Integrating Variety Data into Large-Scale Crop Yield Models

Joshua D. Woodard, Cornell University (jdw277@cornell.edu)
Diane R. Wang, Cornell University (drw44@cornell.edu)
Anna McClung, United States Department of Agriculture (Anna.McClung@ars.usda.gov)
Lewis Ziska, United States Department of Agriculture (Lewis.Ziska@ars.usda.gov)
Tridib Dutta, Cornell University (td276@cornell.edu)
Susan McCouch, Cornell University (srm4@cornell.edu)

Selected Proceedings Paper prepared for presentation at the 2016 Agricultural & Applied Economics Association Annual Meeting, Boston, Massachusetts, July 31-August 2

Integrating Variety Data into Large-Scale Crop Yield Models

Crop yield distribution estimation has long been a major focus of agricultural policy, insurance, and risk management research (see e.g. Woodard and Sherrick, 2014). Yet, explicit incorporation of variety or genetic data in the estimation of yield distributions remains elusive in large scale contexts both within agricultural economics as well as the broader crop sciences. Existing studies in economics tend to simply incorporate time trends in order to proxy for improvements in technology, and at best have only incorporated coarse suggestive genetic proxies (e.g., state level variables on qualitative traits such as numbers of acres with a particular trait, but not actual varietal or genetic marker data); meanwhile, the vast majority of work in plant breeding and crop sciences focuses more so on data from smaller scale experiments.

Most major crop plants have been bred to optimize yields under a relatively stable set of environmental conditions. With climate change threatening that stability, efforts to predict how crops can be adapted to withstand the new range of climatic variation can take lessons from the historical record. Documenting the relationship between past weather events and yield performance of varieties surveyed at the local level and at a large scale provides an opportunity to explore the impact of future climate projections on yield performance explicitly accounting for trends in breeding and environment interactions.

The purpose of this study is to investigate methods for integrating crop variety data into crop yield distribution models in the context of the broader U.S. rice production market. Through a unique data collection effort, we have assembled a dataset of large scale yield, weather, soil type, and varietal data for planted acreage for the last four and a half decades. This dataset represents virtually all rice grown in the southern U.S. since 1970 (approximately 125 million acres). This research is the first to our knowledge that attempts to model the relationships between varieties, yield, soil, and weather using such large scale market data from the U.S. rice market.

Data and Methods

County level rice yield data are obtained from the *United States Department of Agriculture National Agricultural Statistics Service* for 1970-2015 for Arkansas, Louisiana, Mississippi, and Texas (C=114 counties, N= 2932 observations). County level variety acreage data are obtained from the *Rice Technical Working Group Proceedings Variety Acreage Surveys* for the predominant 131 varieties in this region during the sample period (97.2% of all acreage planted). Average monthly temperature, cumulative growing degree days (GDDs), and precipitation data during the rice growing season were obtained from the PRISM Climate Group at Oregon State University (PRISM, 2016). PRISM data are published as gridded 4km resolution daily temperature and precipitation data, which we aggregate to obtain county averages.[1] Climate

[1] Only monthly PRISM data exist prior to 1980; to obtain cumulative monthly growing degree day (GDDs) estimates for the prior period of 1970-1980, we estimate regressions for each month of cumulative GDDs on average minimum and maximum monthly temperature with county level fixed effect intercepts, and district level effects on

change projection data (CMIP5) are obtained from WorldClim.Org (Hijmans et al., 2005), 2.5 minutes spatial resolution, for years 2050 (average for 2041-2060) and 2070 (average for 2061-2080), for four greenhouse gas scenarios (or representative concentration pathways, RCPs), for monthly average minimum and maximum temperature, and monthly total precipitation; the dataset includes runs from 19 different climate models; we aggregate the data at the county level. All data employed in this study including the processed data, and all data sourcing and processing scripts are available at Ag-Analytics.Org (Woodard, 2016a, Woodard, 2016b).

For each observation (year/county) variables were constructed for equal to the percent of planted acreage for each variety for that year/county. Regressions were conducted both including and excluding varieties, as well as several variants to evaluate robustness to different weather variables and fixed effects. A time trend was included as a proxy for technology; in models with varieties included, this time trend then reflects all other (potentially correlated) effects of technology net of varietal shifts. For the first set of results which exclude varieties, standard OLS regression is employed, with log yield as the dependent variable, and all other variables in levels (e.g., weather). Both agricultural district and county fixed effects models are investigated.

For the second set of models which include variety, due to the large number of independent variables when including varieties, standard OLS will not suffice since. In very high dimensional data sets, the presence of a large number of variables diminishes the interpretability of the model, and is prone to overfitting. Our dataset is semi-high dimensional relative to the number of observations (over one hundred and thirty crop variety variables) and therefore it is necessary to employ some form of penalized or regularized regression approach. We employ Least Absolute Shrinkage and Selection Operator (LASSO) regression for this purpose. LASSO and similar penalized regression techniques have been successfully employed in applications in diverse fields, including candidate gene studies for selection of genetic markers such as single nucleotide polymorphisms (SNPs), detection of gene-gene interaction, and prediction of genomic regions associated with traits of interest. Specifically, LASSO involves penalizing the size of the regression coefficients by minimizing the following equation:

$$\boldsymbol{\beta}^* = argmin_{\boldsymbol{\beta}} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_1$$

where $\boldsymbol{\beta}^*$ is a vector of estimated regression coefficients, $\boldsymbol{y}$ is an $N \times 1$ the dependent variable (yields) vector, and $\mathbf{X}$ is an $N \times K$ matrix of independent variables (e.g., weather, soil, varieties, where each column represents a different variable of interest, and rows are observations), and λ is a hyperparameter which controls the penalty. λ is tuned by finding the value of λ via cross-validation that minimizes prediction error. The concept of the penalty hyperparameter is to control the magnitude of the coefficients from becoming too large, which typically happens in unregulated OLS regression due to presence of multicollinearity. A greater penalty shrinks the coefficients towards zero. Due to presence of L_1-norm (||•||_1), some of the coefficients that are

---

the temperature variables; the R-squared values were nearly above 0.99 for the growing season months used in this study; alternative approaches did not have a substantive impact on the analysis.

close to zero are in fact shrinks exactly to zero, rendering it appropriate to use LASSO for variable selection in very high-dimensional datasets.

Results

Yield Regression Results - Varieties Excluded

Several models were fit to explore the effects of climatic variables on historical rice yield in southern U.S. The first two models investigate the effects of within-season temperature (April-September) on rice yield (Table 1). Model 1 incorporates separate mean temperature variables for each month and accounts for the effect of different months during the year. A time trend is also included, along with district level fixed effect (there are 12 agricultural districts in the rice producing states). Mean temperatures during the months of April, July, and September have highly significant effects on rice yield in addition to the time trend variable, with expected yields increasing an average of 1.3% per year. April is positive, indicating that temperatures which are too low can be deleterious during the early season, while heat is the predominant risk during the summer and harvest months. A boost in average temperature during early stages of the rice crop helps speed up developmental growth that may allow plants to reach full canopy coverage earlier in the season. It is possible that this greater photosynthetic capacity underlies the significant relationship of early season temperature and yield. Meanwhile, July mean temperature displays the greatest negative effect of all temperature variables; this may reflect the well-established fact that rice plants are most sensitive to heat stress during flowering stage.

Models 3 and 4 investigate the effects of within-season precipitation, using a similar OLS approach as Models 1 and 2 (Table 2). Due to the production system of rice, it is not as sensitive to precipitation events as to extreme temperature events during the summer months, but in some cases precipitation is significant, likely primarily due to the fact that temperature and precipitation are highly negatively correlated.

Models 5, 6, 7 and 8 include non-linear terms for temperature, and also include soil, as well as interactions of temperature with soil and time. In general, the non-linear terms for temperature are significant, but there is not a substantial increase in fit. The time and temperature interaction term is significant and positive, indicating that effect of adverse heat events has declined relatively through time, as might be expected due to gains in management and technology. Likewise, the interaction term of soil and temperature is significant, indicating that the risk of heat is moderated to some extent by soil quality.

Models 9 and 10 present similar results for GDDs instead of temperature. Not surprisingly, the results are similar, as is the fit, since the GDD threshhold is usually surpassed during the summer months during which heat is a risk (and it is in fact simply a linear transformation in that range above the threshold).

Table 1 - OLS Yield Regression Results (Temperature)

	Model 1	**Model 2**
Time Trend	0.013815***	0.013682***
Apr Temp.	0.010943***	
May Temp.	-0.00486*	
June Temp.	0.00037152	
July Temp.	-0.036934***	
Aug Temp.	-0.0051857*	
Sep Temp.	0.0073874***	
Avg. Apr-Sep. Temp		-0.0074675***
N	2932	2932
Adj. R^2	0.7114	0.6801
Sigma^2	0.0137	0.0152

Note: Table presents results from regression of log county yield on time trend and mean temperature for April-Sept, T = 1970 - 2015, N = 2932 observations, using District level fixed effects. Significance levels:=10%. **=5%, ***=1%.*

Table 2 - OLS Yield Regression Results (Precipitation)

	Model 3	**Model 4**
Year	0.013224***	0.013305***
April Precip.	-0.000270***	
May Precip.	-0.000173***	
June Precip.	-9.5354e-05**	
July Precip.	3.7885e-05	
Aug. Precip.	-7.8932e-05	
Sept. Precip.	-0.0001431***	
Avg(Apr-Sep)		-0.0008117***
N	2932	2932
Adj. R^2	0.6931	0.6884
Sigma^2	0.0146	0.0148

Note: Table presents results from regression of log county yield on time trend and total precipitation for April-Sept, T = 1970 - 2015, N = 2932 observations, using District level fixed effects. Significance levels:=10%. **=5%, ***=1%.*

Table 3 - OLS Yield Regression Results (Precipitation, Soil, and Interactions)

	Model 5	**Model 6**	**Model 7**	**Model 8**
Time Trend	0.013815***	-0.011497*	0.013502***	0.014283***
July Temp.	0.01028	-2.091151***	-0.195852*	-0.204934*
NCCPI	2.6423***	0.097168***	0.097003***	1.705356
July Temp.-Sq.		0.004058**	0.002809	0.002970*
July Precip	-0.000244***	-0.000228***	-0.006427	-0.000250***
*Time *July Temp*		0.000917***		
*July Temp*Soil*	-0.09126***			
*Time *Precip.*			0.000003	
*Time *Soil*				-0.000808
N	2932	2932	2932	2932
Adj. R^2	0.7095	0.7103	0.7077	0.7076
Sigma^2	0.0138	0.0138	0.0139	0.0139

Note: Table presents results from regression of log county yield on time trend, NCCPI soil index, mean temperature July Temperature and Temperature squared, and year interaction, T = 1970 -2015, N = 2932 observations, using District level fixed effects. Significance levels:=10%. **=5%, ***=1%.*

Table 4 - OLS Yield Regression Results (Growing Degree Days)

	Model 9	**Model 10**
Time Trend	0.013968***	0.013712***
Apr growing degree days	0.0002135***	
May growing degree days	-8.3645e-05	
June growing degree days	1.1681e-05	
July growing degree days	-0.00067354***	
Aug. growing degree days	-8.9459e-05	
Sep. growing degree days	0.00013997***	
Avg(Apr-Sep) growing degree days		-0.00014157***
N	2932	2932
Adj. R^2	0.7119	0.6801
Sigma^2	0.0137	0.0152

Note: Table presents results from regression of log county yield on time trend and growing degree days for April-Sept, T = 1970 -2015, N = 2932 observations, using District level fixed effects. Significance levels:=10%. **=5%, ***=1%.*

Yield Regression Results - Varieties Included

Table 5 presents results from LASSO regression of log county yield on time trend, NCCPI soil quality index, mean temperature and total precipitation. Parameter values are approximately the percent change in yield per unit change in the variable. For example, Time Trend variable result indicates that estimated yield increased of 0.85% per year due to factors other than varietal changes and climate (e.g., management, other environmental changes, etc.). This is an interesting result, as it is significantly lower than when varieties are excluded.

When varieties are excluded, trend effects associated with varietal improvements through time are subsumed within the time proxy; however, in the LASSO results with varieties included, the value is still positive but declines significantly. Specifically, the value in the LASSO model is about 0.85% per year due to technology not associated with variety improvements, but is about 1.5% per annum in total. Thus, approximately half of the technology gains during the period can likely be attributed to genetic improvements associated with shifts in varieties planted. To our knowledge, this is the first such study to differentiate management and genetic effects in this manner. Figure 1 graphs the estimated effect by variety, against when the variety was introduced. We only included varieties which were included in the bootstrapped model at least 80% of the time. While not universally positive for all varieties introduced through time, there are apparent and abrupt large gains in a number of major varieties through time, as well as an average upward slope. This indicates that in fact that varieties (net of climate change impacts, other technological advancements, or shifts in production across better or worse soil) are on average improving significantly through time. This is perhaps not surprising given the investment in breeding programs, the many advancements in breeding technology, as well as simply the nature of the artificial selection process.

The rest of the results as it regards impacts of soil quality (as indicated by NCCPI, which is mapped to soil type), temperature, and precipitation, are fairly consistent with the OLS models excluding varieties. Figures 2 and 3 graphically present the estimated coefficient values and their bootstrapped error bars for temperature and precipitation (respectively) by month.

Table 6.1 and 6.2 present the top ranking varieties by coefficient value for long grain varieties, and also report the percent of time it was included in the bootstrapped model, whether it was a Clearfield variety, as well as breeding program and year of introduction of the variety. Only varieties which were included 80% of the time or more where included in the tabulation. Table 6.3 presents similar results for medium and short grain varieties. There is a fairly large spread in terms of varietal performance.

Table 5- LASSO Yield Model Results

Variable	Parameter Value	Bootstrapped Standard Error
Time Trend	0.00848**	0.00055
NCCPI	0.12489**	0.00793
April Temperature	0.00801**	0.00040
April Precip.	-0.00018**	0.00001
May Temperature	0.00000	0.00000
May Precip.	-0.00014**	0.00001
June Temperature	0.00207*	0.00083
June Precip.	-0.00016**	0.00001
July Temperature	-0.02562**	0.00088
July Precip.	-0.00024**	0.00002
August Temperature	-0.00697**	0.00085
August Precip.	-0.00017**	0.00001
September Temperature	0.00000	0.00000
September Precip.	-0.00007**	0.00001

*Note: Table presents results from LASSO regression of log county yield on time trend, NCCPI soil quality index, mean temperature and total precipitation for April-Sept., and varieties (V = 131), for T = 1970 -2015, N = 2983 observations, using District level fixed effects. Variety results for this model are summarized in Table 6. Parameter values are approximately the percent change in yield per unit change in the variable. Significance levels: *=5%, **=1%.*

Figure 1- LASSO Parameter Estimate by Year of Variety's Introduction (Long Grain)

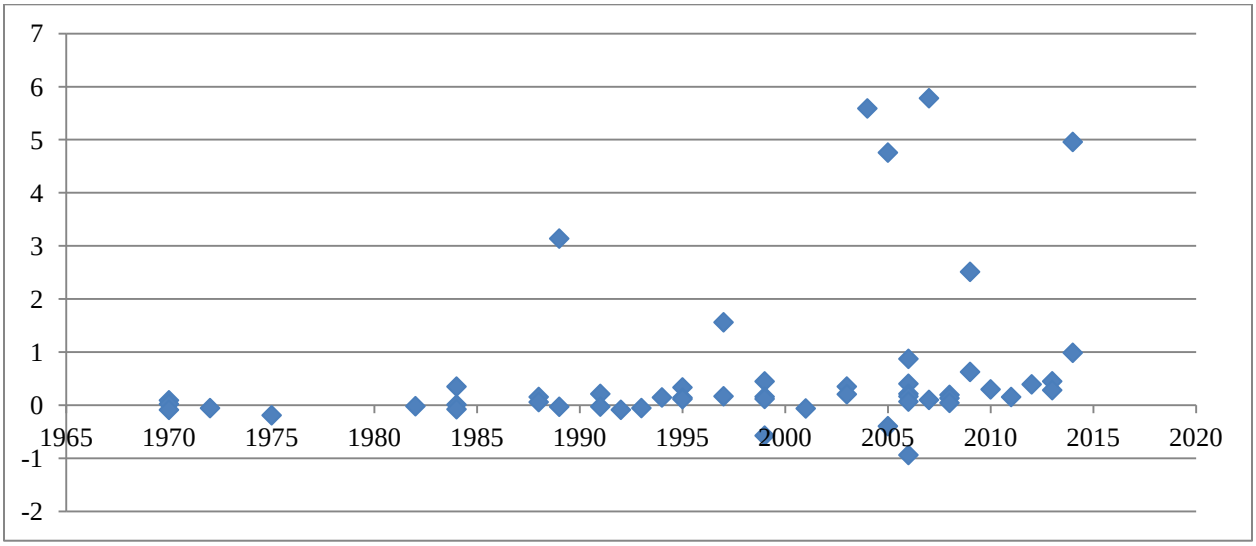

Figure 2- LASSO Temperature Parameter Estimate by Month of Season

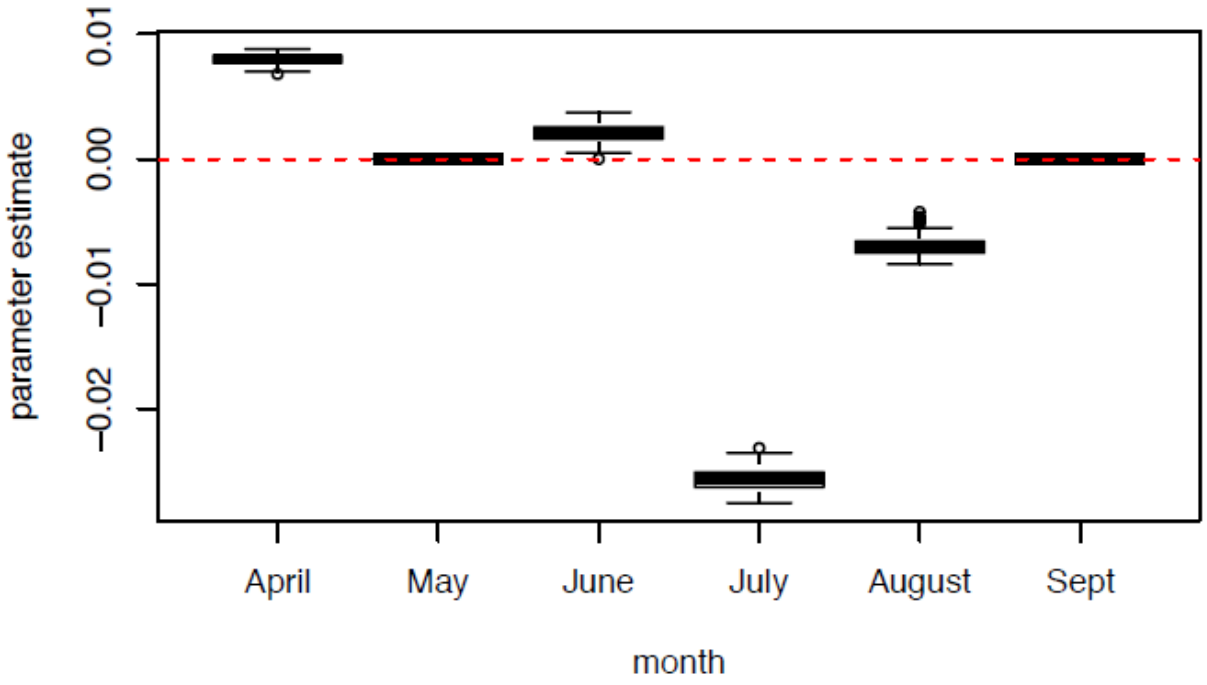

Figure 3- LASSO Precipitation Parameter Estimate by Month of Season

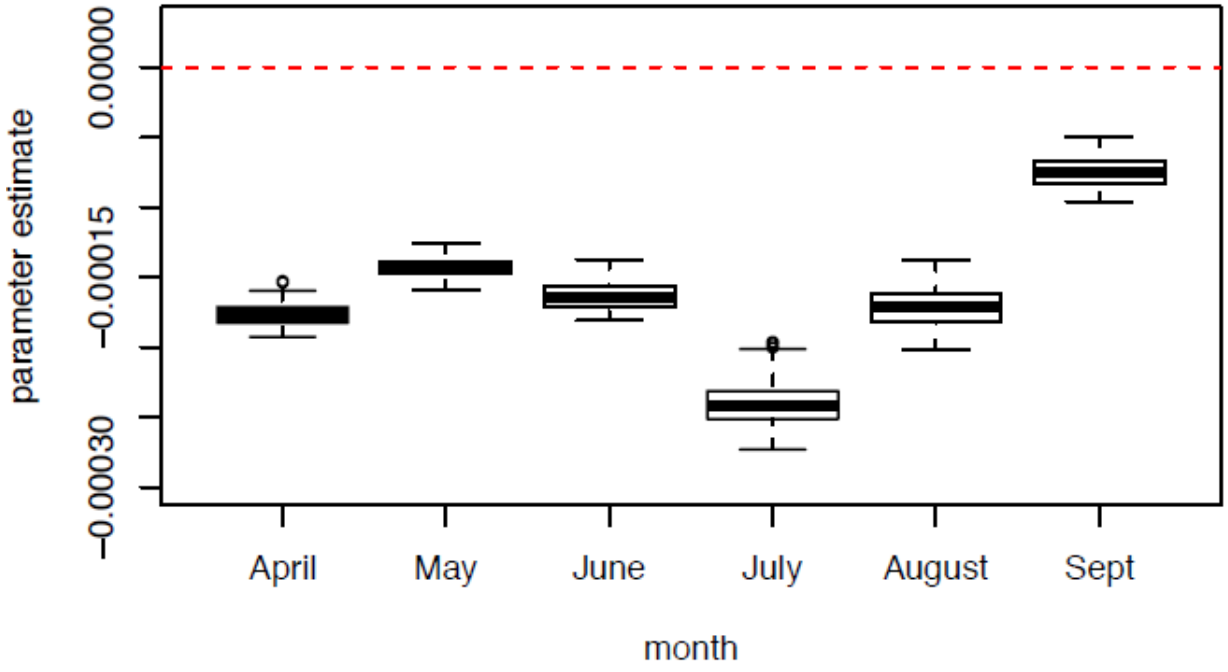

Table 6.1 - LASSO Regression Results, Ranking of Long Grain Varieties

Variety	Year of Introduction	Parameter Val.	Frequency Included in Model	CL vs non-CL	Long/Med./Short grain vs. Other	Breeding Program	conventional vs. specialty
HIDALGO	2007	5.784	88%	non	long	USDA/TX	specialty
XP710	2004	5.591	100%	non	long	RiceTec	conventional
XL760	2014	4.960	99%	non	long	RiceTec	conventional
XP712	2005	4.759	81%	non	long	RiceTec	conventional
V7817	1989	3.139	100%	Non	long	PRIVATE CO TX	conventional
CLXL746	2009	2.513	100%	CL	long	RiceTec	conventional
LITTON	1997	1.562	100%	non	long	MS	conventional
ANTONIO	2014	0.988	98%	non	long	TX	conventional
XP729	2006	0.874	98%	non	long	RiceTec	conventional
XP746	2009	0.627	93%	non	long	RiceTec	conventional
MERMENTAU	2013	0.450	100%	non	long	LA	conventional
DIXIEBELLE	1999	0.448	100%	non	long	USDA/TX	specialty
XL723	2006	0.405	100%	non	long	RiceTec	conventional
XL753	2012	0.393	100%	non	long	RiceTec	conventional
XL8	2003	0.352	98%	non	long	RiceTec	conventional
SKYBONNET	1984	0.351	100%	non	long	USDA/TX	conventional
JODON	1995	0.335	97%	non	long	LA	conventional
CL111	2010	0.298	100%	CL	long	LA	conventional
JAZZMAN 2	2013	0.284	100%	non	long	LA	specialty
CLXL730	2006	0.215	100%	CL	long	RiceTec	conventional
JACKSON	1991	0.213	100%	non	long	MS	conventional
FRANCIS	2003	0.207	100%	non	long	AR	conventional
XP744	2008	0.195	95%	non	long	RiceTec	conventional
JEFFERSON	1997	0.168	100%	non	long	USDA/TX	conventional
PRESIDIO	2006	0.166	100%	non	long	USDA/TX	conventional

Table 6.2- LASSO Regression Results, Ranking of Long Grain Varieties

Variety	Year of Introduction	Parameter Val.	Frequency Included in Model	CL vs non-CL	Long/Med./Short grain vs. Other	Breeding Program	conventional vs. specialty
PRISCILLA	1999	0.160	100%	non	long	MS	conventional
GULFMONT	1988	0.156	100%	non	long	USDA/TX	conventional
CL142 AR	2011	0.155	95%	CL	long	AR	conventional
LAGRUE	1994	0.146	100%	non	long	AR	conventional
KAYBONNET	1995	0.142	100%	non	long	AR	conventional
CLXL745	2008	0.132	100%	CL	long	RiceTec	conventional
COCODRIE	1999	0.119	100%	non	long	LA	conventional
WELLS	1995	0.115	100%	non	long	AR	conventional
CLXL729	2007	0.100	100%	CL	long	RiceTec	conventional
BLUEBELLE	1970	0.092	100%	Non	long	USDA/TX	conventional
TRENASSE	2006	0.068	89%	non	long	LA	conventional
L202	1988	0.059	99%	non	long	CA	conventional
CL151	2008	0.044	98%	CL	long	LA	conventional
BLUEBONNET	1970	0.022	97%	Non	long	USDA/TX	conventional
NEWBONNET	1984	0.012	85%	non	long	AR	conventional
OTHERS	1982	-0.019	96%	non	long	unknown	conventional
ALAN	1991	-0.027	98%	non	long	AR	conventional
KATY	1989	-0.029	86%	non	long	AR	conventional
CYPRESS	1993	-0.054	100%	CL	long	LA	conventional
LABELLE	1972	-0.055	100%	non	long	USDA/TX	conventional
CL121	2001	-0.061	84%	CL	long	LA	conventional
BOND	1984	-0.076	86%	Non	long	AR	conventional
LACASSINE	1992	-0.087	99%	non	long	LA	conventional
STARBONNET	1970	-0.087	100%	Non	long	AR	conventional
LEBONNET	1975	-0.190	100%	non	long	USDA/TX	conventional
XP723	2005	-0.395	100%	non	long	RiceTec	conventional
MADISON	1999	-0.573	100%	non	long	USDA/TX	conventional
CLXP730	2006	-0.938	95%	CL	long	RiceTec	conventional

Table 6.3 - LASSO Regression Results, Ranking of Medium and Short Grain Varieties

Variety	Year of Introduction	Parameter Val.	Frequency Included in Model	CL vs non-CL	Long/Med./Short grain vs. Other	Breeding Program	conventional vs. specialty
ROSES	1970	4.045	100%	Non	medium	USDA/TX, AR	conventional
PEARL	1970	3.374	100%	Non	short	CA	conventional
CAFFEY	2014	1.009	98%	non	medium	LA	conventional
NORTAI	1973	0.297	100%	Non	short	AR	conventional
RISOTTO	2003	0.242	98%	non	medium	unknown	specialty
CL261	2010	0.240	94%	CL	medium	LA	conventional
NOVA	1970	0.239	100%	Non	medium	AR	conventional
BENGAL	1993	0.033	98%	non	medium	LA	conventional
MARS	1979	-0.016	99%	non	medium	AR	conventional
SATURN	1970	-0.081	100%	non	medium	AR	conventional
VISTA	1972	-0.123	97%	Non	medium	LA	conventional
NATO	1970	-0.131	100%	Non	medium	LA	conventional
MELROSE	1978	-0.135	100%	non	medium	Alexandria seed c	conventional
RICO I	1989	-0.196	100%	non	medium	TX	conventional
BRAZOS	1975	-0.203	100%	Non	medium	USDA/TX	conventional
MEDARK	1989	-0.285	96%	non	medium	AR	conventional
LAFITTE	1997	-1.106	100%	non	medium	LA	conventional

Conclusions and Future Work

Using a unique, large scale dataset with semi-high resolution variety, weather, acreage, and soil data, this study explored the feasibility of modeling variety specific effects for the U.S. rice market. Significant variation was found among crop variety performance, and on average varieties have improved through time as it regards yield impacts. We find that about half of all technology gains in the U.S. rice market can be explained by the introduction of, and shifting towards, new rice varieties and adaptation related to adoption of better performing varieties, suggesting that the market actively adapts. This is perhaps not surprising given the investments in these markets, the rational actions of economic agents in choosing varieties, as well as advancements in breeding programs. This is somewhat in contrast with the extant knowledge in some literatures which assume production risk is increasing in agriculture, and that it is non-adaptive, but which ignore the intersection of fairly well known processes occurring in the market (such as advancements in biotechnology, active breeding programs, farmer adaptation, etc.).

Continued expansion of the human population, accompanied by rising incomes and expectations, drives increasing demand for food, fiber, and energy at a time when sustained fossil fuel use and deforestation have raised the concentration of atmospheric carbon dioxide [CO_2] and other trace gases in the atmosphere. Levels of carbon dioxide (CO_2) and other gases that absorb in the infra-red portion of the spectrum are increasing faster than scientists had predicted (Canadell et al., 2007). Thus, developing interdisciplinary frameworks offers a unique model system within which to link the disciplines of genetics and physiology, climate change, risk analysis, and economics. Because rice also serves as a model organism for crop plants, results from this study may have wider implications for new strategies and approaches to studying the relationship between genomic signatures and climate resilience in other major cereal crops such as maize and wheat, as well as lead to new estimation paradigms.

The next steps of the research will involve genotyping the seeds for these historical varieties obtained through a partnership with USDA. Seeds will be genotyped using Genotyping-By-Sequencing (GBS) to recover Single Nucleotide Polymorphism (SNP) markers. GBS is a multiplexed genotyping technology based on Restriction Site Associated DNA (RAD) tags (Elshire, 2011) that was originally developed for human genomics, and since been optimized for use on numerous plant species including rice. This technology is ideal for populations with low levels of diversity, making it a good choice for genotyping the relatively small tropical japonica genepool from which most US varieties are derive.

When working with genetic marker data, one will typically have up to several thousand possible markers to choose from. This creates a problem since the number of explanatory variables is vastly greater than the number of observations. In our case, we have about 20,000 county/annual/variety observations, 3,500 county/annual yield observations, but well over 100,000 genetic markers (a small subset of the entire genome, although these have been reduced

down to about 4,000 markers after applying standard filtering techniques to eliminate markers with inadequate data quality or which lack diversity in the population).

Traditional modeling approaches make the assumption that the genetic markers are uncorrelated, and then run successive OLS regressions with each genetic marker (usually employing family-wise error rate corrections; Sebastiani and Solovieff, 2011), in order to identify the most potentially important markers. More recent approaches employ Bayesian models in order to improve modeling efficiency. LASSO (estimation such as that used here), PCA based methods, as well as machine learning techniques have also been employed to reduce the number of dimensions to enable the feasibility of model identification. Indeed, the physics of reproductive processes lead to what is known as genetic linkage, or a propensity for nucleotides which are physically close together to be inherited (or move) together in a population. From a empirical standpoint, the implication is that markers located close together on the genome tend to be very highly correlated within any population of varieties. These biophysical actualities render the problem conducive to dimension reduction methods, but do lead to interesting modeling challenges. We will further explore a variety of methods for variable selection and selection of functional form , including out-of-sample (i.e., cross validation) optimization methods, Bayesian models (e.g., random coefficient models, variable selection models, etc.), LASSO, and mixed-models methods in this future work.

References

Canadell, J.G., Le Quéré, C., Raupach, M.R., Field, C.B., Buttenhuis, E.T. et al. (2007), 'Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks', *Proceedings National Academy Sciences* (USA), 104: 18866-18870.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6(5): e19379. doi:10.1371/journal.pone.0019379.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, 2016.

Rice Technical Working Group Proceedings Variety Acreage Surveys, (1970-2015).

Sebastiani, P and Solovieff , B (2011) Genome Wide Association Studies, *Problem Solving Handbook in Computational Biology and Bioinformatics*, pp 159-175.

Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, Vol 58, No. 1, pp. 267–288, 1996.

Woodard, J.D., B.J. Sherrick, "Estimation of Mixture Models using Cross-Validation Optimization: Implications for Crop Yield Distribution Modeling,"*American Journal of Agricultural Economics*93-4(2011):968-982.

Woodard J.D. (2016), Data Science and Management for Large Scale Empirical Applications in Agricultural and Applied Economics Research," *Applied Economics Perspectives and Policy*

Woodard J.D. (2016), "Big Data and Ag-Analytics: An Open Source, Open Data Platform for Agricultural & Environmental Finance, Insurance, and Risk," *Agricultural Finance Review*.