11. Better Predictions of Land Values Using Machine Learning

Nathan Hendricks <nph@ksu.edu>

Nathan Hendricks is an Associate Professor in Agricultural Economics at Kansas State University. He holds B.S. and M.S. from KSU and a Ph.D. from University of California, Davis. His research is in the areas of production, policy, and the intersection of production and the environment. He teaches an undergraduate course on global food systems, a graduate course on agricultural policy, and a graduate team-taught course in quantitative methods.

Emrah Er <eremrah@ksu.edu>

Emrah Er is a Ph.D. candidate in Agricultural Economics at Kansas State University. He holds B.A. from Anadolu University, Turkey, M.S. from Ankara University, Turkey, and M.A. from North Carolina State University. His research is in the areas of land use, environmental economics, and applied econometrics.

Mykel Taylor <mtaylor@ksu.edu>

Dr. Taylor's research and extension programs are focused in the areas of crop marketing and farm management. She grew up on a cattle ranch in Montana and attended Montana State University majoring in Agribusiness Management. She has worked in extension positions at both Kansas State University and Washington State University. Some of her current research areas include measuring basis risk for commodity grains, understanding the implications of the 2014 Farm Bill, and analyzing trends in Kansas agricultural land values, rental rates, and leasing arrangements.

Abstract/Summary

Machine learning is a modern set of methods in data analysis that can be used to give better predictions. We report findings from our research that employs machine learning to predict land values using parcel-level transactions in Kansas. The emphasis is on predicting the variation in land values across fields rather than predicting the variation in land values over time. The presentation will compare the predictive ability of machine learning techniques to traditional methods and also discuss which characteristics of the parcel (soil, climate, proximity to urban areas, etc.) provide the greatest predictive ability.

Better Predictions of Land Values Using Machine Learning?

Nathan Hendricks, Emrah Er, and Mykel Taylor

Risk and Profit Conference
August 18–19, 2016

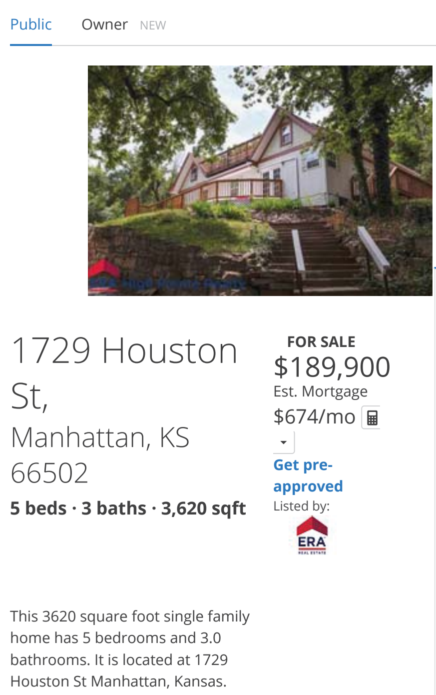
Our Objective

Improve understanding of variations in land values across different parcels at a particular point in time.

Why Predict Land Values?

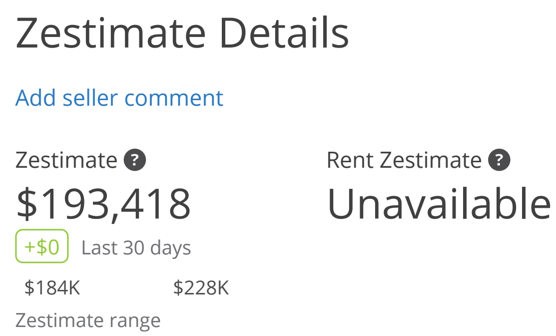
- Inform land purchasing decisions
- Government buy-out programs (e.g., CRP) could differentiate payments by land quality in order to reduce expenditures

Zillow Listing



Source: `http://www.zillow.com/`

Zestimate



Motivation Data Methods Preliminary Results Future Extensions

AcreValue by Granular

"AcreValue aggregates and analyzes terabytes of data about soils, climate, crop rotations, taxes, interest rates, and corn prices to calculate the estimated value of an individual field. With its map-based web interface, AcreValue provides easy access to this data with a simple click on a field. AcreValue is currently available in all U.S. states except Alaska, Hawaii, Alabama, and Florida. Estimated values are currently available in Iowa, Illinois, Indiana, and Minnesota."
Source: `https://www.acrevalue.com/about/`

- Two examples of AcreValue Reports
- Note: We make no endorsement or critique of the services provided by AcreValue.

PARCEL (PVD)

- Sales between 1985 and 2015.
- Only dryland parcels
- **Drop:** No Parcel ID, Acres $\leq$ 35, Sold within same month, Not arms-length.
- 22,768 unique parcels.
- Price Per Acre = (Sale Price - Total Improvements)/Total Acres.

Climate (PRISM)

- Source: PRISM gridded data
- April-September precipitation
- Extreme heat (degree days above 30 degrees C)

Soils

- Know the soil types associated with each parcel
- Collect information related to those soil types from SSURGO by NRCS

Commute Time to Urban Areas

- Source: Google Maps Distance Matrix API
- Match Populated Areas with Parcels.
- Populated areas with population of 10k, 20k, 40k, 50k, 100k, 200k, 500k, and 1m.

- CPI to deflate land values
- Calculate average percent difference in price for each year compared to 2015
- Adjust all prices to 2015 levels
- Similarly, adjust all prices to 4th quarter

Hedonic Model

$$ln(PricePerAcre)_i = f(x_i) + \varepsilon_i$$

- x_i: characteristics of parcel i

Assessing the Model

- Step 1: **Train** the model by taking a portion of the data to fit the model
- Step 2: **Test** the model by seeing how well it predicts a new sample of data that was not used in training (out-of-sample prediction)

Methods of Modeling the Data

- Regression
- Ridge Regression
- Lasso
- Random Forest

Interpretability versus Flexibility

- Regression is easy to interpret, but may not predict well with complex nonlinear relationships
- A Random Forest is difficult to interpret, but allows complex nonlinear relationships between several variables

COEFFICIENTS FROM SIMPLE REGRESSION

- Another inch of precipitation increases land value by 3%
- A 10% increase in slope decreases land value 0.6%
- A 10% increase in soil organic carbon increases land value by 1.7%
- Effect of soil water holding capacity seems to be captured mostly by soil organic carbon

COEFFICIENTS FROM SIMPLE REGRESSION - CONT'D

- Greater bulk density (more susceptible to compaction) decreases land value
- Having no commute time to 50k city increases value of land by 85% compared to commute 1 hour or greater
- Each minute less of commute time increases value by 1.4%.

Most Important Soil Predictors Selected by Lasso (No Particular Order)

- Slope
- Soil Organic Carbon
- Bulk Density
- Cation-exchange capacity
- Low pH
- Sandy soils
- NCCPI (National Commodity Crop Productivity Index) Rating
- Root zone depth
- Non-irrigated capability classification

Most Important Soil Predictors Indicated by Random Forest (Order of Importance)

1. Slope
2. Soil Organic Carbon
3. Saturated Hydraulic Conductivity (Ksat)
4. Percent Silt
5. NCCPI Rating
6. Soil water holding capacity
7. Organic matter

Median Absolute Errors of Prediction

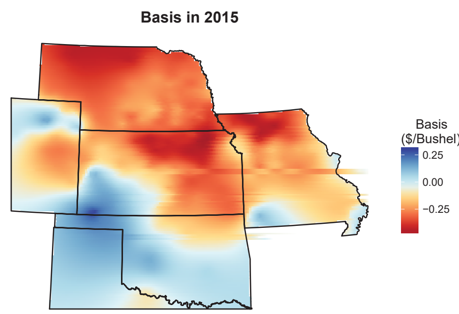
OLS	Ridge	LASSO	Random Forest
0.433	0.435	0.436	0.420

Future Extensions

- Region-specific models
- Use average sales of nearby parcels (or parcels with similar soil) as explanatory variable
- Include additional explanatory variables

Basis (AgManager.info)

- Corn & Wheat.
- 1998 to 2015.
- Calculate 3 year moving average.



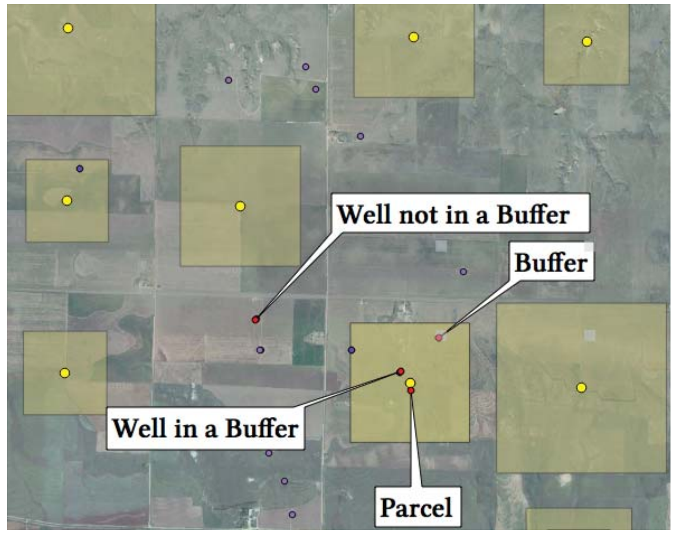
Motivation Data Methods Preliminary Results Future Extensions

Mill Levies (PVD)

- County level mill levies.
- 1987 to 2015.

Oil & Gas (KGS & EIA)

- County level annual production.
- Location of the wells.
- 3 different well types; Oil, Gas, Oil & Gas.
- Number of wells in buffer (around parcel).



Hydrological (KGS)

- Section level data.
- Depth to water, saturated thickness, hydraulic conductivity

Water Rights (WRIS)

- Merge with location of parcel and authorized place of use (not perfect)
- Authorized Inches per Acre, Seniority of water right

Motivation Data Methods Preliminary Results Future Extensions

Conclusion

- In current application, little benefits from machine learning algorithms
- Benefits from machine learning may be larger as we make more localized predictions
- Lot of other factors affect value of land for a particular individual that we cannot model

Open Questions

- What type of information is of greatest benefit?
- How small do prediction errors have to become in order for predictions to be meaningful?
- Can we open up the "black box"?
- How can the results be shared in a meaningful way?