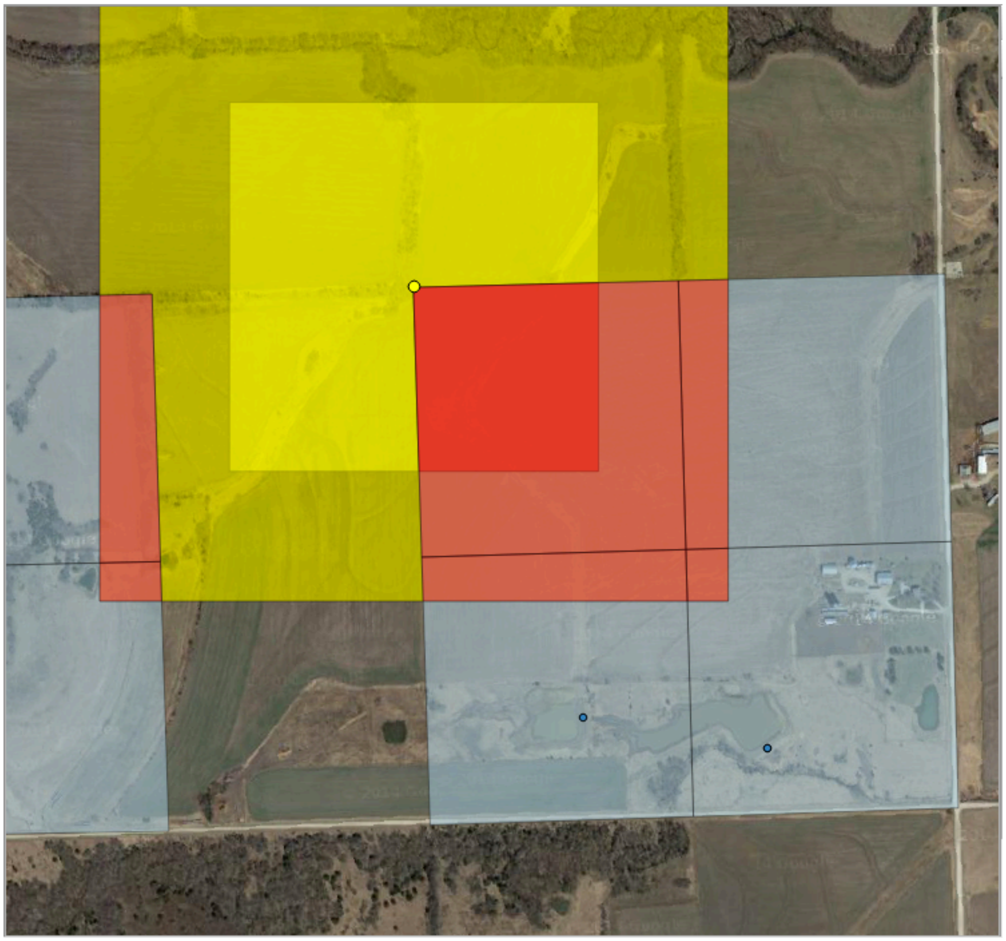

Figure 2.10: Intersection of Buffers around Parcels and Water Rights

Table 2.8: Summary Statistics for Water Rights Data

Variable	Mean	Std.Dev.	Min.	Max.
Source of Water Supply - Groundwater	0.8827	0.3120	0.000	1.0000
Source of Water Supply - Surface Water	0.1173	0.3120	0.000	1.0000
Senior Rights	0.1744	0.3116	0.000	1.0000
Junior Rights	0.8080	0.3269	0.000	1.0000
Priority Years	32.7383	14.3551	-2.474	70.2986
Authorized Inches per Acre	12.2187	26.7417	0.000	1875.9799

2.9 Weather Data

Weather data is obtained from PRISM Climate Data (PRISM Climate Group 2016). This data set includes daily maximum and minimum temperatures along with the daily precipitation for years 1981 to 2015. The data set is distributed as raster files and this allows us to extract temperatures and precipitation for each parcel using the geolocation of parcels.

After extracting temperatures and precipitation values for each parcel, following Snyder (1985), we first calculate degree days for different threshold levels using Equation 2.1.

$$DD = \frac{(M - T) \times (\pi/2 - \theta) + W \times cos(\theta)}{\pi} \tag{2.1}$$

where $M = (T_{max} + T_{min})/2$, T is threshold temperature, $W = (T_{max} - T_{min})/2$, and $\theta = arcsin(\frac{T-M}{W})$.

Extreme degree days, degree days that are greater than 30°C, measure the extreme temperatures that are not beneficial to crop growth whereas growing degree days, degree days that are between 10°C and 30°C, measures the days that are beneficial to crop growth. We expect extreme degree days to decrease but growing degree days to increase the land prices.

Besides degree days variables, we also calculate variables for evapotranspiration and vapor pressure deficit. Evapotranspiration is the combination of two different processes. Evaporation is defined as the loss of water from soil. Transpiration is the loss of water contained in the plants. These two processes occur simultaneously and are affected by different factors such as radiation, temperature, humidity, wind speed, crop type, soil salinity, etc. Vapor pressure deficit, on the other hand, is defined as the difference between the amount of moisture in the air and the amount of moisture when the air is saturated. We follow Hendricks (2018), to calculate evapotranspiration and vapor pressure deficit variables. Evapotranspiration was calculated as the reference evapotranspiration which is independent of any crop or soil characteristics. It is evapotranspiration of a well-watered, actively grown grass (Hendricks 2018, 551) If evapotranspiration is high this means that the soil and the plant are losing water faster, which may cause water strees. We expect to see lower land values in areas with high evapotranspiration rates. High values of vapor pressure deficit can cause plant stress, so we can expect to see lower land prices where vapor pressure deficit is high.

Table 2.9: Summary Statistics for Weather Data

Variable	**Mean**	**Std.Dev.**	**Min.**	**Max.**
Precipitation	543.8233	125.0802	308.9287	773.6423
Minimum Temperature	13.8627	1.4287	10.0039	16.3036
Maximum Temperature	27.7847	0.7579	25.8845	30.1585
Vapor Pressure Deficit	1.1424	0.1237	0.8799	1.4145
Evapotranspiration	861.3364	51.1052	748.1499	975.3592
Growing Degree Days	1986.1964	119.7528	1615.2207	2269.2711
Extreme Degree Days	76.8763	19.8410	33.1848	143.3922

All variables from this data are calculated within the growing season (April 1 – September 30) for each year and then we aggregate them to calculate the climate averages. Table 2.9 shows the summary statistics for the selected weather variables. One may note the values in minimum temperature. The positive value for the minimum of minimum temperature is a result of aggregation over a long period of time. Negative extreme values are cancelled out when we average the temperature values.

`precintcon` (Povoa and Nery 2016) package in R (R Core Team 2018) provides functions to analyze the precipitation intensity, concentration and anomaly. The package provides functions for calculation of the following quantities; Concentration Index (CI), Precipitation Concentration Index (PCI), Precipitation Concentration Degree (PCD), and Precipitation Concentration Period (PCP). For each parcel in the data, we calculate these values. We also calculate the Shannon diversity index discussed in Tremblay et al. (2012). The Shannon diversity index is defined as

$$\text{SDI} = \frac{[-\sum \pi \ln(\pi)]}{\ln(n)} \tag{2.2}$$

where $\pi = Rain/PPT$ is the fraction of daily rainfall relative to the total rainfall in a given time period and n is the number of days in that period. SDI $= 1$ implies complete evenness (i.e., equal amounts of rainfall in each day of the period) whereas SDI $= 0$ implies complete unevenness (i.e. all rain).

Variability in rainfall is known to affect crop yields. Periods of excess rain or periods of drought cause crop stress, which in turn affects the yield and therefore land prices. With even rainfall distribution we expect to see higher yields, which in turn will increase land prices.

Chapter 3

Machine Learning for Prediction: An Application to Agricultural Land Values in Kansas

3.1 Introduction

Machine learning methods in economics are applied to various areas, such as economic growth prediction (Basuchoudhary, Bang, and Sen 2017), bankruptcy prediction (Zięba, Tomczak, and Tomczak 2016), demand prediction (Bajari et al. 2015), forecasting electricity prices (Ludwig, Feuerriegel, and Neumann 2015), and wine price prediction (Yeo, Fletcher, and Shawe-Taylor 2015). One popular area is price predictions in real estate markets using hedonic models (Limsombunchao 2004; Caplin et al. 2008; Yoo, Im, and Wagner 2012; Belloni, Chernozhukov, and Hansen 2014b; Mu, Wu, and Zhang 2014; Park and Bae 2015; Nowak and Smith 2016; Ho 2017).

In agricultural economics, machine learning methods can be used for different applications such as; weather forecasting, crop yield prediction and crop selection, irrigation systems, crop disease prediction, and agricultural policy and trade (Coble et al. 2018). Recent studies in agricultural economics literature applied these techniques to predict farm size change (Oudendag, Szlávik, and Veen 2012), profitability in dairy farming (Yli-Heikkilä et al. 2015), demand for new credit (Ifft, Kuhns, and Patrick 2018), and to determine if a consumer is vegetarian (Lusk 2017).

Using a hedonic pricing model, we apply different machine learning techniques to predict irrigated agricultural land prices in Kansas and compare the results to standard econometric methods. The results show that machine learning methods give better predictions compared to standard econometric methods. Using data from Property Valuation Division we train different machine learning algorithms and make predictions. We find that the extreme gradient boosting algorithm (xgboost) gives the best out of all predictions. The median absolute error for xgboost model is 0.13 whereas for simple OLS model it is 0.65. Median absolute error of 0.13 means that half of our predictions are 13% of the real price and half of them are off by more than 13%. Further, using point of diversion data from WRIS we train another extreme gradient boosting model to predict the agricultural land values in Groundwater Management Districts. The median absolute error for in-sample forecasts is around 0.02 indicating that half of our predictions are 2% of the real price. Using this model, we perform out-of-sample predictions for point of diversion data. Results show that the predicted nominal price varies between \$804 and \$4642 per acre.

3.2 Model

Following Rosen (1974), the hedonic model can be described as follows. Let $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$ denote n different attributes of a differentiated market good. For farmlands these attributes include soil quality, location, improvements, amenity levels, etc. The fundamental hedonic model assumes a functional form that relates attributes of a good to its price and can simply be represented as $p = f(\mathbf{Z})$. The marginal effect of one specific characteristic of the good on the price of the good can be found by simply taking the partial derivative of this function, $\hat{p}_i = \frac{\partial f(\mathbf{Z})}{\partial z_i}$. $\hat{p}_i$ gives "the additional amount that a purchaser must pay to move to a bundle with one more unit of that characteristic, holding all other things constant" (Miranowski and Hammes 1984, 746).

$\hat{p}_i$ gives us the marginal implicit price of z_i under certain assumptions. In the hedonic model, it is assumed that there is a single market and consumers have the knowledge of all available options and maximize their utility by choosing from continuous various bundle of characteristics. According to Miranowski and Hammes (1984), considering a state (in our case Kansas) as a single market is reasonable. The model also assumes "a perfectly competitive market with no significant transaction costs" which is in equilibrium (Palmquist 2005, 797).

One problem related with the empirical applications of hedonic models is to choose the functional form. Decision on a functional form has little theoretical justifications. In the literature, previous researchers used different functional forms such as, linear, semi-log, double-log, and Box-Cox. Cropper, Deck, and McConnell (1988) found that the linear and the Box-Cox perform the best in the presence of misspecification. Palmquist (2005), also suggests that the quadratic Box-Cox functional form performs poorly in case of omitted or incorrectly measured variables and recommends a linear Box-Cox functional form.

Selection of explanatory variables in hedonic models is also not guided by theory and therefore very subjective. A common strategy adopted by researchers in selecting variables is to include all available characteristics in the model and use a top-down selection approach (Schöni 2014).

Using machine learning techniques allows us to include a large number of potential predictors in the hedonic model. Some of these predictors, such as the average land prices in a county, are not part of the hedonic literature. We include these predictors because our aim is to predict agricultural land prices, not to make causal inferences.

3.3 Methodology

In data analysis, our aims are to summarize data, estimate models from data, test various hypothesis, and make predictions. Machine learning, a field in computer science, is mostly concerned with prediction. Machine learning algorithms use data to predict some variables as a function of other variables, which are usually called "features."

We can categorize machine learning methods into 2 broad categories; Unsupervised Learning and Supervised Learning. In unsupervised learning methods, we observe the features but we have no response variable. In this case, we can use methods such as clustering or neural networks. In supervised learning methods, however, we both have features and response data. The model in this case refers to a mathematical formulation in which the response y_i is an unknown function of x_i variables and an error term. The aim is to estimate this unknown function in order to make predictions and inference. This functional form can be estimated by parametric or non-parametric methods. The simplest and widely known example for parametric methods is the linear regression. In linear regression we assume that the function to be estimated is linear in x_i variables. In non-parametric methods, on the other hand, we do not make any assumptions about the functional form, but try to fit the function to data as close as possible without being too rough or wiggly. Well known examples of non-parametric methods are splines and generalized additive models.

Most relevant algorithms for economists fall under the supervised learning category. Even though there are many different algorithms in supervised learning category, we can focus on 2 broad categories, namely

1) Linear Model Selection and Regularization, and
2) Tree-Based Methods.

3.3.1 Ordinary Least Squares

In ordinary least squares (OLS) estimation, we minimize the residual sum of squares (RSS) by estimating the regression parameters. RSS can be written as,

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{3.1}$$

where n denotes the number of observations and p denotes the number of independent variables. βs that minimize the RSS, $\widehat{\beta}$ are then used to make predictions with the following formula.

$$\widehat{y_i} = \widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta}_j x_{ij} \tag{3.2}$$

In our simple OLS specification we use priority date, authorized irrigation per acre, logarithm of slope, logarithm of average soil organic carbon, average national commodity crop productivity index, average root zone available water storage, percent of dryland, percent of irrigated land, precipitation, predevelopment depth to water, and predevelopment saturated thickness as control variables.

3.3.2 Penalized Regression

Penalized regression methods or shrinkage methods are very similar to linear regression methods. In shrinkage methods, we regularize (constrain) or shrink the coefficient estimates. Two well known examples of shrinkage methods are ridge regression and least absolute shrinkage and selection operator (LASSO).

Ridge regression is very similar to OLS but with a slight modification. Instead of minimizing RSS, we minimize RSS plus some penalty term. This can be written as,

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.3}$$

where $\lambda \geq 0$ is called the tuning parameter. The second term in Equation 3.3 is called the shrinkage penalty and it shrinks the parameter estimates towards zero. Note that when the tuning parameter is zero, we have OLS regression. On the other hand, when the tuning parameter approaches infinity, coefficient estimates approach zero. It should also be noted that, the parameter estimates from ridge regression depends on the choice of λ which is crucial. Cross-validation is one way to choose the λ parameter.[1] For different λ values, we can calculate the cross-validation error and choose the one with the smallest cross-validation

[1] "[I]t is important to note that this choice [λ that cross-validation chooses] may not immediately equate to good performance when prediction is not the end goal" (Belloni, Chernozhukov, and Hansen 2014b, 33).

error. Another widely used penalized regression method is LASSO which was proposed by Tibshirani (1996). LASSO is an improvement on ridge regression. In ridge regression, the penalty term will shrink coefficients to zero but not set them equal to zero unless the tuning parameter $\lambda = \infty$. In a setting with lots of independent variables, ridge estimation will create some challenges when interpreting the estimation results. However in LASSO, some parameters are shrunk to zero. The formula for LASSO is similar to ridge regression with a slight modification.

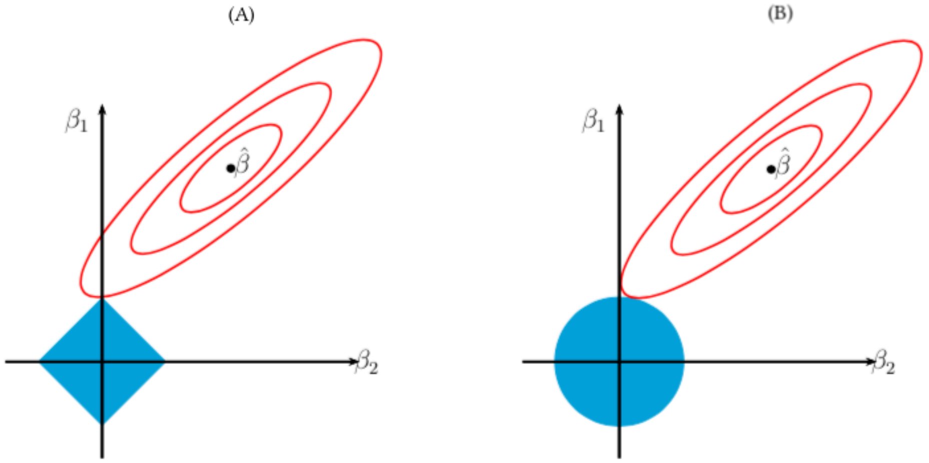

Figure 3.1: Contours of the error and constraint functions for the LASSO (A) and ridge regression (B) (James et al. 2013, 320)

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.4}$$

Since LASSO shrinks some coefficients to zero, it also performs variable selection. Choice of which method (ridge or LASSO) to use depends on the setting of the problem and cross-validation can be used to determine which approach performs best.

Figure 3.1 illustrates the difference between ridge and LASSO regressions. Points marked as $\widehat{\beta}$ show the OLS estimation results where RSS is minimized and red ellipses represent the RSS contours, in other words combinations of $\widehat{\beta}$s that give the same RSS. In penalized regressions we try to find the minimum RSS under some constraints. Blue shaded regions represent the constraints for ridge regression and LASSO for $p = 2$ and $\widehat{\beta}$s that fall in these regions satisfy the constraints.[2] It should be clear from the graphs that ridge and LASSO estimations are given by the tangency of the ellipses and the blue shaded regions. Since in LASSO we have a diamond shaped constraint region, usually the tangency occurs at the corners. This means that some coefficients are shrunk to zero which illustrates the variable selection properties of LASSO (James et al. 2013). Predictions in ridge and LASSO regressions are made in a similar fashion to OLS predictions shown in Equation 3.2.

3.3.3 Regression Trees

In machine learning, decision trees can be used instead of generalized linear models. The goal in decision trees is to grow a tree based on some decision rules and split the data into groups in order to reach good out-of-sample predictions. Building regression trees consists of two steps. In the first step we divide our feature space into J distinct and non-overlapping regions. The goal in this step is to find regions that minimize $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \widehat{y}_{R_j} \right)^2$. In the second step, for every observation in the region R_j we make the same prediction which is simply the mean of the response variable (James et al. 2013).

Figure 3.2 illustrates a simple example of a regression tree. The first node shows that there are 8766 observations in the data. We use "validity1" variable to split the data into two subsets. In node two, we have 2546 observations that all have "validity1" less than -0.46. This corresponds to 29% of the whole data set. We then split data in this node based on "decade.3" variable which leads us to nodes 4 and 5 in which we have 1501 and 1045 observations, respectively. The predicted (mean) price per acre of a parcel in node 4 is 6 whereas in node 5 is 6.6.

[2] Note that we can write ridge regression problem as

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_{ij} \leq s$$

So when $p = 2$, the restriction becomes $\beta_1^2 + \beta_2^2 \leq s$ which defines a region of a circle. Same thing applies to LASSO. In that case we get $|\beta_1| + |\beta_2| \leq s$ which defines a diamond shaped region.

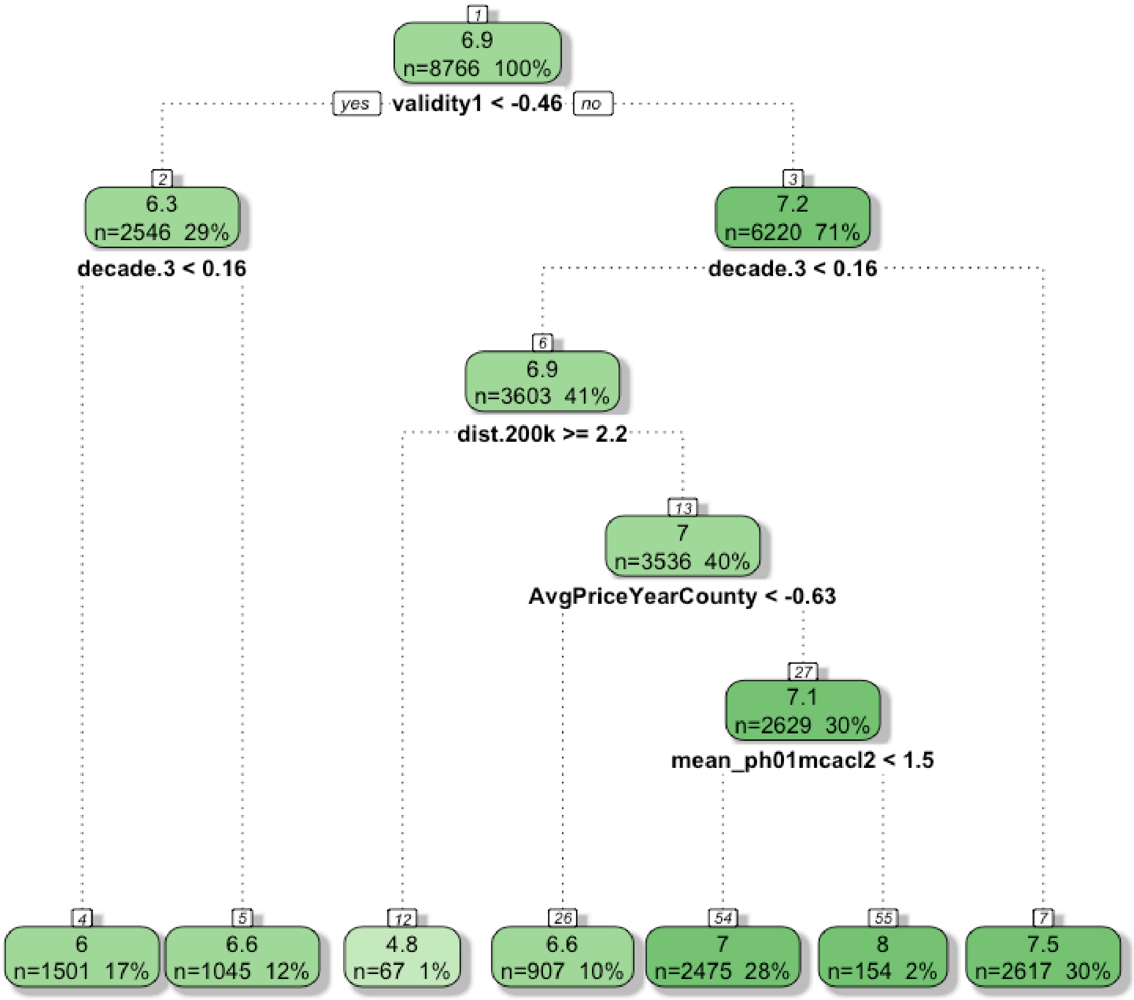

Figure 3.2: Example of a Regression Tree

These models work well if there are nonlinearities in the data, but they have a tendency of over-fitting the data. In regression trees, we face a bias-variance trade-off meaning that if we grow a big tree, we will have a low bias but will end up with high variance. To overcome this over-fitting problem, we use pruning. In pruning, we create a tuning parameter based on cross-validation and choose the best model that gives us the minimum out-of-sample prediction error (James et al. 2013).

3.3.4 Bagging

There are also some other ways to increase the performance in decision trees. First of these methods is called bagging (Breiman 1996). Since we have high variance problem while constructing the trees, we can use bootstrapping to decrease this variance. Bootstrapping is basically a re-sampling method used in statistics. In decision trees framework, we use bootstrapping to choose different samples with replacement from the training data set and build trees which are not pruned. Averaging across different bootstrapped unpruned trees is called bagging (bootstrap aggregation). Since unpruned trees have high variance but low bias, by averaging them we can reduce the variance of the prediction and increase the prediction accuracy. Even though bagging gives us the accuracy we want, there is the disadvantage of interpretability (James et al. 2013).

3.3.5 Random Forests

An improvement of bagged trees is called the random forests. In random forests, we basically construct uncorrelated trees. Similar to what we do in bagging, we build bootstrapped trees, but whenever a split is considered, we choose a random sample of $m \approx \sqrt{p}$ predictors from a full set of p predictors[3]. Even though choosing a subsample of predictors sounds irrational, as James et al. (2013) mentions, it has a clever rationale. Suppose that there exists a very strong predictor in the data. Then in the collection of bagged trees, we will have this strong predictor in most or all of the trees in the very top split. This will lead to a situation where all trees look similar which in turn will lead to highly correlated trees. In this case, bagging will not lead to lower variances. Random forests overcome this problem by forcing each split to consider only a subset of predictors. "Therefore, on average $(p{-}m)/p$ of the splits will not even consider the strong predictor, and so other predictors will have more of a chance" of being included (James et al. 2013, 320).

[3] Note that if we build the random forest by choosing $m = p$ then we will have bagging.

3.3.6 Boosting

Boosting (Freund and Schapire 1997) is very similar to bagging but trees are grown sequentially. Each tree that is grown uses information from the previously grown trees. In boosting we don't use bootstrapping, but fit the tree on a modified version of the original data set. In this ensemble technique, errors are corrected by sequentially adding new models to the existing models until no more improvements can be made. Boosting algorithms include some parameters which slow down the learning process. This slow learning process gives us better predictions (James et al. 2013). Another method related to boosting is the gradient boosting which was developed in several papers such as Breiman (1996); Friedman, Hastie, and Tibshirani (2000); Friedman (2001). In this method, new models are created to predict the errors of the previous models. The gradient descent algorithm is used to optimize an arbitrary differentiable loss function while adding new models, it can be thought as a combination of gradient descent and boosting.

Extreme Gradient Boosting (xgboost), developed by Chen and Guestrin (2016), is an efficient, flexible and portable variant of the gradient boosting model of Friedman, Hastie, and Tibshirani (2000) and Friedman (2001). Xgboost has been a winning tool for several Machine learning competitions (Adam-Bourdarios et al. 2015; Chen and Guestrin 2016). Xgboost relies on the same principles with gradient boosting but compared to gradient boosting, it is more efficient and faster since it uses sparsity aware algorithms and better processor utilization.

3.4 Results

3.4.1 Preprocessing

Before estimating the machine learning models, we need to preprocess the data and create new features/variables. The accuracy of predictions depend on zero and near zero variance features in the data set. In order to increase the accuracy, we first analyze these variables. Results show that there are no critical variables that have zero or near zero variances. Some soil characteristics and travel time variables have zero or near zero variances so we drop these variables. Since the algorithms require no missing values in the data set, we impute missing values with k-nearest neighbor methods.

When we analyze the linear dependence among variables we see that interaction terms and some soil variables are linearly dependent. We drop linearly dependent variables from the data set. Centering and scaling predictors are highly recommended (Kuhn and Johnson 2013), so we use Yeo-Johnson transformation, which is similar to Box-Cox transformation but allows for negative values, to transform the variables and then center and scale them. We also drop percent of native grassland variable from the dataset to make grassland our baseline. After preprocessing, we end up with 251 predictors. Before applying feature selection, we divide the data into training and test sets. We randomly chose 80% of the data as the training and the remaining 20% as the test set.

3.4.2 Feature Selection

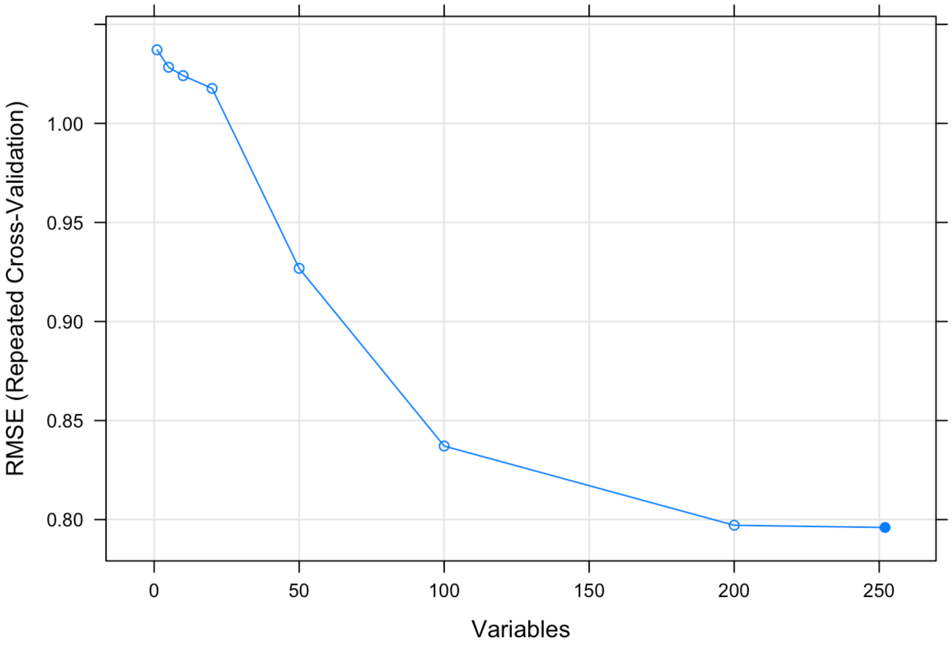

Figure 3.3: Optimal Number of Features Selected

Feature or variable selection is one of the most important steps in machine learning estimation. Feature selection allows us to decrease the time spent for training models since training time increases exponentially with number of features. It also eliminates the risks of overfitting. We first use a recursive feature selection with five fold repeated cross validation to find the best subset of variables. Figure 3.3 shows the root mean square of regressions with different number of variables. Results show that the optimal number of variables are 251. This suggests that we do not need any feature selection at all and therefore we use all the variables in the model estimations.

3.4.3 Estimation

We use different machine learning algorithms for estimation and prediction. In ridge and LASSO estimations, selection of the tuning parameters are crucial. We estimate the model twice to find the optimal parameter values. Using ten fold repeated cross validation, we first find the optimal regularization parameters for ridge regression.

Figure 3.4 shows the optimal regularization parameter for the first ridge estimation. Using the parameter values from this estimation, we re-estimate the model using ten fold repeated cross validation. Figure 3.5 shows the optimal values for the regularization parameters for ridge regression in the second estimation. At this point we decide that 0.04033 is the optimal λ value for the ridge regression. We carry a similar estimation procedure to find the optimal regularization parameter for the LASSO regression. Figure 3.6 shows the optimal value for the regularization parameters for LASSO estimation. We decide that 0.0000000009 is the optimal λ value for the LASSO regression. Using this parameter value, we estimate the LASSO regression. Since LASSO has the property of variable selection, it drops 16 variables out of 251 from the model which corresponds to nearly 6% of all predictors. In order to make predictive accuracy comparisons we also estimate an OLS model with all variables, and a simple OLS model of selected variables with and without year fixed effects. In regression trees estimation, similar to ridge regression and LASSO, we first find the optimal complexity parameter. Complexity parameter can be defined as the cost of adding another variable to

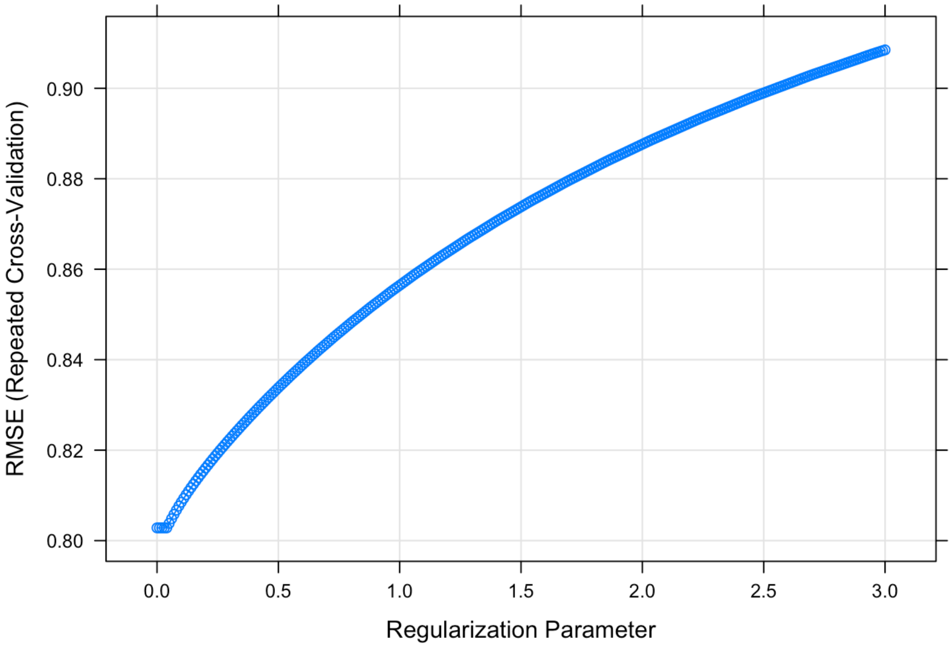

Figure 3.4: Optimal Regularization Parameter for Ridge Regression - First Step

the model. Again using ten fold repeated cross validation we find the optimal value for the complexity parameter. Figure 3.7 shows that the optimal value for the complexity parameter is 0. After deciding on the optimal value for the complexity parameter, we re-estimate the model for pruning. Figure 3.8 shows the set of possible cost-complexity prunings of a tree. Therneau and Atkinson (2018) suggest that the optimal complexity parameter for pruning to be the leftmost value for which the mean lies below the horizontal line shown in Figure 3.8. Hastie, Friedman, and Tibshirani (2001), on the other hand, suggest the one-standard error rule. According to this rule “we choose the most parsimonious model whose error is no more than one standard error above the error of the best model” (Hastie, Friedman, and Tibshirani 2001, 244). So we prune the estimated tree with both approaches. In random forest estimation, the parameter to be optimized is the number of randomly selected predictors. Using cross validation, we find that the optimal number of predictors is 251. Using all the predictors in the data, we estimate the random forest model. Similarly for stochastic gradient boosting model, we use cross validation to fine tune the model parameters. We find

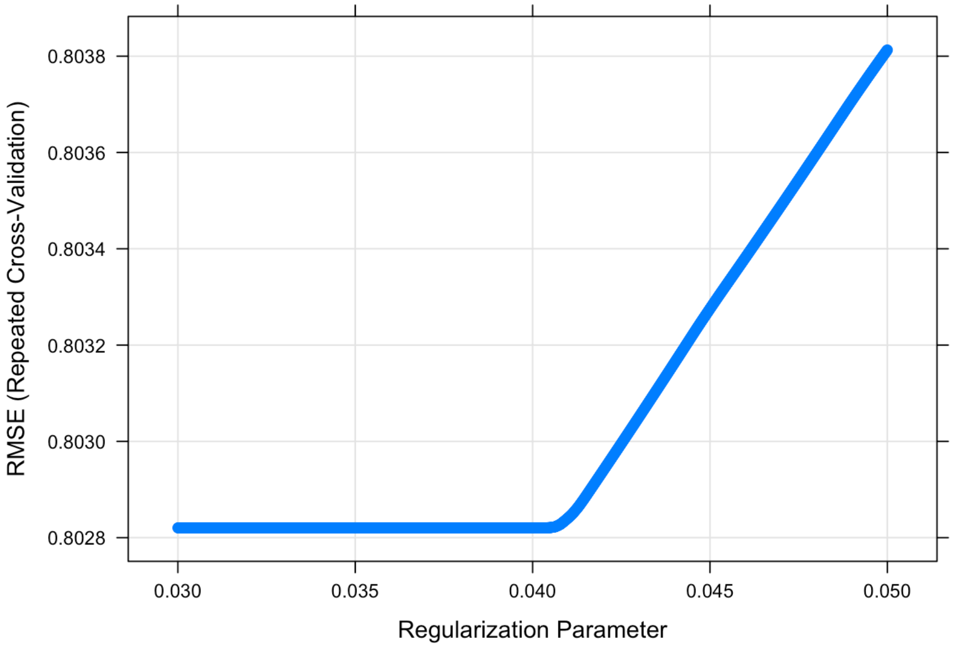

Figure 3.5: Optimal Regularization Parameter for Ridge Regression - Second Step

that the optimal number of boosting iterations as 500, maximum tree depth as 6, shrinkage (the steps taken in the gradient descent) as 0.1, and minimum terminal node size as 1. Using these optimal values we estimate the model. We train the extreme gradient boosting model by setting three different sets of parameters[4]. All these parameters are used in order to make good choices about model complexity and predictive power. We use five fold cross validation five times to find the optimal tuning parameters. We find that the optimal number of rounds for boosting is 2000, optimal learning rate is 0.1, minimum loss reduction is 0, maximum tree depth is 6, and minimum number of instances needed in each node is 3. Using these optimal parameter values, we re-estimate the model.

[4] Please see https://xgboost.readthedocs.io/en/latest/parameter.html for a complete list of parameters to be tuned.

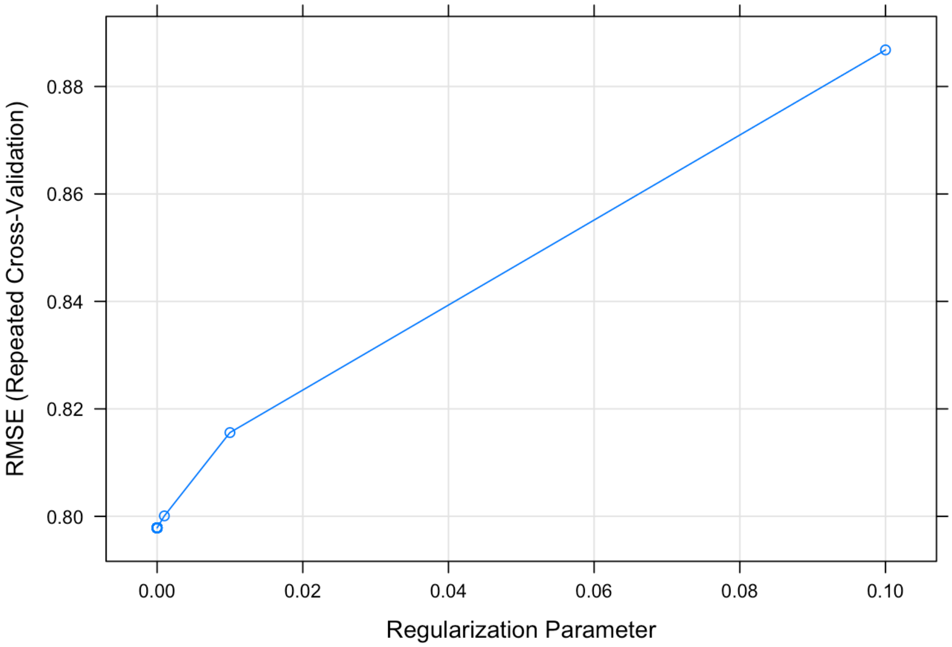

Figure 3.6: Optimal Regularization Parameter for LASSO Regression

3.4.4 Model Comparisons

Predictive Accuracy

In model evaluation studies, both the root mean square error (RMSE) and the mean absolute error (MAE) are regularly used. RMSE shows the absolute fit of the model to the data and is a good measure of predictive performance if the main purpose of the model is prediction (Yoo, Im, and Wagner 2012). However, Willmott and Matsuura (2005) suggest that RMSE may not be a good indicator of average model performance and might be misleading. They suggest that MAE should be used for model performance comparisons. On the other hand, Chai and Draxler (2014) suggest that a combination of similar metrics should be used for model performance evaluation. R^2 values of the models can also be compared but while doing that we should be cautious since R^2 is a measure of correlation and not accuracy (Kuhn and Johnson 2013). Another metric we can use is the median absolute error (MedAE) that is calculated by taking the median of all absolute prediction errors. One specific property of MedAE is, being robust to outliers (Bonnin 2017). Table 3.1 shows the predictive accuracy

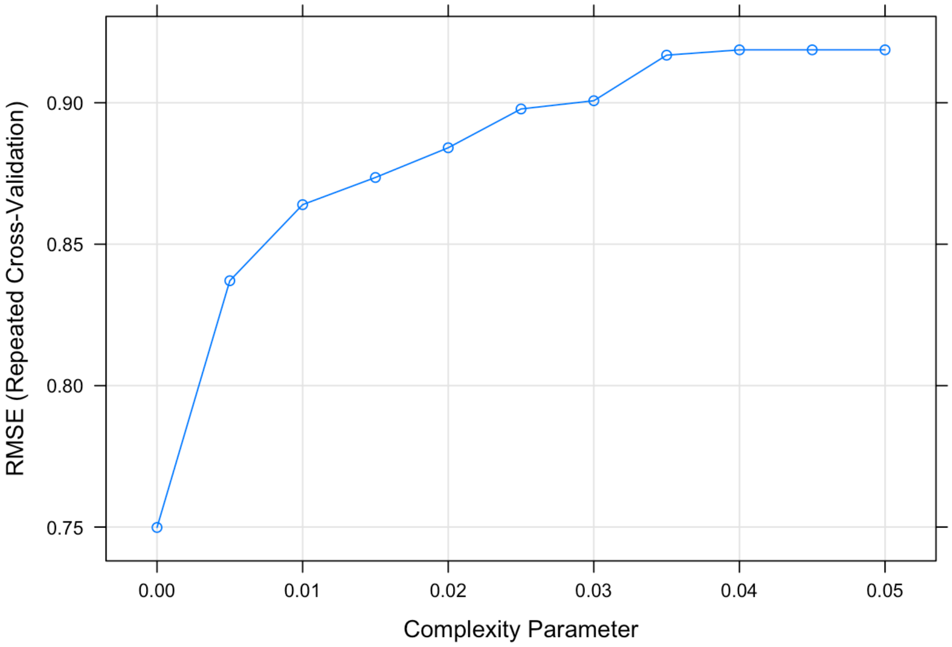

Figure 3.7: Optimal Complexity Parameter for Regression Trees

measures for different predictions. We find that MAE in xgboost is 0.2643. This implies that, on average, the difference between our model's predictions and the true log price was 0.2643. The mean log price in the training data is 6.9186. If we predicted the value 6.9186

Table 3.1: Predictive Accuracy Measures for Different Models

Model	R.Squared	RMSE	MAE	MedAE
Extreme Gradient Boosting	0.8203	0.4407	0.2643	0.1312
Random Forest	0.7463	0.5251	0.3456	0.2103
Regression Trees - TA	0.5251	0.7384	0.5139	0.3439
Regression Trees - HFT	0.5075	0.7396	0.5375	0.3877
Gradient Boosting	0.4568	0.7765	0.5723	0.4149
LASSO	0.4135	0.7940	0.5990	0.4494
OLS	0.4189	0.7905	0.5996	0.4513
Ridge Regression	0.4211	0.7893	0.6022	0.4641
Simple OLS with Year Fixed Effects	0.1622	0.9486	0.7285	0.5859
Simple OLS	0.0412	1.0150	0.7988	0.6528

Note:
TA: Therneau and Atkinson (2018), HFT: Hastie, Friedman, and Tibshirani (2001)

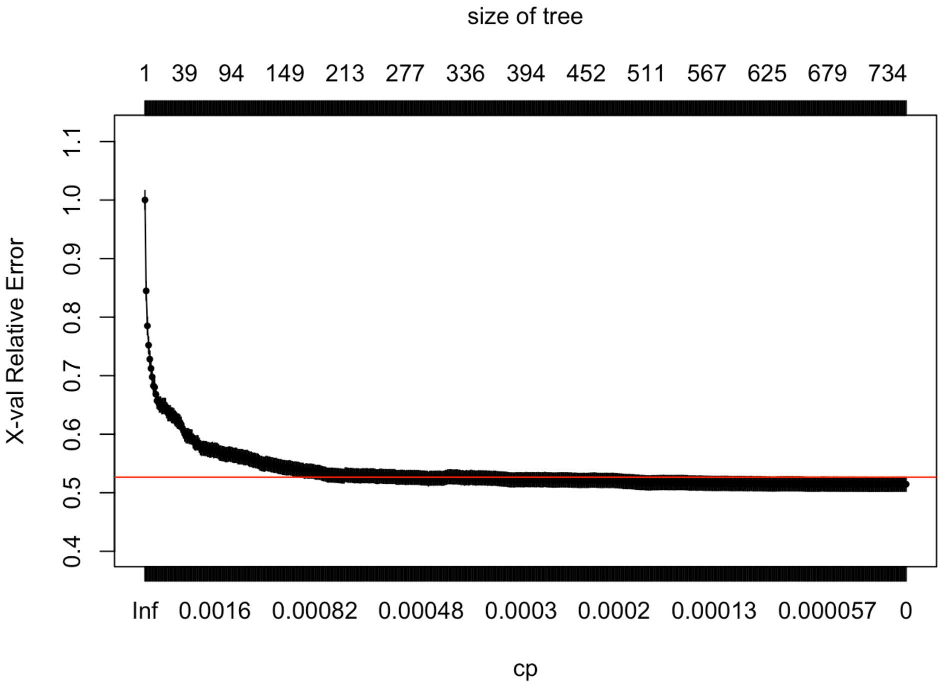

Figure 3.8: Optimal Complexity Parameter for Regression Tree Pruning

for every parcel sample, we would have a mean absolute error of only about 0.8224 (Lantz 2015, 214). Comparing MAE from xgboost and other models shows that we have a huge improvement in prediction. The estimated MedAE for xgboost is 0.13, meaning half of the predictions are 13% of the real price and half of them are off by more than 13%.

Visual Comparison

Figure 3.9, Figure 3.10, and Figure 3.11 show the predicted versus the observed values for simple OLS model, LASSO model, and extreme gradient boosting model, respectively. Using these figures we can visually assess the predictive power of the models. As seen in Figure 3.9, the points are scattered around too much. We want the points to be closer to the blue line. This is an indication that the predictive power of the simple OLS model is not good. Figure 3.10, shows the LASSO model. Compared to simple OLS model the predictions from LASSO are better. Finally when we check, Figure 3.11 we see that extreme gradient boosting model is the best among the models estimated. Even though we have some outliers most of the

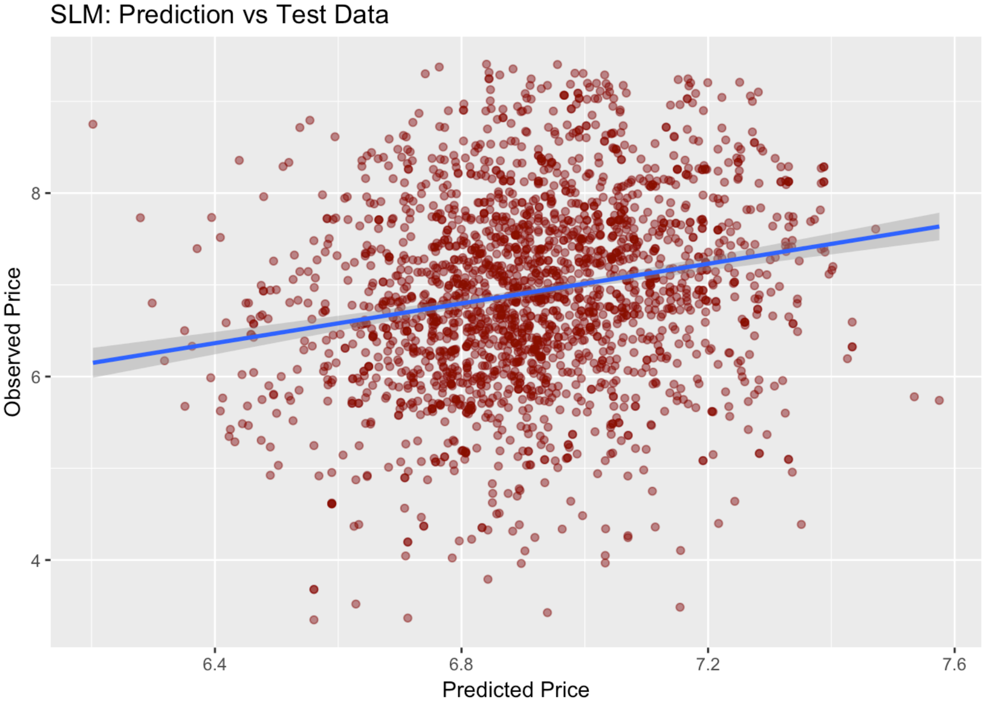

Figure 3.9: Predicted vs. Observed Values (Simple Linear Model)

predictions are very close to observed values.

Feature Importance

Feature importance provides information about the contribution of each variable to prediction. Feature importance only measures the importance of variables to prediction, therefore they should not be interpreted as impacts on probabilities or as regression coefficients. Highly ranked features contribute more to prediction, but this does not necessarily mean that low ranked features are not important. Having a low rank in feature importance does not imply that the feature is a bad predictor (Ifft, Kuhns, and Patrick 2018). Figure 3.12, Figure 3.13, and Figure 3.14 show the variable importance results from ridge regression, LASSO, and xgboost. As seen from Figure 3.12 the most important variables are `validity1`, `AvgPriceYearCounty`, and `decade3`. `validity1` is the variable showing if the parcel sale is defined as valid. `AvgPriceYearCounty` shows the average price of all sales in a county for the same year that the parcel was sold. `decade3` denotes if the sale occurred between 2005

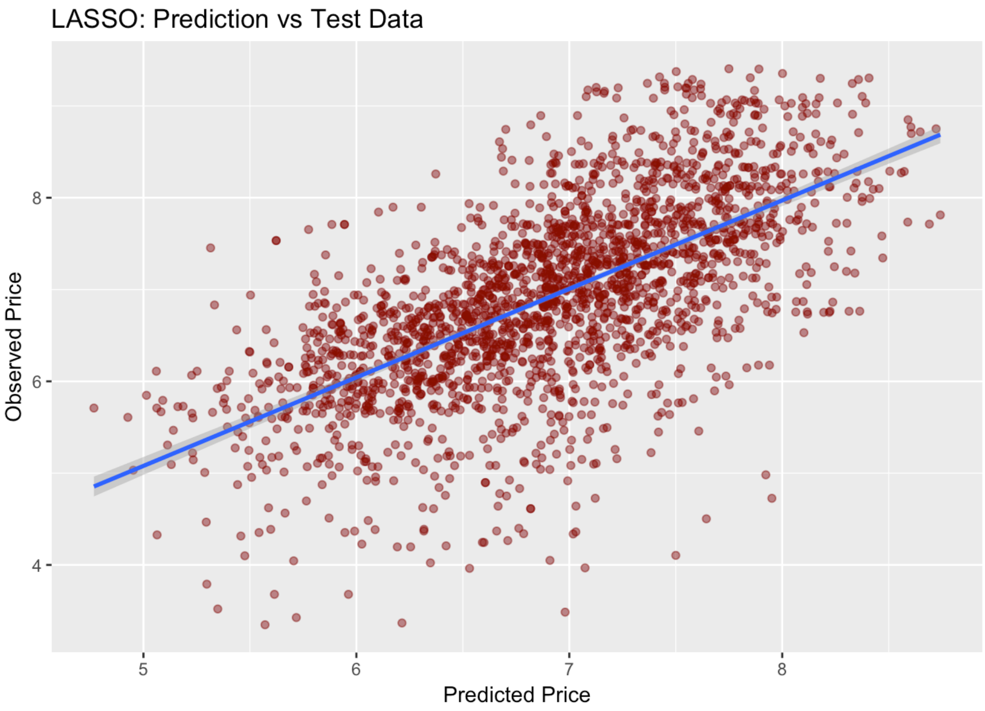

Figure 3.10: Predicted vs. Observed Values (LASSO Regression)

and 2015. Ridge regression results also show 6 different soil properties as most important features.

Figure 3.13 shows that the most important variables in LASSO estimation are `decade3`, `pre_dtw`, and `mean_aws_100_150`. `pre_dtw` shows the predevelopment depth to water whereas `mean_aws_100_150` shows the available water storage estimate in mm. in 100 - 150 cm. depth. LASSO results indicate that 3 different time variables are important in prediction whereas ridge regression shows that 11 time variables are important. LASSO results also show that growing degree days and 5 other weather variables are important features. We also see that size of the parcel is also important in LASSO estimation. Figure 3.14 shows that there are 4 different clusters of predictors that have similar importance values. `validity1` is again the most important feature. Similar to ridge and LASSO results, we see that `decade3` and `AvgPriceYearCounty` are in the top predictors. Contrary to ridge and LASSO, xgboost results show that priority days (`priority_days`) and authorized irrigation per acre (`aipa`) are also important features.

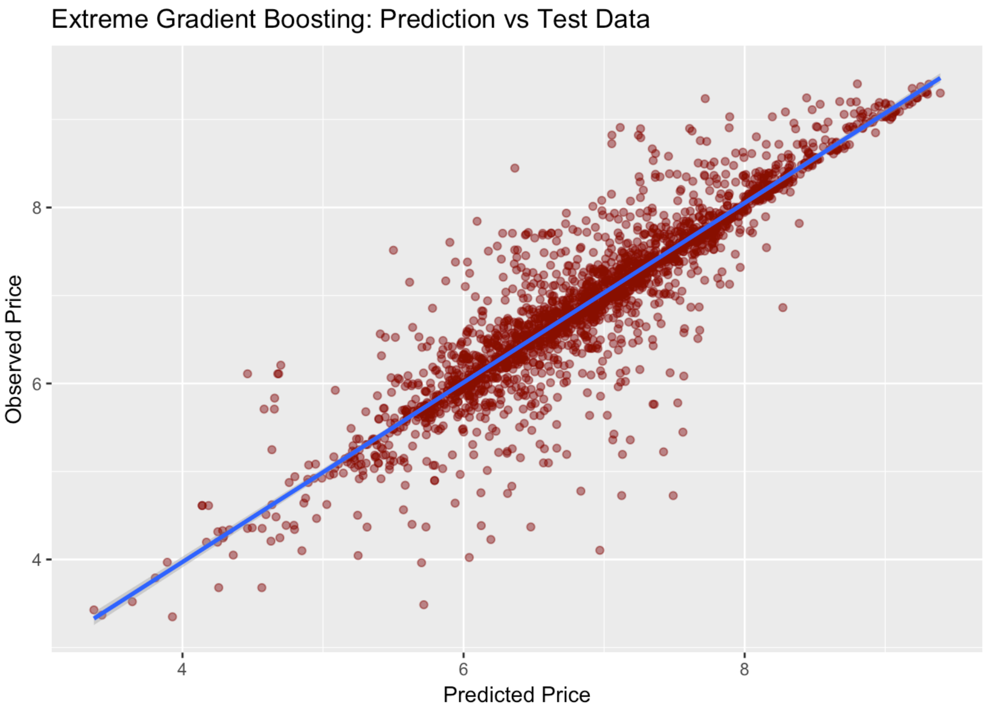

Figure 3.11: Predicted vs. Observed Values (Extreme Gradient Boosting Model)

3.4.5 Prediction

In this section, we combine PVD data with the points of diversion data to predict the land values at points of diversion locations. First we split the data into training and test sets. We choose PVD data to be the training set and the property valuation division data to be the test set. We randomly split our training data again to create training and validation samples. We train the xgboost model using the training data, which is 80% of the property valuation division data. Using the validation data, which is the 20% of the property valuation division data, we make in-sample predictions.

Table 3.2 shows the in-sample predictive accuracy measures for the validation data. The root mean squared error of the model for validation data is around 0.27. We also find that mean absolute error is around 0.02, meaning half of the predictions are 2 percent of the real price and half of them are off by more than 2 percent. Figure 3.15 shows the variable importance graph for the new estimation. As seen from the graph, validity of the sale, sale being in years between 2005 and 2015, and distance to 200K cities are the most important

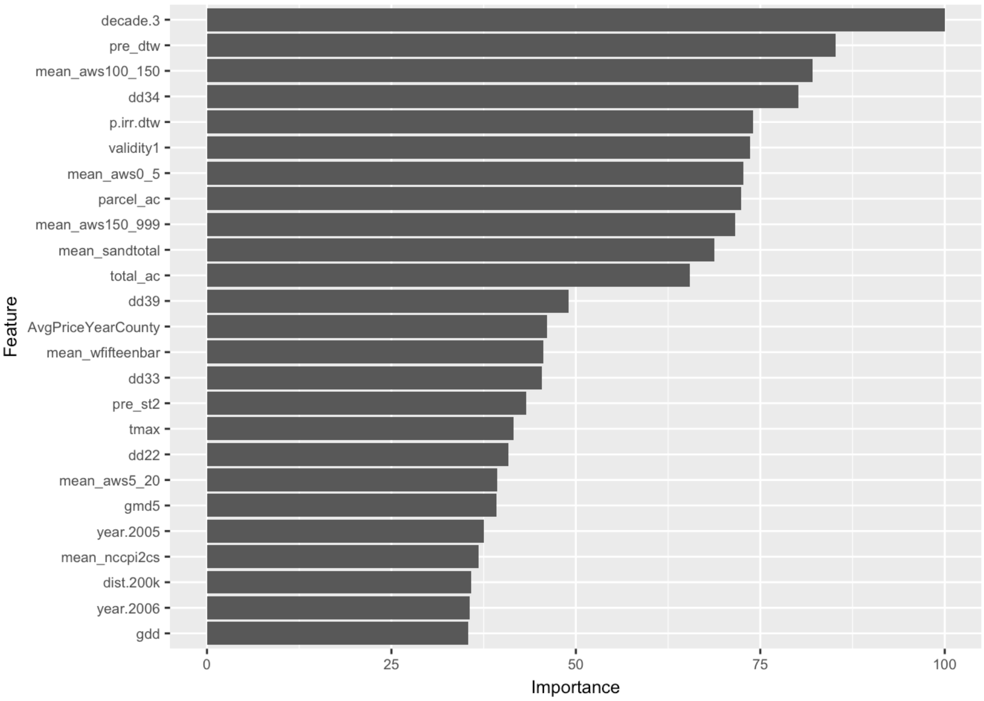

Figure 3.12: Variable Importance (Ridge Regression)

Table 3.2: In-sample Predictive Accuracy Measures for Points of Diversion Data

Model	R.Squared	RMSE	MAE	MedAE
Extreme Gradient Boosting	0.9351	0.2721	0.1045	0.0189

predictors. We also see that multiple soil and water rights characteristics are also important factors that improve the power of the prediction. After training the model, we use it to perform out of sample forecasts using the points of diversion data. Since we do not have any price information for this data we make some assumptions to predict the prices. We assume that all parcels in point of diversion data are sold in 2014. We also assume that all the sales are valid sales. We map the results of predictions in Figure 3.16 and Figure 3.17. Figure 3.16 shows the plot of predicted nominal prices for the parcels in points of diversion data. Predicted nominal prices vary between $804 and $4,642 per acre. Figure 3.17 shows the predicted nominal price averages in each GMD. As seen from the plot the most expensive lands are located in the south-eastern Kansas. We find that GMD 2 has the lowest mean

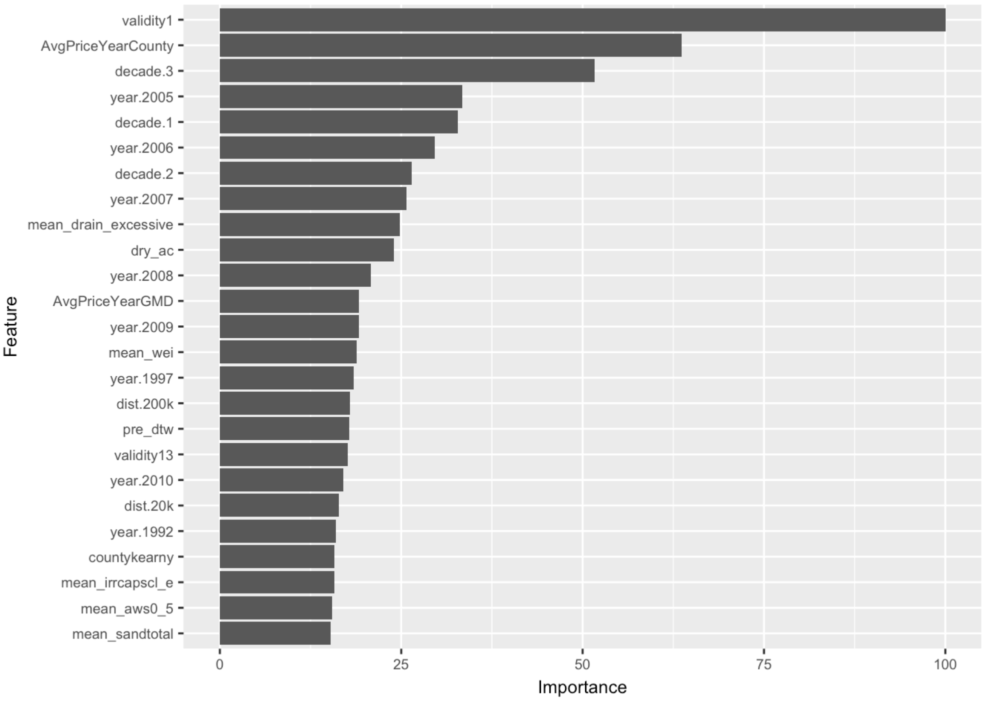

Figure 3.13: Variable Importance (LASSO Regression)

price per acre ($2,028) and GMD 5 has the highest mean price per acre ($2,350).

3.5 Conclusion

In this study we used machine learning techniques to predict the agricultural land values in Kansas and compare the results with the standard econometric methods. We introduced many parcel and geography related variables into the hedonic pricing model. Using feature selection algorithms we found the optimal subset of variables and then estimated different machine learning models.

We found that compared to standard econometric methods, machine learning methods give better predictions and the extreme gradient boosting algorithm is the best among all algorithms. The results show that half of the predictions from the simple OLS model are 65% of the real price whereas half of the predictions from xgboost model are 13% of the real price.

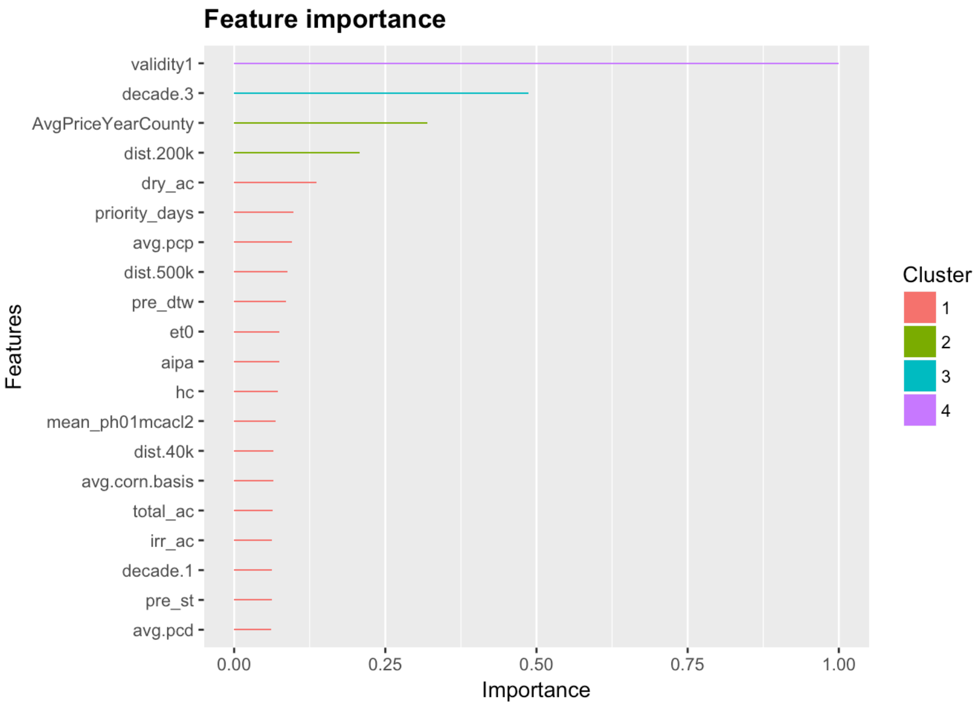

Figure 3.14: Variable Importance (Extreme Gradient Boosting Model)

Out-of-sample predictions from xgboost model using point of diversion data show that the predicted nominal agricultural land price in Kansas varies between $804 and $4,642 per acre. We also see variation of predictions in different groundwater management districts. Mean nominal price per acre varies between $1,178 and $4,643 in different groundwater management districts.

Our results show that machine learning algorithms can increase the predictive power of the models. Our results are valuable for researchers, farmers, investors, and policymakers. Researchers and policymakers can take advantage of easy access to big data and apply these techniques to make better predictions. Farmers and investors can use price predictions during their decision process whether to buy land. Governments can also benefit from the results. For example, for a water program that aims to decrease irrigation of farmers, governments can use price prediction to find the optimal parcels to buy.

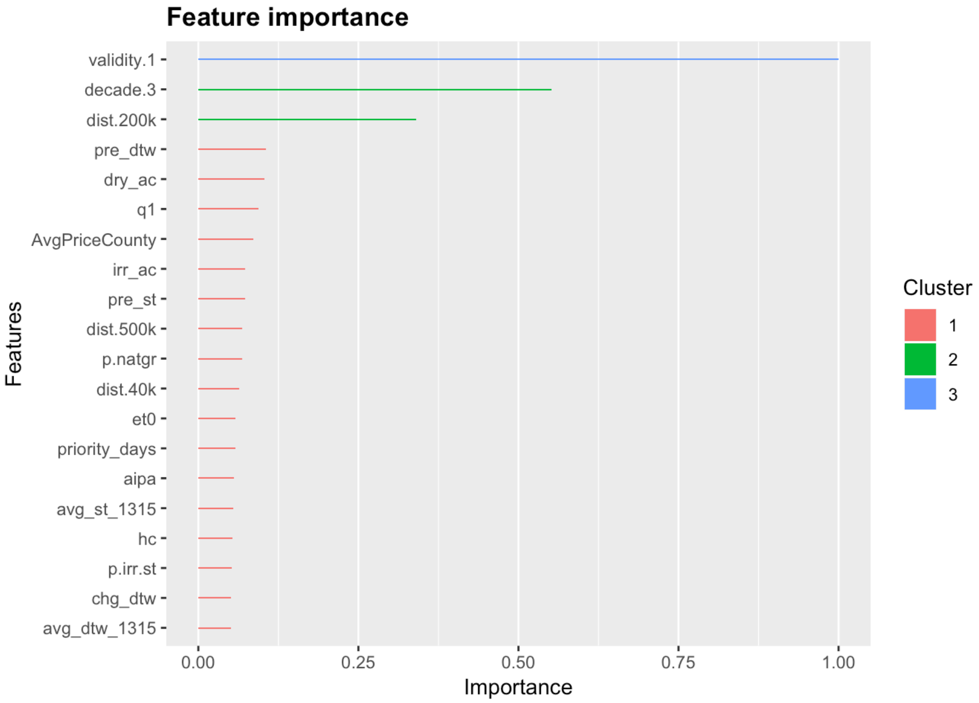

Figure 3.15: Variable Importance (Extreme Gradient Boosting Model) for Points of Diversion

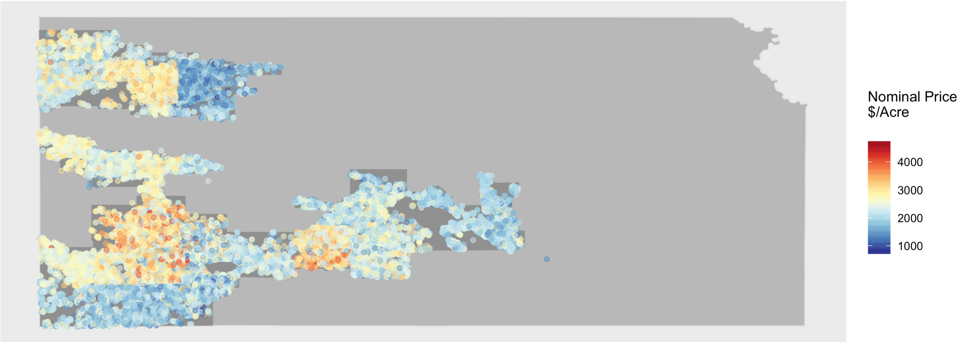

Figure 3.16: Nominal Predicted Prices for Points of Diversion

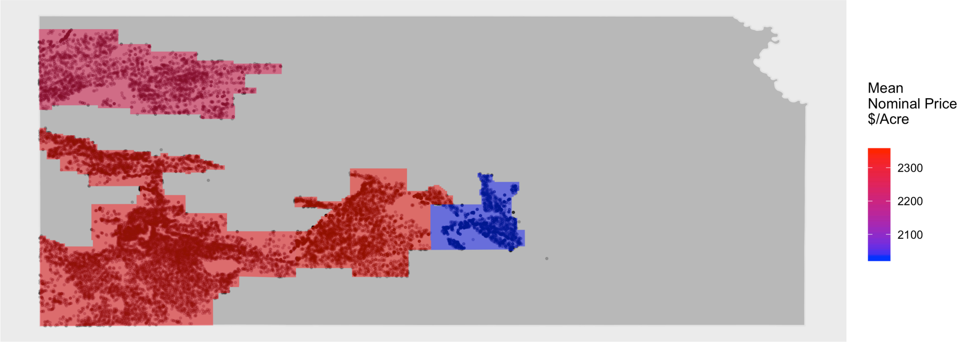

Figure 3.17: Nominal Predicted Prices for Points of Diversion - Groundwater Management District Averages

Chapter 4

The Effects of Water Rights on Agricultural Land Values in Kansas

4.1 Introduction

Land values are significant both to farmers, landowners, and policymakers; therefore understanding the determinants of land values and predicting them is important (Nickerson and Zhang 2014; Burns et al. 2018). As stated in Jenkins et al. (2007), buyers of agricultural land face different prices in different geographical regions even though the characteristics of the lands are similar. These differences can be attributed to differences in soil quality, annual rainfall, urban influences, and other factors (Burns et al. 2018). Therefore, numerous studies in the literature investigated the determinants of agricultural land values either using farm income variables or farm characteristics (Burt 1986; Featherstone and Baker 1987; Just and Miranowski 1993; Moss 1997; Shi, Phipps, and Colyer 1997; Goodwin, Mishra, and Ortalo-Magné 2003; Tsoodle, Golden, and Featherstone 2006).

Crop yield is one of the most important factors that affects the agricultural land values. Crop yield is highly dependent on irrigation, irrigation and water rights associated with the land are expected to be capitalized in agricultural land values. The price difference in agricultural land values, therefore, can partially be explained by the water rights associated to the land.

Several studies in the literature also analyzed the effect of water rights (i.e., the right to irrigate a parcel) on land prices with hedonic pricing methods (Crouter 1987; Torell, Libbin, and Miller 1990; Faux and Perry 1999; Jenkins et al. 2007; Petrie and Taylor 2007; Buck, Auffhammer, and Sunding 2014; Hornbeck and Keskin 2014; Mukherjee and Schwabe 2014; Brent 2016; Ifft, Bigelow, and Savage 2018). In these studies authors used different variables to capture the effect of water rights. For example, Torell, Libbin, and Miller (1990) used depth of water available for pumping, Butsic and Netusil (2007) used a dummy variable if the land has a water right, Buck, Auffhammer, and Sunding (2014) used surface water delivery right per acre, Hornbeck and Keskin (2014) used access to Ogallala groundwater, Mukherjee and Schwabe (2014) used a water portfolio (having access to multiple sources of water), Brent (2016) used water volatility, and Ifft, Bigelow, and Savage (2018) used irrigation restrictions. To the best of our knowledge, none of these studies have estimated the value of specific attributes of the water right.

Hedonic pricing methods have some problems that are usually overlooked. As stated in Schöni (2014), selection of explanatory variables in hedonic models is not guided by theory and therefore very subjective. Majority of the hedonic models in the literature include a very limited set of explanatory variables and therefore suffer from omitted variable bias. For example, Jenkins et al. (2007) estimated the link between price and water rights by using Ordinary Least Squares. They only regressed a dummy variable of water rights on land prices and found that water rights increases the land price. Even though the estimated coefficient in the study is statistically significant, due to lack of important explanatory variables, such as soil properties, weather, etc., their analysis suffers from omitted variable bias.

Another problem with the previous studies is the small sample size of the data used. For example, Crouter (1987) found no significant effect of water rights on land prices but the sample size was only 53. Similarly, Faux and Perry (1999) and Butsic and Netusil (2007) used 225 and 113 observations, respectively.

The Property Valuation Division (PVD) of the Kansas Department of Revenue (KDOR) does not provide water right attributes information in their data; however, they provide the geolocation of the parcels. The water data from Water Rights Information System (WRIS) (2015) include these attributes along with geographical information. Using Geographical Information System (GIS) techniques, parcel sale data and water rights data are merged and water rights attributes for each parcel were determined. In Kansas, the right to use water is based on “first in time - first in right” principle. The objective of this paper is to determine whether water right seniority is capitalized into land values.

A possible concern during the estimation of hedonic models would be the omitted variable bias. In our case, for example, one may include water rights attributes in the estimation but not fully control for the soil characteristics. Since people developed irrigation on good soils earlier, we expect a positive correlation between soil characteristics and water right seniority. Omitting soil characteristics during estimation will cause our coefficients to be biased. Because of this concern, we need to control for other characteristics of the land, such as soil properties, hydrological properties, etc. On the other hand, adding too many variables in the model will cause overfitting problems.

In the literature, researchers usually use fixed effects, instrumental variables, or quasi-randomness to overcome the omitted variable bias, but these methods require strong assumptions (Ho 2017). To overcome the omitted variable bias and overfitting tradeoff, we use recent machine learning methods to estimate the causal effect of water rights on irrigated land prices in Kansas. The paper takes advantage of the variable selection properties of least absolute shrinkage and selection operator (LASSO) to reduce omitted variable bias (Belloni, Chernozhukov, and Hansen 2014b). By introducing many parcel and geography related variables and letting the LASSO estimator choose the explanatory variables, we overcome the problem of omitted variable bias and overfitting. Our data consist of 7,005 observations of irrigated agricultural land sales in Kansas between 1985 and 2015.

The effects of groundwater irrigation on land values are mixed in the literature. Hartman and Taylor (1989a), Hartman and Taylor (1989b), and Sunderland, Libbin, and Torell (1987) find that groundwater irrigation has no significant effect on land prices (Islam 2010; Brozovic and Islam 2010; Ifft, Bigelow, and Savage 2018). Whereas Torell, Libbin, and Miller (1990), Brozovic and Islam (2010), Hornbeck and Keskin (2014), and Mukherjee and Schwabe (2014) find that access to water increases farmland value.

The results show that water rights have a positive impact on irrigated land prices in Kansas. An additional year of water right seniority causes irrigated land value to increase nearly $17 per acre. The analysis shows that price per acre is a nonlinear function of seniority. Further analysis based on different Groundwater Management Districts also shows that there are geographical differences in the response of price to seniority. We find that the impact of seniority appears to be strongest in GMD 3 and GMD 5.

4.2 Prior Appropriation Water Rights in Kansas

Different property rights doctrines and institutions govern the water rights in the United States. Prior appropriation doctrine in Colorado, Kansas, New Mexico, South Dakota, and Wyoming, gives an individual the rights to use water based on the priority of the date that water use was established. Since the right to use water is based on priority, this doctrine is also known as "first in time - first in right" principle.

The prior appropriation doctrine prohibits junior right holders, individuals who claimed their right later in time, from using water when the available resources are very limited and there is a probability of senior right holders to drop below their allocated right limits.

Until 1945, water rights in Kansas were governed by the absolute ownership doctrine –a doctrine that gives landowners the absolute right to extract water. In June 28, 1945, after multiple conflicts between water users, Kansas Water Appropriation Act has been enacted and Kansas adopted prior appropriation doctrine (Peck 2007; Lawell 2017). After 1945 until 1970s, with the development of agriculture in Kansas, several pumping permits were issued. This caused a rapid decline in groundwater resources and in 1972 Kansas legislature decided to create five groundwater management districts (GMDs) to regulate the water extraction (Peck 2007; Lawell 2017).

Prior appropriation rights were rarely exercised in Kansas. However, in recent years, depletion of the reservoirs and droughts gave rise to new orders and lawsuits. There are two cases that we are aware of: establishment of Intensive Groundwater Use Control Areas

(IGUCA) and Haskell County impairment lawsuit. In 1990, due to declining groundwater levels in Walnut Creek Basin, IGUCA process was initiated and within two years it was established (Golden and Leatherman 2017). With IGUCA, 22,700 acre-feet was decided as the safe (long-term sustainable yield) level for withdrawal and two types of water rights (senior and junior appropriation rights) were defined. Senior rights are the rights with priority dates on or prior to October 1, 1965 whereas junior rights are the rights with priority dates subsequent to October 1, 1965. According to IGUCA, all vested rights are left at their authorized quantities. Appropriation for senior rights were reduced to an amount that is "reasonable" for the area. Junior rights were allocated the remaining portion of the 22,700 acre-feet.

Similarly, the water scarcity in the southwestern Kansas caused farmers, who want to experience their senior rights, to file lawsuits. For example, in a recent rule Haskell County District Court Judge Linda Gilmore decided to shut down a company's two junior wells.[1]

The senior water rights provide security of water use to farmers when the water is scarce. When we take depletion of the reservoirs, droughts, and rules into account, we expect the security that senior rights provide to be capitalized into land values.

4.3 Methodology

4.3.1 Causal Estimation with Machine Learning

The main focus of machine learning algorithms is prediction and they may lead to wrong conclusions if we are interested in casual inference (Belloni, Chernozhukov, and Hansen 2014b; Leeb and Pötscher 2008a, 2008b). It should be clear that with many explanatory variables as much as the number of observations, OLS estimate may give a nearly perfect fit. In this case, even though the in-sample prediction is perfect, estimated models may give poor out-of-sample prediction. One way to model this problem is to use regularization methods. By using regularization we restrict the estimates and fix the over-fitting problem. When

[1] Source: Judge rules in favor of southwest Kansas farm family's senior water rights.

the over-fitting problem is fixed, useful out-of-sample forecasts can be obtained (Belloni, Chernozhukov, and Hansen 2014b). It should be noted that even though a low variance occurs with regularization, estimates tend to be biased towards zero. To overcome this problem, post-LASSO proposed by Belloni et al. (2012) and Belloni and Chernozhukov (2013) is used.

Post-LASSO estimation consists of 2 steps. In the first step LASSO's variable selection properties are used to drop variables from the model. In the second step, OLS is estimated on the remaining variables. It has been shown that Post-LASSO estimators are often better than LASSO estimators in terms of convergence and bias (Belloni et al. 2012; Belloni and Chernozhukov 2013; Belloni, Chernozhukov, and Hansen 2014b).

In causal inference the treatment effect (α) is of interest. In the water rights context the model is as follows,

$$\log\left(\text{Real Price per Acre}\right)_i = \beta_0 + \alpha\text{Priority Date}_i + \sum_{j=1}^{p}\beta_j x_{ij} + \varepsilon_i \tag{4.1}$$

where conditional independence and approximate sparsity are assumed and $p \gg n$.[2] Priority Date$_i$ is the number of years after the enactment of Kansas Water Appropriation Act, and x_{ij}'s are all other control variables included in the model.

Since we are interested in causal relationships, we can use double-selection LASSO proposed in Belloni, Chernozhukov, and Hansen (2014b). One may think to apply LASSO to Equation 4.1 and force α to stay in the model. Then use remaining variables and Priority Date$_i$ to estimate an OLS regression to make causal inferences about the treatment. The first problem related to this approach is the omitted-variable bias. Since LASSO is about prediction and not about the specific parameters, LASSO will drop any controls that are highly correlated to the treatment variable which will lead to an omitted-variable bias. Second problem with this approach is that the model is constructed in a way to predict the outcome given treatment and other independent variables (Belloni, Chernozhukov,

[2] Approximate sparsity imposes a restriction that only s variables among all of x_{ij}, where s is much smaller than n, have associated coefficients β_j that are different from 0 (Belloni, Chernozhukov, and Hansen 2014b, 32).

and Hansen 2014b). This problem can be overcome by transforming Equation 4.1. We can substitute Equation 4.2 into Equation 4.1 to get Equation 4.5.

$$\text{Priority Date}_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \nu_i \tag{4.2}$$

$$\log\left(\text{Real Price per Acre}\right)_i = \beta_0 + \alpha\left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \nu_i\right) + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{4.3}$$

$$= \left(\alpha\theta_0 + \beta_0\right) + \sum_{j=1}^{p} \left(\alpha\theta_j + \beta_j\right) x_{ij} + \left(\alpha\nu_i + \varepsilon_i\right) \tag{4.4}$$

$$= \gamma_0 + \sum_{j=1}^{p} \gamma_j x_{ij} + \epsilon_i \tag{4.5}$$

Equation 4.2 and Equation 4.5 can be estimated by LASSO. A problem may arise at this point. One may think to use one of these equations for variable selection which again will lead to omitted variable bias. Suppose one chooses the Equation 4.5 for variable selection. LASSO may ignore some variables that have strong predictive power for the treatment variable Priority Date$_i$. So we should apply LASSO to both Equation 4.2 and Equation 4.5. In this 3 step procedure, we first use variable selection on Equation 4.2. This estimation allows us to identify the control variables that have a strong predictive power for the treatment variable. In the second step we use variable selection on Equation 4.5. This estimation allows us to identify the control variables that have a strong predictive power for the outcome. In the last step, we estimate the model with OLS, as in Equation 4.1, using the treatment variable and union of all other variables we selected in the previous steps. This Post-Double-Selection approach decreases the omitted variable bias that may be caused by variable selection by LASSO (Belloni, Chernozhukov, and Hansen 2014b, 2014a) and we can interpret the α parameter as causal effect.

Standard errors for the estimated coefficients can be calculated in different ways. We use year, county, decade, GMD, and parcels for clustered standard errors. We also estimate Conley standard errors for different cutoff distances and decide that 300 km is the optimal cutoff distance[3].

4.3.2 Generalized Additive Models

Generalized Additive Models (GAMs) (Hastie and Tibshirani 1986, 1990) are an extension of the standard linear model. GAMs assume that the mean of the response variable depends on additive predictors through some nonlinear link function. We can write GAMs for the regression model in the water rights context as follows;

$$\log\left(\text{Real Price per Acre}\right)_i = \beta_0 + \alpha f(\text{Priority Date}_i) + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{4.6}$$

where $f(\dot{})$ is a smooth nonlinear function like polynomials or splines. By introducing nonlinear functions in the model, GAMs allow one to model nonlinear relations that standard linear regression can miss (James et al. 2013). GAMs estimation will allow us to see if the water right seniority has a nonlinear effect on irrigated agricultural land prices.

4.4 Results

4.4.1 Post-Double-Selection Results

In June 28, 1945, the Kansas Water Appropriation Act was enacted and changed the water allocation law in Kansas (Peck 1994). The Priority Date variable in the models we estimated, denotes the number of years after the enactment of Kansas Water Appropriation Act so a larger Priority Date means that the water right associated with the land is newer. Our expected sign for Priority Date is therefore negative. Figure 4.1 shows the histogram for the Priority Years.

One point related to data used needs a clarification at this point. Our data includes land sales over time and some of these sales are repeated sales however they constitue a very

[3] Conley standard errors are calculated using Fiona Burlig's lecture notes which are available at https://www.fionaburlig.com/s/ARE_212_Section_10-kxr6.pdf.

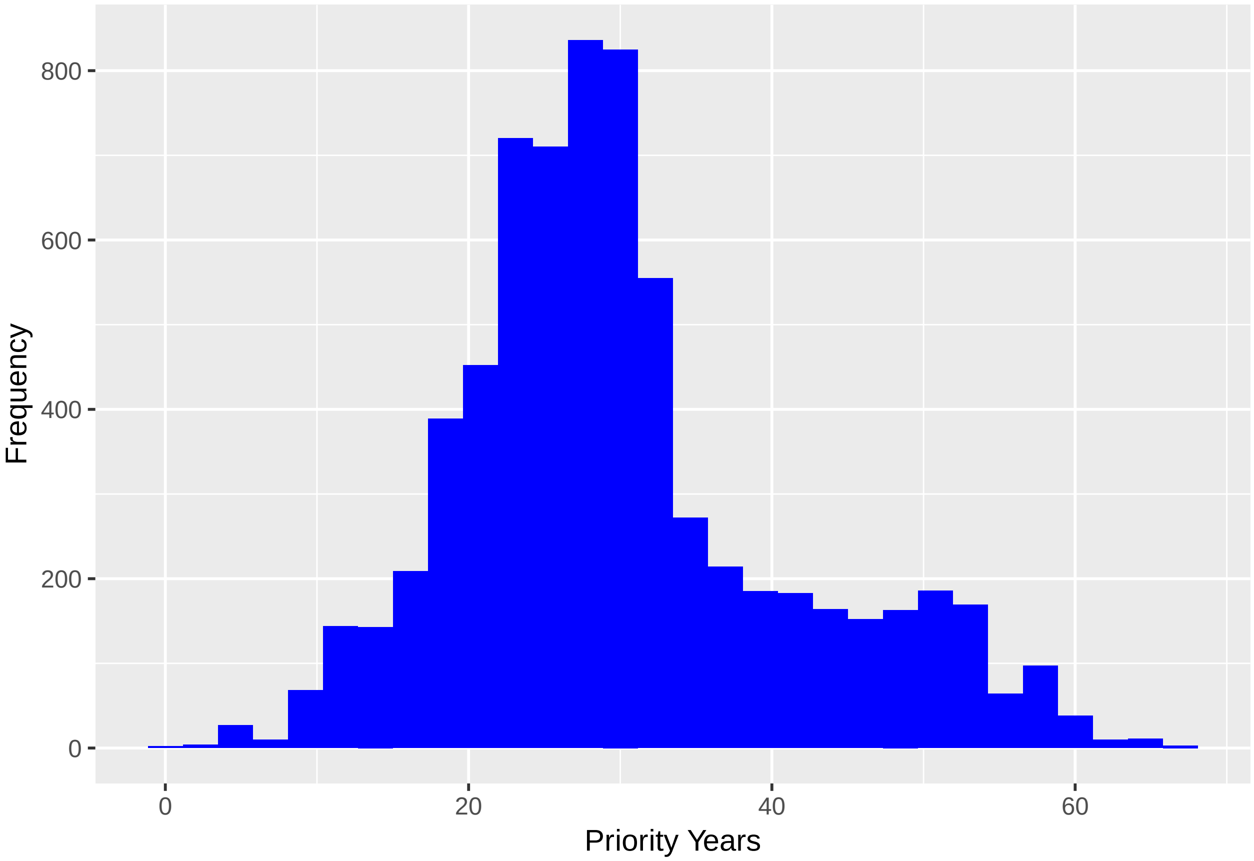

Figure 4.1: Histogram of Priority Years

small portion of all sales. We cannot estimate a fixed effects model since the seniority of the water rights does not change over time.

Post-Double-Selection LASSO results for the whole sample are reported in Table 4.1. LASSO model selects 53 variables in total; 12 county dummies, 16 year dummies, 11 soil characteristics, 4 interaction terms, 2 basis variables, 1 precipitation index, 1 weather variable, 1 hydrological variable, 3 PVD variables, and 1 water variable along with our variable of interest, which is Priority Date. As seen from Table 4.1, the estimated coefficient is negative, as expected, and highly statistically significant ($p < 0.01$). A negative value for the estimated coefficient actually indicates that having an older water right has a positive effect on land price. Since the coefficient on Priority Date is 0.0076, we can say that a one year increase in Priority Date will in turn decrease the real price per acre by 0.76%.

The mean price in the sample is around $945 per acre so a 0.76% increase will correspond to an increase of $7.182 per acre in 1982-1984 dollars. When we convert this to 2015 dollars it corresponds to a value of nearly $17 per acre for an additional year of water right seniority. For a parcel with priority date of 1965 the predicted price per acre is around $2,340 whereas for a parcel with priority date of 1985 the predicted price per acre is around $2,011 in 2015 dollars.

Table 4.1: Post-Double-Selection LASSO Regression Results

Variable	Estimate	Std. Error	t-value
Priority Date (Years)	-0.0076	0.0014	-5.3657

Note:
Other parameter estimates are not reported.

4.4.2 Nonlinear Estimation Results

We also conducted a nonlinear estimation using Generalized Additive Models (GAMs) using the variables selected by the double-selection LASSO.[4] Figure 4.2 shows the plot from the estimation result of the GAM. The black line denotes the cubic spline fit which shows the relationship between Priority Date and Price of Land.

As seen from the plot, for very low values of Priority Date (up to around 1960) we see that the land price is decreasing with Priority Date though there is a high degree of uncertainty. From 1960 to 1980, it seems that Priority Date have no effect on Land Price. After 1980 again we see a decresing effect of Priority Date on the land price. We see that if priority years get lower, the price of land is decreasing. This may happen because the lands without any water rights might be bought for non-farming reasons.

4.4.3 Heterogeneous Estimation Results

After LASSO estimation for the whole date range, LASSO for decades are estimated. LASSO estimations for each decade are reported in Table 4.2. As seen from the table, estimated

[4] The model is estimated with `mgcv` (Wood 2018) package in R (R Core Team 2018) using the variables selected in LASSO estimation.

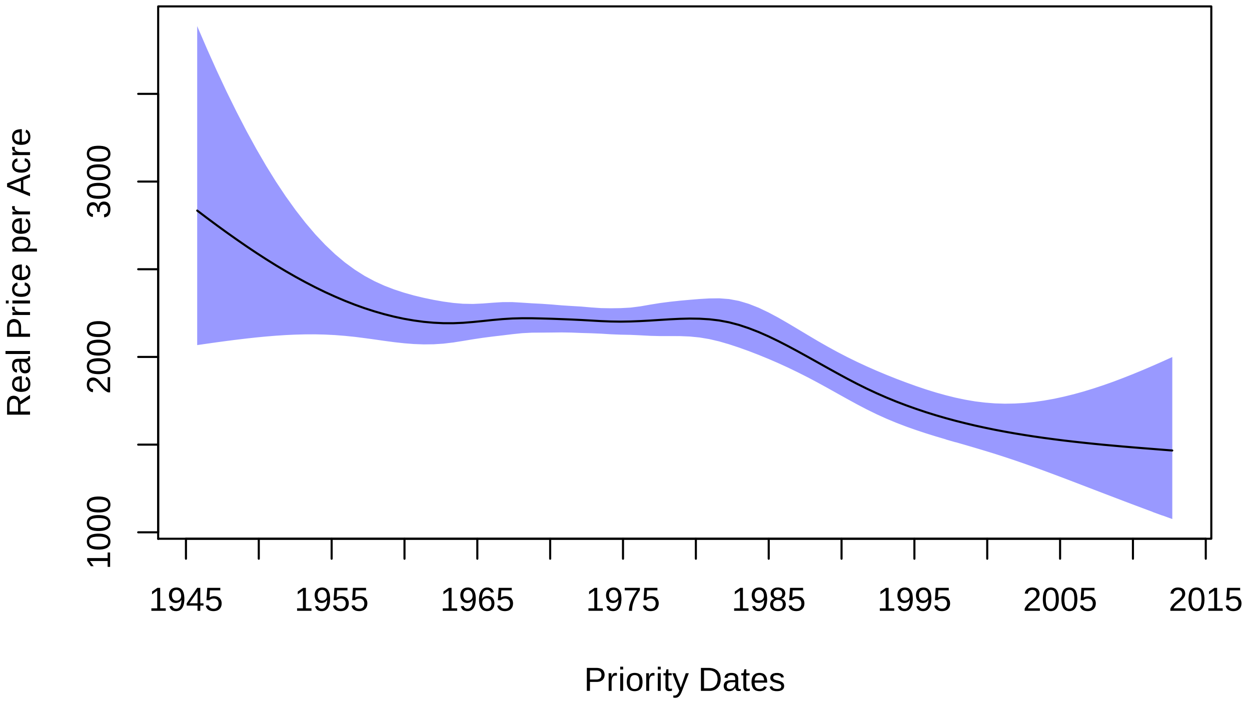

Figure 4.2: Generalized Additive Model Plot

coefficient in 2005-2015 period is statistically insignificant but has the expected sign. Estimated coefficient is higher in magnitude for 1995-2004 period. Compared to whole sample the magnitude of the coefficient of interest is higher in first two decades.

Table 4.2: Post-Double-Selection LASSO Regression Results by Decade

Decade	Estimate	Std. Error	t-value	p-value
1985-1994	-0.0090	0.0025	-3.5736	0.0004
1995-2004	-0.0113	0.0024	-4.7433	0.0000
2005-2015	-0.0005	0.0023	-0.2191	0.8266

We can convert estimation results into dollar values for each decade. Table 4.3 shows the mean price per acre and estimated dollar effect per acre in 1982–1984 dollars for different decades. Similarly, Table 4.4 shows the estimated dollar effect per acre along with 95% confidence interval in 2015 dollars.

Table 4.3: Estimation Results in 1982-1984 Dollars by Decade

Decade	Mean Price	Dollar Effect
1985-1994	633.6069	-5.6848
1995-2004	791.4427	-8.9810
2005-2015	1364.9162	-0.6763

For example, in Table 4.3 the mean price in 1985–1994 interval is around \$633 per acre so a 0.9% increase will correspond to an increase of \$5.6848 per acre in 1982–1984 dollars. When we convert this into 2015 dollars, shown in Table 4.4, it corresponds to a value of nearly \$13 per acre for an additional year of water right seniority.

Table 4.4: Estimation Results in 2015 Dollars by Decade

Decade	Dollar Effect	CI Lower	CI Upper
1985-1994	-13.4727	-13.5390	-13.4064
1995-2004	-21.2844	-21.3842	-21.1846
2005-2015	-1.6029	-1.6100	-1.5958

Groundwater Management Districts (GMDs), which are local units of government, provide water-use administration and planning along with information. Primary use of ground water in these areas is irrigation. There are five GMDs in Kansas (See Figure 4.3). For each of these districts we estimated LASSO regressions. The results are reported in Table 4.5.

Table 4.5: Post-Double-Selection LASSO Regression Results by GMD

Groundwater Management District	Estimate	Std. Error	t-value	p-value
GMD 1	0.0203	0.0082	2.4667	0.0136
GMD 2	-0.0017	0.0050	-0.3387	0.7349
GMD 3	-0.0129	0.0020	-6.4414	0.0000
GMD 4	-0.0044	0.0033	-1.3216	0.1863
GMD 5	-0.0086	0.0034	-2.5409	0.0111

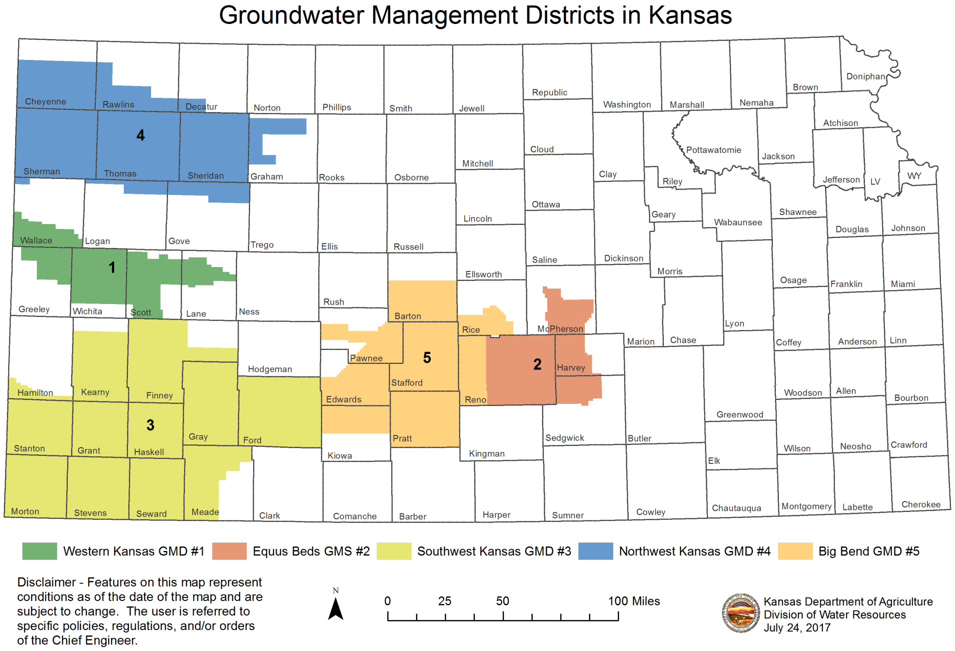

Figure 4.3: Groundwater Management Districts

Results suggest that the effect of Priority Date has different effects on price in different regions of the state. As seen from the table in some GMDs we have statistically insignificant estimation results. Effect of Prioriy Date is statistically significant in GMD 3 and GMD 5 with the expected signs. This result shows that the impact of seniority appears to be strongest in GMD 3 and GMD 5. GMD 3 and GMD 5 are areas where the water reservoirs have substantially depleted in the last decade.

Table 4.6: Estimation Results in 1982-1984 Dollars by GMD

GMD	Mean Price	Dollar Effect
GMD 1	670.6124	13.6409
GMD 2	928.4708	-1.5819
GMD 3	1080.5415	-13.9091
GMD 4	666.5652	-2.9002
GMD 5	852.3138	-7.3629

We can also convert these estimation results into dollar values for each GMD. Table 4.6 shows the mean price per acre and estimated dollar effect per acre in 1982–1984 dollars in each GMD. Similarly, Table 4.7 shows the estimated dollar effect per acre along with 95% confidence interval in 2015 dollars.

Table 4.7: Estimation Results in 2015 Dollars by GMD

GMD	Dollar Effect	CI Lower	CI Upper
GMD 1	32.3282	31.8057	32.8507
GMD 2	-3.7490	-3.7860	-3.7120
GMD 3	-32.9639	-33.0930	-32.8347
GMD 4	-6.8733	-6.9177	-6.8290
GMD 5	-17.4498	-17.5660	-17.3335

For example, in Table 4.6 the mean price in the GMD 3 is around $1,080 per acre so a 1.29% increase will correspond to an increase of $13.9091 per acre in 1982–1984 dollars. When we convert this into 2015 dollars, shown in Table 4.7, it corresponds to a value of nearly $33 per acre for an additional year of water right seniority.

Table 4.8: Clustered and Conley Standard Errors for Post-Double-Selection LASSO Regression

	Estimate	Std. Error
Year Cluster	-0.0076	0.0026
County Cluster	-0.0076	0.0027
Decade Cluster	-0.0076	0.0034
GMD Cluster	-0.0076	0.0031
Parcel Cluster	-0.0076	0.0016
Conley 300km.	-0.0076	0.0011

Since "failure to control for within-cluster error correlation can lead to very misleadingly small standard errors, and consequent misleadingly narrow confidence intervals, large t-statistics and low p-values" (Cameron and Miller 2015, 2), we also estimated Clustered and

Conley Standard Errors for LASSO. Clustered and Conley Standard Errors are reported in Table 4.8. Without any clustering the estimated standard error is found to be 0.0014. Clustering with different variables, nearly doubled the standard errors in some cases but as seen from the table, estimation results still remain statistically significant.

4.4.4 OLS Results

Lastly we estimated OLS regressions using some specific variables that are believed to be highly affecting land prices as usually done in the literature. Similar to a majority of the hedonic models in the literature OLS estimations include a very limited set of explanatory variables and therefore suffer from omitted variable bias. OLS results are reported in Table 4.9. Model (5) in the table shows the simplest regression of Priority Date on Price. The estimated coefficient in this model has the expected sign and the magnitude is higher compared to that of LASSO.

As seen from the table, estimated coefficients in all models are highly statistically significant. Model (4) adds some parcel level characteristics to the Model (5) but does not control for year or county fixed effects. As a result of adding new variables into the model, we see increased explanatory power and our coefficient of interest (Priority Days) remains highly significant. Models (1), (2), and (3) controls for County and Year fixed effects. Comparison of these models with Model (4) shows that controlling for fixed effects increases the explanatory power. When we control both for year and county fixed effects, we see that coefficient on Priority Date decreases compared to simple model, Model (5), but the explanatory power is highly increased. When we compare the estimated coefficient of Priority Dates to that of the LASSO model, we see that OLS models estimate the effect to be higher in magnitude. In Model (1), we find that the estimated coefficient of Priority Date is -0.0084. Since the mean price in the sample is around \$945 per acre, a 0.84% increase will correspond to an increase of \$7.938 per acre in 1982-1984 dollars. When we convert this to 2015 dollars it corresponds to a value of nearly \$19 per acre for an additional year of water right seniority which is \$2 more than what we found in LASSO estimation.

Table 4.9: OLS Regression Results

	log(Real Price per Acre)				
	(1)	(2)	(3)	(4)	(5)
Priority Date (Years)	-0.0084^{***}	-0.0113^{***}	-0.0096^{***}	-0.0121^{***}	-0.0099^{***}
	(0.0013)	(0.0014)	(0.0013)	(0.0014)	(0.0011)
Authorized Irrigation per Acre	-0.0043^{**}	-0.0036^{*}	-0.0067^{***}	-0.0064^{***}	
	(0.0019)	(0.0020)	(0.0019)	(0.0020)	
log(Slope)	0.0205	0.0246	-0.0468^{***}	-0.0532^{***}	
	(0.0154)	(0.0164)	(0.0133)	(0.0141)	
log(Average Soil Organic Carbon)	0.1463^{***}	0.1337^{**}	0.0584	0.0300	
	(0.0562)	(0.0598)	(0.0511)	(0.0541)	
Average National Commodity Crop Productivity Index (Overall)	0.1182	0.1907	-0.3838^{*}	−0.1694	
	(0.2714)	(0.2897)	(0.1971)	(0.2088)	
Average Root Zone Available Water Storage	-0.0022^{***}	-0.0022^{***}	-0.0012^{***}	-0.0017^{***}	
	(0.0006)	(0.0006)	(0.0005)	(0.0005)	
Percent of Dry Land	-0.4866^{***}	-0.4153^{***}	-0.6283^{***}	-0.5538^{***}	
	(0.1170)	(0.1246)	(0.1147)	(0.1215)	
Percent of Irrigated Land	0.4179^{***}	0.5968^{***}	0.4092^{***}	0.5870^{***}	
	(0.1011)	(0.1076)	(0.0981)	(0.1039)	
Precipitation	0.0002	−0.0014	0.0003	−0.0002	
	(0.0013)	(0.0014)	(0.0003)	(0.0004)	
Predevelopment Depth to Water	-0.0009^{***}	-0.0012^{***}	0.0004	−0.0002	
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	
Predevelopment Saturated Thickness	−0.0002	0.0001	0.0006^{***}	0.0007^{***}	
	(0.0002)	(0.0002)	(0.0001)	(0.0001)	
Constant	6.7708^{***}	6.6200^{***}	7.5815^{***}	7.1742^{***}	7.1342^{***}
	(0.7878)	(0.8351)	(0.4040)	(0.4237)	(0.0358)
County Fixed Effects	Yes	Yes	No	No	No
Year Fixed Effects	Yes	No	Yes	No	No
Observations	7,005	7,005	7,005	7,005	7,005
R^2	0.2350	0.1218	0.1772	0.0636	0.0111
Adjusted R^2	0.2264	0.1157	0.1724	0.0621	0.0109
F Statistic	27.2829^{***}	20.0996^{***}	36.5741^{***}	43.1451^{***}	78.3668^{***}

Notes: Standard Errors are in parentheses. * denotes $p < 0.1$, ** denotes $p < 0.05$, and *** denotes $p < 0.01$.

4.5 Conclusion

In this study we used machine learning techniques to estimate whether water right seniority is capitalized into land values in Kansas. LASSO method and its variable selection properties allows us to account for the usually overlooked omitted variable bias problem in hedonic models. Using different OLS and LASSO specifications and by introducing many parcel and geography related variables, we found that water rights have highly statistically significant positive effects on land prices.

The results show that an additional year of water right seniority causes irrigated land value to increase nearly \$17 per acre. Our GAM analysis showed that price per acre is a nonlinear function of the water rights.

Further analysis based on different Groundwater Management Districts and decades also showed that there are geographical and temporal differences in the effects of water rights. The impact of seniority is stronger in GMD 3 and GMD 5. For example in GMD 3, the effect of water rights is nearly doubled compared to our state wide effect. We find that an additional year of seniority causes land prices to increase by nearly \$33 in GMD 3.

OLS estimation results have the expected signs but the magnitude of the estimated coefficient is higher compared to that of LASSO estimation. This result suggests that OLS estimations suffer from omitted variable bias as expected.

Results have implications for different water resource managing bodies and farmers. Understanding the response of price to water right seniority in different regions may help Kansas Division of Water Resources to better govern the water allocation throughout the state. It may also help farmers while they are establishing their buying and selling decisions.

Colophon

The document was written in R (R Core Team 2018), using R Markdown (Allaire et al. 2018) and LaTeX, and rendered into PDF using wildcatdown (Ismay et al. 2018), knitr (Xie 2018b), and bookdown (Xie 2018a). This document was typeset using the XeTeX typesetting system, and the Kansas State University Thesis class. Under the hood, the Kansas State University Thesis LaTeX template is used to ensure that documents conform precisely to submission standards. Other elements of the document formatting source code have been taken from the LaTeX, Knitr, and RMarkdown templates for UC Berkeley's graduate thesis, and Dissertate: a LaTeX dissertation template to support the production and typesetting of a PhD dissertation at Harvard, Princeton, and NYU.

The computational environment that was used to generate this version is as follows:

```
Session info ----------------------------------------------------------------------------

 setting  value
 version  R version 3.5.1 (2018-07-02)
 system   x86_64, darwin15.6.0
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 tz       Europe/Istanbul
 date     2018-11-05

Packages --------------------------------------------------------------------------------

 package       * version    date       source
 abind           1.4-5      2016-07-21 CRAN (R 3.5.0)
 assertthat      0.2.0      2017-04-11 CRAN (R 3.5.0)
 backports       1.1.2      2017-12-13 CRAN (R 3.5.0)
```

```
base          * 3.5.1      2018-07-05 local
bindr           0.1.1      2018-03-13 CRAN (R 3.5.0)
bindrcpp      * 0.2.2      2018-03-29 CRAN (R 3.5.0)
bookdown        0.7        2018-02-18 CRAN (R 3.5.0)
broom         * 0.5.0      2018-07-17 CRAN (R 3.5.0)
caret         * 6.0-80     2018-05-26 CRAN (R 3.5.0)
cellranger      1.1.0      2016-07-27 CRAN (R 3.5.0)
checkmate       1.8.5      2017-10-24 CRAN (R 3.5.0)
Ckmeans.1d.dp * 4.2.1      2017-07-09 CRAN (R 3.5.0)
class           7.3-14     2015-08-30 CRAN (R 3.5.1)
cli             1.0.1.9000 2018-10-10 Github (r-lib/cli@56538e3)
codetools       0.2-15     2016-10-05 CRAN (R 3.5.1)
colorspace      1.3-2      2016-12-14 CRAN (R 3.5.0)
compiler        3.5.1      2018-07-05 local
corrplot      * 0.84       2017-10-16 CRAN (R 3.5.0)
crayon          1.3.4      2017-09-16 CRAN (R 3.5.0)
CVST            0.2-2      2018-05-26 CRAN (R 3.5.0)
data.table    * 1.11.4     2018-05-27 cran (@1.11.4)
datasets      * 3.5.1      2018-07-05 local
ddalpha         1.3.4      2018-06-23 CRAN (R 3.5.0)
DEoptimR        1.0-8      2016-11-19 CRAN (R 3.5.0)
devtools        1.13.6     2018-06-27 CRAN (R 3.5.0)
digest          0.6.16     2018-08-22 cran (@0.6.16)
dimRed          0.1.0      2017-05-04 CRAN (R 3.5.0)
dplyr         * 0.7.6      2018-06-29 cran (@0.7.6)
DRR             0.0.3      2018-01-06 CRAN (R 3.5.0)
DT            * 0.4        2018-01-30 CRAN (R 3.5.0)
e1071         * 1.7-0      2018-07-28 CRAN (R 3.5.0)
evaluate        0.11       2018-07-17 CRAN (R 3.5.0)
forcats       * 0.3.0      2018-02-19 CRAN (R 3.5.0)
foreach       * 1.4.4      2017-12-12 CRAN (R 3.5.0)
Formula         1.2-3      2018-05-03 CRAN (R 3.5.0)
geometry        0.3-6      2015-09-09 CRAN (R 3.5.0)
```

```
geosphere       1.5-7      2017-11-05 CRAN (R 3.5.0)
ggfortify     * 0.4.5      2018-05-26 CRAN (R 3.5.0)
ggmap         * 2.6.1      2016-01-23 CRAN (R 3.5.0)
ggplot2       * 3.0.0      2018-07-03 CRAN (R 3.5.0)
git2r           0.23.0     2018-07-17 CRAN (R 3.5.0)
glue            1.3.0      2018-07-17 CRAN (R 3.5.0)
gower           0.1.2      2017-02-23 CRAN (R 3.5.0)
graphics      * 3.5.1      2018-07-05 local
grDevices     * 3.5.1      2018-07-05 local
grid            3.5.1      2018-07-05 local
gridExtra       2.3        2017-09-09 CRAN (R 3.5.0)
gtable          0.2.0      2016-02-26 CRAN (R 3.5.0)
haven           1.1.2      2018-06-27 CRAN (R 3.5.0)
hdm           * 0.2.0      2016-06-17 CRAN (R 3.5.0)
hms             0.4.2      2018-03-10 CRAN (R 3.5.0)
htmltools       0.3.6.9003 2018-08-26 Github (rstudio/htmltools@a9e969f)
htmlwidgets     1.2        2018-04-19 CRAN (R 3.5.0)
httr            1.3.1      2017-08-20 CRAN (R 3.5.0)
ipred           0.9-6      2017-03-01 CRAN (R 3.5.0)
iterators     * 1.0.10     2018-07-13 CRAN (R 3.5.0)
itertools     * 0.1-3      2014-03-12 CRAN (R 3.5.0)
jpeg            0.1-8      2014-01-23 CRAN (R 3.5.0)
jsonlite        1.5        2017-06-01 CRAN (R 3.5.0)
kableExtra    * 0.9.0      2018-05-21 CRAN (R 3.5.0)
kernlab         0.9-27     2018-08-10 CRAN (R 3.5.0)
knitr         * 1.20       2018-02-20 CRAN (R 3.5.0)
labeling        0.3        2014-08-23 CRAN (R 3.5.0)
lattice       * 0.20-35    2017-03-25 CRAN (R 3.5.1)
lava            1.6.3      2018-08-10 CRAN (R 3.5.0)
lazyeval        0.2.1      2017-10-29 CRAN (R 3.5.0)
lmtest        * 0.9-36     2018-04-04 CRAN (R 3.5.0)
lubridate     * 1.7.4      2018-04-11 CRAN (R 3.5.0)
magic           1.5-8      2018-01-26 CRAN (R 3.5.0)
```

```
magrittr       1.5        2014-11-22 CRAN (R 3.5.0)

mapproj        1.2.6      2018-03-29 CRAN (R 3.5.0)

maps           3.3.0      2018-04-03 CRAN (R 3.5.0)

MASS           7.3-50     2018-04-30 CRAN (R 3.5.1)

Matrix         1.2-14     2018-04-13 CRAN (R 3.5.1)

memoise        1.1.0      2017-04-21 CRAN (R 3.5.0)

methods      * 3.5.1      2018-07-05 local

mgcv         * 1.8-24     2018-06-23 CRAN (R 3.5.1)

missForest   * 1.4        2013-12-31 CRAN (R 3.5.0)

mlbench      * 2.1-1      2012-07-10 CRAN (R 3.5.0)

ModelMetrics   1.2.0      2018-08-10 CRAN (R 3.5.0)

modelr         0.1.2      2018-05-11 CRAN (R 3.5.0)

munsell        0.5.0      2018-06-12 CRAN (R 3.5.0)

nlme         * 3.1-137    2018-04-07 CRAN (R 3.5.1)

nnet           7.3-12     2016-02-02 CRAN (R 3.5.1)

parallel       3.5.1      2018-07-05 local

pillar         1.3.0      2018-07-14 CRAN (R 3.5.0)

pkgconfig      2.0.1      2017-03-21 CRAN (R 3.5.0)

pls            2.6-0      2016-12-18 CRAN (R 3.5.0)

plyr           1.8.4      2016-06-08 CRAN (R 3.5.0)

png            0.1-7      2013-12-03 CRAN (R 3.5.0)

prodlim        2018.04.18 2018-04-18 CRAN (R 3.5.0)

proto          1.0.0      2016-10-29 CRAN (R 3.5.0)

purrr        * 0.2.5      2018-05-29 cran (@0.2.5)

R6             2.2.2      2017-06-17 CRAN (R 3.5.0)

randomForest * 4.6-14     2018-03-25 CRAN (R 3.5.0)

RANN         * 2.6        2018-07-16 CRAN (R 3.5.0)

Rcpp           0.12.18    2018-07-23 CRAN (R 3.5.0)

RcppRoll       0.3.0      2018-06-05 CRAN (R 3.5.0)

readr        * 1.1.1      2017-05-16 CRAN (R 3.5.0)

readxl         1.1.0      2018-04-20 CRAN (R 3.5.0)

recipes      * 0.1.3      2018-06-16 CRAN (R 3.5.0)

reshape2       1.4.3      2017-12-11 CRAN (R 3.5.0)
```

```
RgoogleMaps     1.4.2      2018-06-08 CRAN (R 3.5.0)
rjson           0.2.20     2018-06-08 CRAN (R 3.5.0)
rlang           0.2.2      2018-08-16 cran (@0.2.2)
rmarkdown       1.10       2018-06-11 CRAN (R 3.5.0)
robustbase      0.93-2     2018-07-27 CRAN (R 3.5.0)
rpart           4.1-13     2018-02-23 CRAN (R 3.5.1)
rprojroot       1.3-2      2018-01-03 CRAN (R 3.5.0)
rsample       * 0.0.2      2017-11-12 CRAN (R 3.5.0)
rstudioapi      0.7        2017-09-07 CRAN (R 3.5.0)
rvest           0.3.2      2016-06-17 CRAN (R 3.5.0)
sandwich      * 2.5-0      2018-08-17 cran (@2.5-0)
scales        * 1.0.0      2018-08-09 CRAN (R 3.5.0)
sfsmisc         1.1-2      2018-03-05 CRAN (R 3.5.0)
skimr         * 1.0.3      2018-06-07 CRAN (R 3.5.0)
sp              1.3-1      2018-06-05 CRAN (R 3.5.0)
splines         3.5.1      2018-07-05 local
stargazer     * 5.2        2015-07-14 Github (markwestcott34/stargazer-booktabs@e93c8d2)
stats         * 3.5.1      2018-07-05 local
stats4          3.5.1      2018-07-05 local
stringi         1.2.4      2018-07-20 CRAN (R 3.5.0)
stringr       * 1.3.1      2018-05-10 CRAN (R 3.5.0)
survival        2.42-6     2018-07-13 CRAN (R 3.5.0)
tibble        * 1.4.2      2018-01-22 CRAN (R 3.5.0)
tidyr         * 0.8.1      2018-05-18 CRAN (R 3.5.0)
tidyselect      0.2.4      2018-02-26 CRAN (R 3.5.0)
tidyverse     * 1.2.1      2017-11-14 CRAN (R 3.5.0)
timeDate        3043.102   2018-02-21 CRAN (R 3.5.0)
tools           3.5.1      2018-07-05 local
utils         * 3.5.1      2018-07-05 local
viridisLite     0.3.0      2018-02-01 CRAN (R 3.5.0)
wildcatdown     0.0.1      2018-10-10 local
withr           2.1.2      2018-03-15 CRAN (R 3.5.0)
xfun            0.3        2018-07-06 CRAN (R 3.5.0)
```

```
xgboost       * 0.71.2    2018-06-09 CRAN (R 3.5.0)

xml2            1.2.0     2018-01-24 CRAN (R 3.5.0)

yaml            2.2.0     2018-07-25 cran (@2.2.0)

zoo           * 1.8-3     2018-07-16 CRAN (R 3.5.0)
```

References

Accuacre. 2018. http://accuacre.com/.

AcreValue. 2018. https://www.acrevalue.com/.

Adam-Bourdarios, C, G Cowan, C Germain-Renaud, I Guyon, B Kégl, and D Rousseau. 2015. "The Higgs Machine Learning Challenge." *Journal of Physics: Conference Series* 664 (7): 072015.

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. *Rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

Bajari, Patrick, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. 2015. "Machine Learning Methods for Demand Estimation." *The American Economic Review* 105 (5): 481–85.

Basuchoudhary, Atin, James T. Bang, and Tinni Sen. 2017. *Machine-Learning Techniques in Economics*. Springer International Publishing.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80 (6): 2369–2429.

Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least Squares After Model Selection in High-Dimensional Sparse Models." *Bernoulli* 19 (2): 521–47.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014a. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–50.

———. 2014b. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50.

Bonnin, Rodolfo. 2017. *Machine Learning for Developers: Uplift Your Regular Applications with the Power of Statistics, Analytics, and Machine Learning.* Packt Publishing Ltd.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–40.

Brent, Daniel A. 2016. "The Value of Heterogeneous Property Rights and the Costs of Water Volatility." *American Journal of Agricultural Economics* 99 (1): 73–102.

Brozovic, Nicholas, and Shahnila Islam. 2010. "Estimating the Value of Groundwater in Irrigation." *Selected Paper for Presentation at the Agricultural and Applied Economic Association.*

Buck, Steven, Maximilian Auffhammer, and David Sunding. 2014. "Land Markets and the Value of Water: Hedonic Analysis Using Repeat Sales of Farmland." *American Journal of Agricultural Economics* 96 (4): 953–69.

Burns, Christopher, Nigel Key, Sarah Tulman, Allison Borchers, and Jeremy Weber. 2018. "Farmland Values, Land Ownership, and Returns to Farmland, 2000-2016." *ERR-245, U.S. Department of Agriculture, Economic Research Service, February 2018.*

Burt, Oscar R. 1986. "Econometric Modeling of the Capitalization Formula for Farmland Prices." *American Journal of Agricultural Economics* 68 (1): 10.

Butsic, Van, and Noelwah R. Netusil. 2007. "Valuing Water Rights in Douglas County, Oregon, Using the Hedonic Price Method." *Journal of the American Water Resources Association* 43 (3): 622–29.

Cameron, A. Colin, and Douglas L Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50 (2): 317–72.

Caplin, Andrew, Sumit Chopra, John V Leahy, Yann Lecun, and Trivikraman Thampy. 2008. “Machine Learning and the Spatial Structure of House Prices and Housing Returns.” Available at SSRN: https://ssrn.com/abstract=1316046.

Chai, T., and R. R. Draxler. 2014. “Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments Against Avoiding RMSE in the Literature.” *Geoscientific Model Development* 7 (3): 1247–50.

Chen, Tianqi, and Carlos Guestrin. 2016. “Xgboost: A Scalable Tree Boosting System.” *CoRR* abs/1603.02754.

Coble, Keith H, Ashok K Mishra, Shannon Ferrell, and Terry Griffin. 2018. “Big Data in Agriculture: A Challenge for the Future.” *Applied Economic Perspectives and Policy* 40 (1): 79–96.

Cropper, Maureen L., Leland B. Deck, and Kenenth E. McConnell. 1988. “On the Choice of Funtional Form for Hedonic Price Functions.” *The Review of Economics and Statistics* 70 (4): 668.

Crouter, Jan P. 1987. “Hedonic Estimation Applied to a Water Rights Market.” *Land Economics* 63 (3): 259.

Esri. 2018. “The SSURGO Tool Box.” Available online at http://resources.arcgis.com/en/communities/hydro/01vn0000001m000000.htm. Accessed [05/03/2018].

Faux, John, and Gregory M. Perry. 1999. “Estimating Irrigation Water Value Using Hedonic Price Analysis: A Case Study in Malheur County, Oregon.” *Land Economics* 75 (3): 440.

Featherstone, Allen M., and Timothy G. Baker. 1987. “An Examination of Farm Sector Real Asset Dynamics: 1910-85.” *American Journal of Agricultural Economics* 69 (3): 532.

Freeman III, A Myrick, Joseph A Herriges, and Catherine L Kling. 2014. *The Measurement of Environmental and Resource Values: Theory and Methods*. London: Routledge.

Freund, Yoav, and Robert E Schapire. 1997. "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55 (1): 119–39.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. "Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors)." *The Annals of Statistics* 28 (2): 337–407.

Golden, Bill, and John Leatherman. 2017. "Impact Analysis of the Walnut Creek Intensive Groundwater Use Control Area." *Journal of Regional Analysis & Policy* 47 (2): 176–87.

Goodwin, Barry K, Ashok K Mishra, and François N Ortalo-Magné. 2003. "What's Wrong with Our Models of Agricultural Land Values? Agricultural Land Values, Government Payments, and Production (Allen Featherstone, Kansas State University, Presiding)." *American Journal of Agricultural Economics* 85 (3): 744–52.

Google. 2018. "Google Distance Matrix API." https://goo.gl/43xKkU.

Guiling, P., B. W. Brorsen, and D. Doye. 2009. "Effect of Urban Proximity on Agricultural Land Values." *Land Economics* 85 (2): 252–64.

Hartman, Loyal M., and R. Garth Taylor. 1989a. "Irrigated Land Values in Eastern Colorado: An Analysis of Farm Sales Prices for Pump Irrigated Land Overlying the Ogallala Aquifer." *Bulletin LTB89-01, Colorado State University Agricultural Experiment Station, Fort Collins, CO.*

———. 1989b. "Irrigated Land Values Is Eastern Colorado." *Colorado State University,* *Agricultural Experiment Station, Technical Bulletin LTB 89-1.*

Hastie, T. J., and R. J. Tibshirani. 1990. *Generalized Additive Models.* Chapman; Hall/CRC.

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. 2001. *The Elements of Statistical Learning.* Springer New York.

Hastie, Trevor, and Robert Tibshirani. 1986. "Generalized Additive Models." *Statist. Sci.* 1 (3): 297–310.

Heath, Ralph C. 1983. "Basic Ground-Water Hydrology." U.S. Geological Survey Water-Supply Paper 2200.

Hendricks, Nathan P. 2018. "Potential Benefits from Innovations to Reduce Heat and Water Stress in Agriculture." *Journal of the Association of Environmental and Resource Economists* 5 (3): 545–76.

Hiemstra, Paul. 2013. *Automap: Automatic Interpolation Package.* https://CRAN.R-project.org/package=automap.

Ho, Jenny. 2017. "Essays on Machine Learning in Applied Microeconomics." PhD thesis, Seattle, Washington: University of Washington.

Hornbeck, Richard, and Pinar Keskin. 2014. "The Historically Evolving Impact of the Ogallala Aquifer: Agricultural Adaptation to Groundwater and Drought." *American Economic Journal: Applied Economics* 6 (1): 190–219.

Ifft, Jennifer, Daniel P Bigelow, and Jeffrey Savage. 2018. "The Impact of Irrigation Restrictions on Cropland Values in Nebraska." *Journal of Agricultural and Resource Economics* 43 (2): 195–S11.

Ifft, Jennifer, Ryan Kuhns, and Kevin Patrick. 2018. "Can Machine Learning Improve Prediction an Application with Farm Survey Data." *International Food and Agribusiness Management Review*, September, 1–16.

Islam, Shahnila. 2010. "Estimating the Value of Groundwater in Irrigation." Master's thesis, Urbana, Illinois: University of Illinois at Urbana-Champaign.

Ismay, Chester, Nick Solomon, Ben Marwick, and Emrah Er. 2018. *Wildcatdown: An R Markdown Thesis Template for the Kansas State University*.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer New York.

Jenkins, Allan, Bruce Elder, Ram Valluru, and Paul Burger. 2007. "Water Rights and Land Values in the West-Central Plains." *Great Plains Research*, 101–11.

Just, Richard E, and John A. Miranowski. 1993. "Understanding Farmland Price Changes." *American Journal of Agricultural Economics* 75 (1): 156–68.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *The American Economic Review* 105 (5): 491–95.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer Science & Business Media.

Lantz, Brett. 2015. *Machine Learning with R*. Packt Publishing Ltd.

Lawell, C-Y Cynthia Lin. 2017. "Property Rights and Groundwater Management in the High Plains Aquifer." Available at www.des.ucdavis.edu/faculty/Lin/water_temporal_property_rts_paper.pdf.

Leeb, Hannes, and Benedikt M Pötscher. 2008a. "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory* 24 (02): 338–76.

———. 2008b. "Recent Developments in Model Selection and Related Areas." *Econometric Theory* 24 (02): 319–22.

Limsombunchao, V. 2004. "House Price Prediction: Hedonic Price Model Vs. Artificial

Neural Network." *American Journal of Applied Sciences* 1 (3): 193–20.

Ludwig, Nicole, Stefan Feuerriegel, and Dirk Neumann. 2015. "Putting Big Data Analytics to Work: Feature Selection for Forecasting Electricity Prices Using the LASSO and Random Forests." *Journal of Decision Systems* 24 (1): 19–36.

Lusk, Jayson L. 2017. "Consumer Research with Big Data: Applications from the Food Demand Survey (FooDS)." *American Journal of Agricultural Economics* 99 (2): 303–20.

McNew, Kevin, and Duane Griffith. 2005. "Measuring the Impact of Ethanol Plants on Local Grain Prices." *Review of Agricultural Economics* 27 (2): 164–80.

Miranowski, John A., and Brian D. Hammes. 1984. "Implicit Prices of Soil Characteristics for Farmland in Iowa." *American Journal of Agricultural Economics* 66 (5): 745.

Moss, C. B. 1997. "Returns, Interest Rates, and Inflation: How They Explain Changes in Farmland Values." *American Journal of Agricultural Economics* 79 (4): 1311–8.

Mu, Jingyi, Fang Wu, and Aihua Zhang. 2014. "Housing Value Forecasting Based on Machine Learning Methods." In *Abstract and Applied Analysis*. Vol. 2014. Hindawi Publishing Corporation.

Mukherjee, M., and K. Schwabe. 2014. "Irrigated Agricultural Adaptation to Water and Climate Variability: The Economic Value of a Water Portfolio." *American Journal of Agricultural Economics* 97 (3): 809–32.

Nickerson, Cynthia J, Mitchell Morehart, Todd Kuethe, Jayson Beckman, Jennifer Ifft, and Ryan Williams. 2012. *Trends in US Farmland Values and Ownership*. US Department of Agriculture, Economic Research Service Wasington, DC.

Nickerson, Cynthia, and Wendong Zhang. 2014. "Modeling the Determinants of Farmland Values in the United States." In *The Oxford Handbook of Land Economics*, edited by Joshua M. Duke and JunJie Wu, 111–38. Oxford University Press.

Nowak, Adam, and Patrick Smith. 2016. “Textual Analysis in Real Estate.” *Journal of Applied Econometrics* 32 (4): 896–918.

Oudendag, Diti, Zoltán Szlávik, and Hennie van der Veen. 2012. “The Use of Machine Learning Techniques to Predict Farm Size Change-an Implementation in the Dutch Dairy Sector.” *American Academic & Scholarly Research Journal* 4 (5): 1.

Palmquist, Raymond B. 2005. “Property Value Models.” In *Handbook of Environmental Economics*, edited by Daniel W Bromley and others, 2:763–819. Elsevier.

Palmquist, Raymond B, and Leon E Danielson. 1989. “A Hedonic Study of the Effects of Erosion Control and Drainage on Farmland Values.” *American Journal of Agricultural Economics* 71 (1): 55–62.

Park, Byeonghwa, and Jae Kwon Bae. 2015. “Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data.” *Expert Systems with Applications* 42 (6): 2928–34.

Peck, J. C. 1994. “The Kansas Water Appropriation Act: A Fifty-Year Perspective.” *U. Kan. L. Rev.* 43: 735.

———. 2002. “Property Rights in Groundwater-Some Lessons from the Kansas Experience.” *Kansas Journal of Law & Public Policy* 12: 493.

———. 2007. “Groundwater Management in the High Plains Aquifer in the USA: Legal Problems and Innovations.” In *The Agricultural Groundwater Revolution: Opportunities and Threats to Development*, edited by M. Geordano and K. G. Villholth, 296–319. CAB International, Wallingford, UK.

Petrie, R. A., and L. O. Taylor. 2007. “Estimating the Value of Water Use Permits: A Hedonic Approach Applied to Farmland in the Southeastern United States.” *Land Economics* 83 (3): 302–18.

Povoa, Lucas Venezian, and Jonas Teixeira Nery. 2016. *Precintcon: Precipitation Intensity, Concentration and Anomaly Analysis*. https://CRAN.R-project.org/package=precintcon.

PRISM Climate Group. 2016. "Oregon State University." Available online at http://prism.oregonstate.edu. Accessed [03/03/2016].

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82 (1): 34–55.

Schloss, J. A., and R. W. Buddemeier. 2000. "Predevelopment Saturated Thickness." Available online at http://www.kgs.ku.edu/HighPlains/atlas/atpst.htm. Accessed [08/20/2018].

Schöni, Olivier. 2014. "Four Essays on Statistical Problems of Hedonic Methods." PhD thesis, Fribourg, Switzerland: University of Fribourg.

Shi, Yue Jin, Timothy T Phipps, and Dale Colyer. 1997. "Agricultural Land Values Under Urbanizing Influences." *Land Economics*, 90–100.

Snyder, Rl. 1985. "Hand Calculating Degree Days." *Agricultural and Forest Meteorology* 35 (1): 353–58.

Soil Survey Staff. 2016. "Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey." Available online at http://websoilsurvey.nrcs.usda.gov/. Accessed [04/20/2016].

Sunderland, David H., James D. Libbin, and L. Allen Torell. 1987. "Estimated Water Values for Tax Depletion Allowance in New Mexico." *Research Report, New Mexico University, College of Agriculture and Home Economics, Agricultural Experiment Station.*

Therneau, Terry, and Beth Atkinson. 2018. *Rpart: Recursive Partitioning and Regression Trees*. https://CRAN.R-project.org/package=rpart.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–88.

Torell, L. Allen, James D. Libbin, and Michael D. Miller. 1990. "The Market Value of Water in the Ogallala Aquifer." *Land Economics* 66 (2): 163.

Tremblay, Nicolas, Yacine M. Bouroubi, Carl Bélec, Robert William Mullen, Newell R. Kitchen, Wade E. Thomason, Steve Ebelhar, et al. 2012. "Corn Response to Nitrogen Is Influenced by Soil Texture and Weather." *Agronomy Journal* 104 (6): 1658.

Tsoodle, Leah J, Bill B Golden, and Allen M Featherstone. 2006. "Factors Influencing Kansas Agricultural Farm Land Values." *Land Economics* 82 (1): 124–39.

UC Davis California Soil Resource Lab. 2018. "Analysis of SSURGO Data in PostGIS: An Overview." Available online at https://goo.gl/LDr3vV. Accessed [05/03/2018].

Water Rights Information System (WRIS). 2015. "Kansas Department of Agriculture, Division of Water Resources."

Willmott, CJ, and K Matsuura. 2005. "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance." *Climate Research* 30: 79–82.

Wilson, B, J Bartley, K Emmons, J Bagley, J Wason, and S Stankiewicz. 2005. "Water Information Management and Analysis System, Version 5, for the Web. User Manual." *Kansas Geological Survey Open File Report 2005-30.*, 37.

Wood, Simon. 2018. *Mgcv: Mixed Gam Computation Vehicle with Automatic Smoothness Estimation*. https://CRAN.R-project.org/package=mgcv.

Xie, Yihui. 2018a. *Bookdown: Authoring Books and Technical Documents with R Markdown*.

https://CRAN.R-project.org/package=bookdown.

———. 2018b. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://CRAN.R-project.org/package=knitr.

Yeo, Michelle, Tristan Fletcher, and John Shawe-Taylor. 2015. "Machine Learning in Fine Wine Price Prediction." *Journal of Wine Economics*, 1–22.

Yli-Heikkilä, Maria, Jukka Tauriainen, Mika Sulkava, and Others. 2015. "Predicting the Profitability of Agricultural Enterprises in Dairy Farming." In *23 Rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Esann 2015: Proceedings/Ed. M. Verleysen.* Ciaco scrl.

Yoo, Sanglim, Jungho Im, and John E Wagner. 2012. "Variable Selection for Hedonic Model Using Machine Learning Approaches: A Case Study in Onondaga County, NY." *Landscape and Urban Planning* 107 (3): 293–306.

Zhang, W., and C. J. Nickerson. 2015. "Housing Market Bust and Farmland Values: Identifying the Changing Influence of Proximity to Urban Centers." *Land Economics* 91 (4): 605–26.

Zięba, Maciej, Sebastian K. Tomczak, and Jakub M. Tomczak. 2016. "Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction." *Expert Systems with Applications* 58 (October): 93–101.

Zillow. 2018. https://www.zillow.com/.